Offline Contextual Bandits in the Presence of New Actions

Anonymous authors
Paper under double-blind review

Abstract

Automated decision-making algorithms drive applications in domains such as recommendation systems and search engines. These algorithms often rely on off-policy contextual bandits or *off-policy learning* (OPL). Conventionally, OPL selects actions that maximize the expected reward from an existing action set. However, in many real-world scenarios, actions, such as news articles or video content, change continuously, and the action space evolves over time compared to when the logged data was collected. We define actions introduced after deploying the logging policy as *new actions* and focus on the problem of OPL with new actions. Existing OPL methods identify optimal actions from the existing set effectively. However, these methods cannot learn and select new actions because no relevant data are logged. To address this limitation, we propose a new OPL method that leverages action features. In particular, we first introduce the Local Combination PseudoInverse (LCPI) estimator for the policy gradient, generalizing the PseudoInverse estimator initially proposed for off-policy evaluation of slate bandits. LCPI controls the trade-off between reward-modeling condition and the condition for data collection regarding the action features, capturing the interaction effects among different dimensions of action features. Furthermore, we propose a generalized algorithm called ***Policy Optimization for New Actions (PONA)***, which integrates LCPI, a component specialized for new action selection, with Doubly Robust (DR), which excels at learning within existing actions. We define PONA as a weighted sum of the LCPI and DR estimators, optimizing both the selection of existing and new actions, and allowing the proportion of new action selections to be adjusted by controlling the weight parameter. Through extensive experiments, we demonstrate that PONA efficiently selects new actions while maintaining the overall policy performance as opposed to most existing methods that cannot select new actions.

1 Introduction

We increasingly rely on data-driven algorithms to optimize decision-making in various domains, including recommendation systems (Gilotte et al., 2018; Saito & Joachims, 2021; Felicioni et al., 2022; Saito et al., 2021b;a), search engines (Li et al., 2015), advertising (Bottou et al., 2013), and medical treatments (Kallus & Zhou, 2018). These problems take the form of contextual bandits, where a policy selects actions based on observed contexts, and the corresponding rewards become available. The goal in such problems is to learn a policy that maximizes the expected reward. *Off-Policy Learning* (OPL) (Swaminathan & Joachims, 2015a; Metelli et al., 2021; Swaminathan et al., 2017; Saito et al., 2024a) enables learning such policies by leveraging previously collected data, eliminating the need for online deployment and avoiding the risk of negatively impacting user experiences through exploration (Saito & Joachims, 2021; Kiyohara et al., 2023). Typically, OPL focuses on identifying the best actions within a predefined set of existing actions. However, in many real-world scenarios, the action space evolves over time as new actions emerge or existing ones are updated. For example, search systems must process new articles daily, and recommendation systems for products or videos continually welcome new items. We define these actions, introduced after the data collection phase, as ***new actions***. In this work, we rigorously address the challenges of OPL regarding the presence of such new actions.

Existing OPL methods can be broadly categorized into policy-based and regression-based approaches (Saito & Joachims, 2021; Saito et al., 2024b). Regression-based methods first predict the expected rewards of

existing actions based on supervised learning on logged data, and then pick the action that has the highest estimated reward for each context (Jeunen & Goethals, 2021). In contrast, policy-based methods update the policy parameter based on the policy gradient (Swaminathan & Joachims, 2015b), which is estimated using techniques such as inverse propensity score (IPS) (Strehl et al., 2010) and doubly robust (DR) (Dudík et al., 2014). These methods are effective at identifying optimal actions within the set of existing ones. However, they are not designed to evaluate the effectiveness of new actions. In particular, the naive applications of existing methods cannot choose new actions at all, failing to give fair opportunities to them.

To address these critical limitations of the typical OPL methods, we consider and formulate a scenario where actions are represented as multi-dimensional action features. There exist many real-life problems where the actions are represented by features (Felicioni et al., 2022; Saito & Joachims, 2022). For example, in the problem of thumbnail optimization (Figure 1), as done at Netflix (Amat et al., 2018), actions correspond to thumbnails described by features such as character type (e.g., male, female, child), title position (e.g., top, middle, bottom), and title size (e.g., large, small). In particular, the left side of the figure depicts existing thumbnails in the predefined action set, while the right side illustrates new thumbnails that are never observed in the logged data. As in this illustration, we consider leveraging the action features to develop a new OPL method that effectively selects new actions (i.e., unobserved combinations of action features) while still achieving competitive expected rewards.

Merely applying existing regression-based or policy-based approaches to observable action features does not resolve the problem of new actions. This is because the combinations of action features that represent new actions are never observed in the logged data. A possible solution might be to draw inspiration from the PseudoInverse (PI) method introduced in the slate bandit setting (Swaminathan et al., 2017; Vlassis et al., 2021; Kiyohara et al., 2024). Slate bandits refer to the setting where a policy aims to optimize the selection of slates, which consist of multiple action features (which are also called sub-actions) (Swaminathan et al., 2017). To deal with the combinatorial slate spaces and resulting variance issues, the PI estimator relies on two conditions (Swaminathan et al., 2017; Vlassis et al., 2021) (1) *the logging policy provides full support for each action feature independently*, and (2) *the expected reward function (q-function) is a linear combination of the intrinsic value of each feature dimension*. Under these conditions, the PI estimator can estimate the policy value or policy gradient without bias even if there are new actions. However, the linearity condition is particularly restrictive because it treats the effects of action features as completely independent, ignoring potentially significant interaction effects and thus introducing bias into the estimation (Kiyohara et al., 2024). Therefore, we relax the linearity condition of PI to deal with local interactions among action features and propose **Local Combination PseudoInverse (LCPI)**, a generalized version of the PI estimator. LCPI enables new policies to account for interactions among different dimensions of action features, allowing a more effective selection of new actions. Although LCPI is better at identifying effective new actions, existing OPL methods, such as the policy-based approach using DR to estimate the policy gradient (Dudík et al., 2014), can be more effective for learning within existing actions. This is because they do not rely on any conditions about the form of the q-function. To leverage the strengths of both approaches, we finally introduce a hybrid algorithm named ***Policy Optimization for New Actions (PONA)***, which integrates LCPI and DR by weighting them through a weight parameter. This design allows PONA to optimize the overall policy performance via the effective selection of existing actions while identifying promising new actions. Furthermore, by adjusting the weight parameter, PONA can control how aggressively it selects new actions, supporting both conservative (greater focus on existing actions) and aggressive (greater focus on new actions) strategies. Results on synthetic and real-world experiments demonstrate that PONA efficiently selects new actions and achieves satisfactory overall performance, whereas existing methods fail to choose new actions.

2 Preliminaries: Conventional OPL

This section formulates the typical problem of OPL and introduces existing OPL methods.

First, let $x \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$ represent a context vector such as user features in recommender systems. Given a context x, a possibly stochastic policy $\pi(a|x)$ selects an action $a \in \mathcal{A}$ such as products or movies. Furthermore, let $r \in \mathbb{R}_{\geq 0}$ denote a reward, such as conversion label and play duration, drawn from an unknown conditional

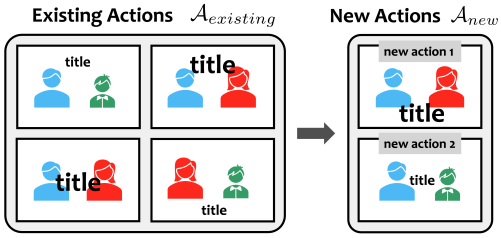

Figure 1: Examples of Existing and New Actions.
Note: There are three types of action features: **character type** (chosen as two from male, female, or child), **title position** (top, center, or bottom), and **title size** (large or small). The left side of the figure illustrates the thumbnails for existing actions, while the right side shows examples of new actions: "Male and female, title position at the bottom, large title" (**new action 1**) and "Male and child, title position at the center, small title" (**new action 2**).

distribution $p(r|x,a)$. In the setting of OPL, we are provided with a logged dataset $\mathcal{D} := \{(x_i, a_i, r_i)\}_{i=1}^n$ collected under a logging policy π_0 such that $(x,a,r) \sim p(x)\pi_0(a|x)p(r|x,a)$.

We define the expected reward under the deployment of a policy π (policy value) as a measure of its overall performance $V(\pi) := \mathbb{E}_{p(x)\pi(a|x)p(r|x,a)}[r] = \mathbb{E}_{p(x)\pi(a|x)}[q(x,a)]$, where $q(x,a) := \mathbb{E}[r|x,a]$ is the expected reward function or *q-function* for a given context x and action a.

The goal in OPL is to learn a new policy π_θ, parameterized by θ, using only the logged dataset $\mathcal{D}$. When learning a policy, we aim to maximize the policy value as $\theta^* = \mathrm{argmax}_{\theta\in\Theta} V(\pi_\theta)$. The following describes two typical approaches to OPL.

Firstly, the **policy-based** approach updates the policy parameter θ using iterative gradient ascent, $\theta_{t+1} \leftarrow \theta_t + \eta \nabla_\theta V(\pi_\theta)$, where

$$\nabla_\theta V(\pi_\theta) := \mathbb{E}_{p(x)}\left[\sum_{a\in\mathcal{A}} \pi_\theta(a|x) q(x,a) \nabla_\theta \log \pi_\theta(a|x)\right], \tag{1}$$

is the policy gradient (PG). Since the true PG is unknown, we need to estimate it using the logged data $\mathcal{D}$. A common approach for this estimation is to use the IPS method (Precup et al., 2000)

$$\nabla_\theta \widehat{V}_{\mathrm{IPS}}(\pi_\theta; \mathcal{D}) := \frac{1}{n}\sum_{i=1}^n \frac{\pi_\theta(a_i|x_i)}{\pi_0(a_i|x_i)} r_i \nabla_\theta \log \pi_\theta(a_i|x_i), \tag{2}$$

where $\pi_\theta(a|x)/\pi_0(a|x)$ is called the *importance weight*. It plays a key role in ensuring the unbiasedness of IPS but often causes the issue of high variance and inefficient policy learning (Saito et al., 2024b).

An improved approach to estimate the PG is using the DR estimator (Dudík et al., 2014). It leverages a q-function estimator $\hat{q}(x,a)$ to reduce the variance in the PG estimation compared to IPS as

$$\nabla_\theta \widehat{V}_{\mathrm{DR}}(\pi_\theta; \mathcal{D}) := \frac{1}{n}\sum_{i=1}^n \frac{\pi_\theta(a_i|x_i)}{\pi_0(a_i|x_i)} \left(r_i - \hat{q}(x_i,a_i)\right) \nabla_\theta \log \pi_\theta(a_i|x_i) + \frac{1}{n}\sum_{i=1}^n \sum_{a\in\mathcal{A}} \pi_\theta(a|x_i)\hat{q}(x_i,a) \nabla_\theta \log \pi_\theta(a|x_i). \tag{3}$$

IPS and DR are both unbiased against the true PG under the *full support* condition (**a condition to ensure sufficient data collection by the logging policy** π_0).

Condition 2.1. *(Full Support) The logging policy π_0 has full support if $\pi_0(a|x) > 0, \forall x \in \mathcal{X}, \forall a \in \mathcal{A}$.*

Secondly, the **regression-based** approach estimates the q-function $q(x,a)$ using off-the-shelf supervised machine learning methods. It then transforms the estimated q-function $\hat{q}_\theta(x,a)$ into a decision-making policy, for example, by applying the softmax function as

$$\pi_\theta(a|x) = \frac{\exp(\hat{q}_\theta(x,a))}{\sum_{a'\in\mathcal{A}} \exp(\hat{q}_\theta(x,a'))}. \tag{4}$$

This approach may fail due to bias issues resulting from the difficulty in accurately estimating the q-function $q(x, a)$. However, it can avoid the issue of high variance compared to the policy-based approach based on IPS and DR (Jeunen & Goethals, 2021).

Existing policy-based or regression-based methods perform effectively in identifying existing actions with high expected rewards (Swaminathan & Joachims, 2015a; Metelli et al., 2021; Saito et al., 2024b). However, these methods cannot learn to select new actions, even if some of the new actions have high expected rewards. This is because we do not observe any data about new actions in $\mathcal{D}$, which severely violates *full support* (Condition 2.1). Therefore, we need to first relax the typical condition of full support to effectively learn policies that can choose promising new actions.

3 Related Work

This section summarizes the key related work and their connections to OPL with new actions. A more thorough discussion on related work can be found in Appendix A.

Off-Policy Evaluation (OPE) and Learning (OPL) (Dudík et al., 2014; Wang et al., 2017; Liu et al., 2018; Farajtabar et al., 2018; Su et al., 2019; 2020; Kallus et al., 2021; Metelli et al., 2021; Saito & Joachims, 2022; Saito et al., 2023) aim to evaluate and optimize decision-making policies using only historical logged data. Various methods and estimators have been developed in OPE to control the bias and variance of the estimation of the performance of new policies (Su et al., 2020; 2019; Swaminathan & Joachims, 2015b;c; Dudík et al., 2014; Farajtabar et al., 2018; Kallus et al., 2021; Kiyohara et al., 2022; 2023; Metelli et al., 2021; Saito & Joachims, 2021; Wang et al., 2017; Kiyohara et al., 2024). In OPL, these estimation techniques enable accurate estimation of the policy gradient (Saito et al., 2024b). Unfortunately, most existing methods in OPL rely crucially on full support (Condition 2.1) and focus solely on selecting the most effective actions within those already present in the logged data (Sachdeva et al., 2020). Thus, they cannot learn and give fair opportunities to new actions that are potentially available in many decision-making problems.

A setting particularly relevant to ours is OPE for slate bandits. It considers selecting a combinatorial action called *slate* composed of multiple sub-actions. For example, in medical treatment scenarios, one may aim to select the optimal combination of drug doses to improve patient outcomes. While these systems often have access to extensive logged data, an accurate OPE is challenging due to the exponential increase in variance associated with slate-wise importance weighting (Swaminathan et al., 2017; Su et al., 2020; Kiyohara et al., 2024). To address this challenge, several versions of *PseudoInverse* (PI) estimators have been proposed (Su et al., 2020; Swaminathan et al., 2017; Vlassis et al., 2021). They use slot-wise importance weights to relax the support condition and reduce the variance of typical estimators such as IPS. They have also been proven to enable unbiased OPE under the *linearity* assumption on the q-function. Linearity requires that the q-function be linearly decomposable, ignoring interaction effects among different slots. While PI often outperforms IPS under linear reward structures, it introduces significant bias when the linearity assumption does not hold (Kiyohara et al., 2024; Shimizu et al., 2024). To effectively handle new actions in OPL by leveraging action features, we extend the idea of PI to estimate the policy gradient by relaxing the linearity assumption on the reward. This relaxation is crucial for dealing with the real-world non-linearities in reward functions during policy learning. It should be noted that the key contributions of our work are substantially different from those of existing studies for slate bandits (Swaminathan et al., 2017; Kiyohara et al., 2024; Vlassis et al., 2021; Chaudhari et al., 2024). We focus on the problem of policy *learning with new actions*. In contrast, the existing literature around slate OPE addresses only the problem of policy *evaluation without considering new actions*.

Another setting similar to new actions is OPL with deficient actions (Sachdeva et al., 2020; Felicioni et al., 2022), which refer to the actions that have zero probability of being selected under the logging policy for *certain* contexts x. Rigorously, the set of deficient actions is written as $\{a \in \mathcal{A} \mid \exists x \in \mathcal{X}, \pi_0(a \mid x) = 0\}$. Deficient action is a broader concept that encompasses the new actions we consider as a (more challenging) special case. Due to this distinction, existing OPL approaches to deal with deficient actions (Sachdeva et al., 2020; Felicioni et al., 2022) cannot handle new actions.

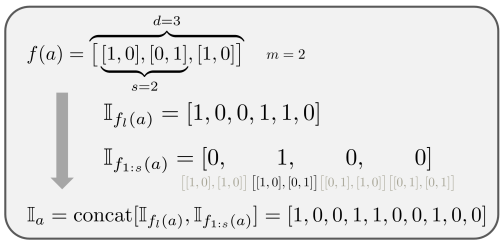

Figure 2: An example of converting action features into a vector with binary values by representing each dimension ($f_l(a)$) by their respective one-hot vectors. The vector representing the independent effect of each dimension ($\mathbb{I}_{f_l(a)}$) is then concatenated with $\mathbb{I}_{f_{1:s}(a)}$, which represents interaction effects.

4 The Proposed Approach

The previous sections discussed the challenges associated with selecting new actions using existing OPL methods. To overcome these challenges, we propose a solution that leverages action features, which are often available in real-world decision-making problems (Saito & Joachims, 2022; Peng et al., 2023; Felicioni et al., 2022).

4.1 Formulation of OPL with Action Features

We now represent an action a by the d-dimensional action features $f(a) = (f_1(a), \ldots, f_l(a), \ldots, f_d(a))$, where $f_l(a) \in \mathcal{F}_l$ denotes the l-th dimension. In particular, we consider the setting where each action feature $f_l(a)$ is discrete and can be represented as a one-hot vector. For example, the genre features such as 'documentary', 'romance', and 'horror' are encoded as $[1,0,0]$, $[0,1,0]$, and $[0,0,1]$, respectively, as shown in Figure 2.

In OPL with action features, we can possibly address the presence of new actions by being able to represent them as combinations of already observed factors. Figure 1 illustrates an example in the context of thumbnail selection with the 3-dimensional action features. The left side of the figure shows the existing action set, which includes thumbnails already considered by the logging policy. The right side introduces new thumbnails as new actions. For example, **new action 1** consists of "male and female characters," "title positioned at the bottom," and "large title size." We can see that we have already observed every individual feature of **new action 1** separately in the existing action set. **Despite this, existing OPL methods struggle to handle these new actions because the logged data does not include specific combinations of action features representing the new actions**. Therefore, we propose a novel method and algorithm that leverage these action features through a new estimator for the PG to identify effective new actions.

Before introducing our proposed approach, we rigorously introduce three key types of action spaces.

1. **The space of *all* actions**: The set of all possible combinations of action features: $\mathcal{A} := \{a \mid a = (f_1, \ldots, f_d),\ \forall f_1 \in \mathcal{F}_1, \ldots, \forall f_d \in \mathcal{F}_d\}$.

2. **The space of *existing* actions**: The set of actions that have some positive probability under the logging policy: $\mathcal{A}_{existing} := \{a \in \mathcal{A} \mid \exists x \in \mathcal{X},\ \pi_0(a|x) > 0\}$.

3. **The space of *new* actions**: The set of actions that are never chosen under the logging policy: $\mathcal{A}_{new} := \{a \in \mathcal{A} \mid \forall x \in \mathcal{X},\ \pi_0(a|x) = 0\}$.

These action spaces satisfy the following relations: $\mathcal{A}_{existing} \cap \mathcal{A}_{new} = \emptyset, \quad \mathcal{A}_{existing} \cup \mathcal{A}_{new} = \mathcal{A}$. In the typical formulation for OPL, the objective is to identify actions that maximize the expected reward from a set of *existing* actions $\mathcal{A}_{existing}$. In contrast, we consider a more general and realistic scenario, where the goal is to select actions that maximize the expected reward from $\mathcal{A}$, which includes new actions $\mathcal{A}_{new}$.

4.2 The Proposed Algorithm: PONA

To design a new algorithm for OPL with new actions, we first draw inspiration from the PI estimator initially proposed for OPE in slate bandits (Swaminathan et al., 2017) as described in Section 3, and provide a generalization.

To introduce PI, we begin with formally defining a one-hot vector for each action feature as $\mathbb{I}_{f_l(a)} \in \mathbb{R}^{dm}$.[1] As shown in Figure 2, this is a flattened vector that represents the one-hot encoding of each action feature. Using this notation, we can first extend the PI estimator (Swaminathan et al., 2017) as an estimator for the PG as follows.

$$\nabla_\theta \widehat{V}_{\mathrm{PI}}(\pi_\theta; \mathcal{D}) = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{a \in \mathcal{A}} \pi_\theta(a|x_i) \nabla_\theta \log \pi_\theta(a|x_i) \mathbb{I}_{f_l(a)}{}^T \right) \Gamma^{\dagger}_{\pi_0, x_i} \mathbb{I}_{f_l(a_i)} r_i, \tag{5}$$

where $\Gamma_{\pi_0,x} := \mathbb{E}_{\pi_0(a|x)}[\mathbb{I}_{f_l(a)} \mathbb{I}^T_{f_l(a)} | x]$, and $M^\dagger$ denotes the Moore-Penrose pseudoinverse of the matrix M.

The PI estimator introduced above relies on the following conditions (4.1 and 4.2) to ensure unbiased estimation of the PG:

Condition 4.1. *(Independent Support) The logging policy π_0 ensures independent support if $\pi_0(f_l|x) > 0, \forall x \in \mathcal{X}, \forall l \in [1, ..., d], \forall f_l \in \mathcal{F}_l$. Note that $\pi_0(f_l|x) = \sum_{a \in \mathcal{A}: f_l(a) = f_l} \pi_0(a \mid x)$ is the marginal probability of observing f_l under π_0.*

This condition requires the logging policy to independently support every dimension of the action features. The condition of Independent Support is weaker than the Full Support condition (Condition 2.1), which requires the logging policy to choose every combination of action features. When new actions exist, the Full Support condition cannot hold, but the Independent Support condition may still hold even in the presence of new actions. **Due to this relaxation of the support condition, the PI estimator can learn a new policy that chooses new actions**.

The next condition of PI is with regard to the q-function.

Condition 4.2. *(Linearity) For each context $x \in \mathcal{X}$, there exists an (unknown) intrinsic reward vector $\phi_{x,l} \in \mathbb{R}^{dm}$ such that*

$$q(x, a) = \sum_{l=1}^{d} q_l(x, f_l(a)) = \mathbb{I}^T_{f_l(a)} \phi_{x,l}.$$

The linearity condition requires that the q-function be expressed as a linear combination of latent value functions for each dimension of the feature. **This condition essentially assumes no interaction effects between different dimensions of the action feature**. While extending the PI estimator to estimate the PG is straightforward as done in Eq. 5, the linearity condition rarely holds and introduces significant bias in PG estimation (Kiyohara et al., 2024; Vlassis et al., 2021).

To overcome this critical limitation of the naive extension of PI, we first generalize Independent Support to account for local interactions of action features as follows.

Condition 4.3. *(Local Combination Support) The logging policy π_0 is said to ensure local combination support if, for the first s dimensions of the action features, $\pi_0(f_{1:s}|x) > 0, \forall x \in \mathcal{X}, \forall f_{1:s} \in \prod_{j=1}^{s} \mathcal{F}_j$ as well as $\pi_0(f_l|x) > 0$ for the rest $s+1 \leq l \leq d$.*

Note that $f_{1:s} = (f_1, ..., f_s)$ represents the first s dimensions of the action features, where $1 \leq s \leq d$. The condition of Local Combination Support requires $f_{1:s}$ to be fully supported under the logging policy.[2] Condition 4.3 remains weaker than the original Full Support condition (Condition 2.1), because Full Support

[1] Here, we assume that each dimension of the action feature has the same number of possible values ($m = |\mathcal{F}_l|$) for ease of exposition. However, this is not strictly necessary, and different dimensions can indeed have varying sizes in practice.

[2] Note that we focus on the first s dimensions just for ease of exposition, as the dimensions of the action feature can be arbitrarily reordered.

requires every dimension of the action features to be simultaneously supported (equivalent to Local Combination Support with $s = d$). When $s = 1$, Local Combination Support reduces to Condition 4.1. Along with this extension, we generalize the linearity condition for the q-function (Condition 4.2) as described below.

Condition 4.4. *(Local Linearity) Define a vector with binary values for representing the first s dimensions of the action features as $\mathbb{I}_{f_{1:s}} \in \mathbb{R}^{m^s}$. The overall action indicator $\mathbb{I}_a$ is then defined as $\mathbb{I}_a := concat[\mathbb{I}_{f_l}, \mathbb{I}_{f_{1:s}}]$. We say that the q-function meets Local Linearity, if for each context $x \in \mathcal{X}$, there exists an (unknown) intrinsic reward vector $\phi_x = concat[\phi_{x,l} \in \mathbb{R}^{dm}, \phi_{x,1:s} \in \mathbb{R}^{m^s}]$ such that:*

$$q(x,a) = \underbrace{\sum_{l=1}^{d} q_l(x, f_l(a)) + q(x, f_{1:s}(a))}_{\mathbb{I}_{f_l(a)}^T \phi_{x,l} \quad + \quad \mathbb{I}_{f_{1:s}(a)}^T \phi_{x,1:s}} = \mathbb{I}_a^T \phi_x.$$

The Local Linearity condition allows the first s dimensions of the action features to have interaction effects with each other. Therefore, this condition is weaker than linearity of PI (Condition 4.2), because linearity does not allow any interaction effects across different action features. When $s = 1$, the condition of local linearity reduces to the linearity condition (Condition 4.2).

Building on these generalized conditions, we propose a new policy gradient estimator, called the Local Combination PseudoInverse (LCPI) estimator, generalizing the PI estimator (Swaminathan et al., 2017) as

$$\nabla_\theta \widehat{V}_{\mathrm{LCPI}}(\pi_\theta; \mathcal{D}) = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{a \in \mathcal{A}} \pi_\theta(a|x_i) \nabla_\theta \log \pi_\theta(a|x_i) \mathbb{I}_a{}^T \right) \Gamma_{\pi_0, x_i}^{\dagger} \mathbb{I}_{a_i} r_i, \tag{6}$$

where $\Gamma_{\pi_0,x} := \mathbb{E}_{\pi_0(a|x)}[\mathbb{I}_a \mathbb{I}_a^T | x]$.

LCPI satisfies unbiasedness against the PG $\nabla_\theta V(\pi_\theta)$ under Conditions 4.3 and 4.4, i.e., $\boldsymbol{E}_{\mathcal{D}}[\nabla_\theta \widehat{V}_{\mathrm{LCPI}}(\pi_\theta; \mathcal{D})] = \nabla_\theta V(\pi_\theta)$.[3] This indicates that LCPI can estimate the reward of new actions while dealing with the interaction effects among action features (the original PI estimator cannot deal with the interaction effects).

Although LCPI is more effective for selecting new actions than PI by dealing with the interaction effects, traditional policy- or regression-based methods may achieve better performance within existing actions. This is because they do not rely on such conditions when focusing only on existing actions. Therefore, there is an interesting trade-off between our LCPI and traditional methods. While traditional methods focus on identifying the optimal action within existing actions, LCPI is superior for choosing new actions based on local linearity. This interesting trade-off leads us to finally develop the **Policy Optimization for New Actions (PONA)** algorithm, which integrates LCPI with the DR estimator[4] as follows.

$$\nabla_\theta \widehat{V}_{\mathrm{PONA}}(\pi_\theta; \kappa, \mathcal{D}) = \kappa \cdot \nabla_\theta \hat{V}_{\mathrm{LCPI}}(\pi_\theta; \mathcal{D}) + (1 - \kappa) \cdot \nabla_\theta \hat{V}_{\mathrm{DR}}(\pi_\theta; \mathcal{D}), \tag{7}$$

where $\kappa \in [0, 1]$ is a hyperparameter that controls the trade-off between focusing on new actions and maximizing the overall policy value by prioritizing existing actions. A larger value of κ increases the weight of the LCPI component, which leads to greater reliance on the local linearity condition and a stronger focus on optimizing new actions. However, this approach sacrifices the effectiveness of identifying the optimal actions within the set of existing actions. In contrast, a smaller value of κ increases the weight of DR, reducing reliance on local linearity and shifting the focus toward optimizing existing actions. This results in a more conservative policy that rarely selects new actions. We can tune the value of κ in a data-driven manner using a cross-validation procedure. Multiple formulations of cross-validation can be considered for tuning κ, depending on the specific objectives. For instance, one can simply maximize the policy value $V(\pi)$ on the validation set when tuning κ. Alternatively, it is possible to impose constraints on the proportions of new

[3] Please see the appendix for the derivation.

[4] Note that PONA is general and can be combined with any existing estimator for the PG other than DR. We use DR as a part of our PONA algorithm because it is often the most effective method among traditional ones (Saito et al., 2024b).

actions under the learned policy, allowing control over how aggressive or conservative the policy should be:

$$\max_{\kappa} \ \hat{V}(\pi_{\theta,\kappa};\mathcal{D}) \quad \text{s.t. } \rho_L \leq \mathbb{E}_{p(x)}[\sum_{a\in\mathcal{A}_{new}} \pi_{\theta,\kappa}(a|x)] \leq \rho_U, \tag{8}$$

where ρ_L and ρ_U define the desired range for the proportion of new actions. In the next section, we will empirically demonstrate that, by performing this cross-validation procedure and changing the values of ρ_L and ρ_U, we can flexibly control the proportion between existing and new actions under the learned policy. Note that, when performing cross-validation, we can estimate the policy value $\hat{V}(\pi_{\theta,\kappa})$ of the policy induced by κ by applying off-the-shelf OPE estimators such as IPS and DR on the validation set.

Remark. Local combination support and local linearity do introduce structural assumptions, but they remain substantially weaker than the full set of conditions required by the PseudoInverse estimator. In practice, the satisfaction of local combination support is often controllable when the logging policy can be designed by practitioners. It is also important to emphasize that our objective is not to obtain fully unbiased gradient estimates nor to perfectly satisfy these conditions. Our goal is to improve upon baseline methods in settings with new actions, even when the assumptions are only approximately met. As demonstrated in our experiments in the following section, the proposed methods consistently outperform the baselines across a wide range of scenarios, including those with significant degrees of condition violations.

5 Empirical Evaluation

This section empirically evaluates our proposed method, PONA, and its special case (LCPI). We provide anonymized code in the supplementary material to facilitate reproducibility.

5.1 Synthetic Experiments

We create synthetic data to evaluate policy learning algorithms and their ability to select effective new actions based on the ground-truth policy value.

We begin by sampling context vectors x from a normal distribution. The action features span $d = 5$ dimensions, with each feature having $m = 3$ possible values. The total number of possible actions is thus $|\mathcal{A}| = 3^5 = 243$. We then define the synthetic q-function as follows:

$$q(x,a) = \sum_{l=1}^{d} \underbrace{q_l(x, f_l(a))}_{:=x^T M_{f_l(a)}} + \underbrace{q_{1:s}(x, f_{1:s}(a))}_{:=x^T M_{f_{1:s}(a)}} + \gamma \cdot \underbrace{q_{1:d}(x, f_{1:d}(a))}_{:=x^T M_{f_{1:d}(a)}},$$

where γ is a parameter that controls the contribution of the last term $q_{1:d}(x, f_{1:d}(a))$. M_{f_l}, $M_{f_{1:s}}$, and $M_{f_{1:d}}$ are parameter matrices sampled from a uniform distribution. The first term of the q-function, $q_l(x, f_l(a))$, represents the independent effect of each dimension of the action feature. The second term, $q_{1:s}(x, f_{1:s}(a))$, captures the interaction effects within the first s dimensions of the action feature. Most methods, including IPS, DR, and our method, can handle the first two terms because they do not involve interaction effects beyond the first s dimensions. Only the original PI method (Swaminathan et al., 2017) cannot estimate the second term due to its reliance on the linearity condition. The last term of the q-function, $q_{1:d}(x, f_{1:d}(a))$, accounts for the interaction effects among all dimensions of the action feature. The proposed methods cannot estimate the last term without bias due to their reliance on local linearity, while existing methods can achieve unbiased estimation of the q-function within existing actions regardless of the interaction effects. By adjusting the γ parameter, we control the degree of the condition's violation and evaluate the robustness of our methods under these violations.

Next, we define the logging policy π_0 by applying the softmax function to the expected reward function $q(x,a)$ within the set of existing actions as follows.

$$\pi_0(a|x) := \begin{cases} \dfrac{\exp(\beta \cdot q(x,a))}{\sum_{a'\in\mathcal{A}_{existing}} \exp(\beta \cdot q(x,a'))} & \text{if} \quad a \in \mathcal{A}_{existing} \\ 0 & \text{if} \quad a \in \mathcal{A}_{new} \end{cases}$$

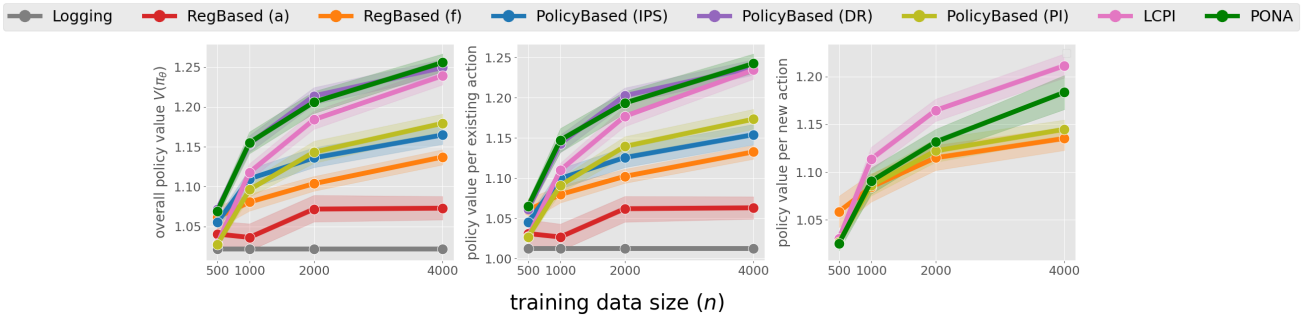

Figure 3: Comparisons of the overall policy value, policy value per existing action, and policy value per new action with varying training data sizes (n). Note that the metrics are normalized by those of the uniform random policy.

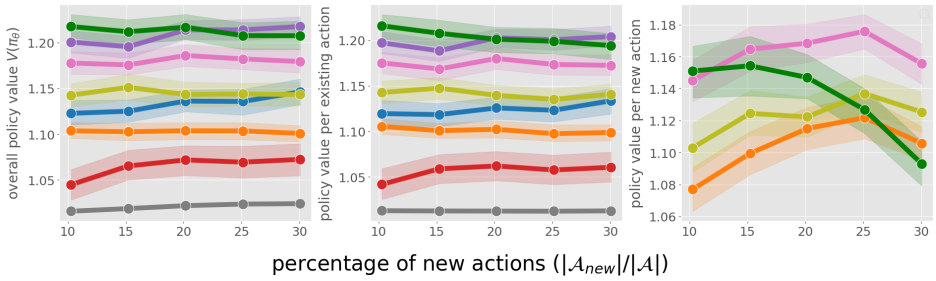

Figure 4: Comparisons of the overall policy value, policy value per existing action, and policy value per new action with varying percentages of new actions ($|\mathcal{A}_{new}|/|\mathcal{A}|$). Note that the metrics are normalized by those of the uniform random policy.

where β is a parameter that adjusts the trade-off between optimality and entropy in the logging policy. We set $\beta = 0.05$ throughout our experiments. Note that the above formulation of the logging policy ensures that the new actions in $\mathcal{A}_{new}$ have zero probability of being selected by the logging policy for any context x.

Compared Methods. We compare PONA and LCPI with the logging policy π_0, the regression-based method using action index (RegBased (a)), the regression-based method using action feature (RegBased (f)), policy-based method w/ IPS (PolicyBased (IPS)), policy-based method w/ DR (PolicyBased (DR)), and policy-based method w/ PI (PolicyBased (PI)). We report RegBased (f) as a baseline method that estimates the q-function using action features as input. It then picks the best action according to the estimated reward function $\hat{q}_\theta(x, f_{1:d}(a))$. RegBased ($f$) differs from RegBased (a) in that it may select new actions. To tune the hyperparameter κ for PONA, we perform a grid search over the range $[0, 0.25, 0.5, 0.75, 1.0]$, selecting the best value in terms of optimizing the policy value $V(\pi)$ on the validation data.

Results. We report **overall policy value**: $\mathbb{E}[\sum_{a \in \mathcal{A}} \pi_\theta(a|x) q(x,a)]$, **policy value per existing action**: $\frac{\mathbb{E}[\sum_{a \in \mathcal{A}_{existing}} \pi_\theta(a|x) q(x,a)]}{\mathbb{E}[\sum_{a \in \mathcal{A}_{existing}} \pi_\theta(a|x)]}$, and **policy value per new action**: $\frac{\mathbb{E}_{p(x)}[\sum_{a \in \mathcal{A}_{new}} \pi_\theta(a|x) q(x,a)]}{\mathbb{E}[\sum_{a \in \mathcal{A}_{new}} \pi_\theta(a|x)]}$ of policies learned by each OPL method, averaged across 200 simulations. **Overall policy value** measures the overall effectiveness of the learned policy. **Policy value per existing action** evaluates the quality of selecting existing actions, while **policy value per new action** assesses that of selection within new actions. **Note that the logging policy π_0, RegBased (a), PolicyBased (IPS), and PolicyBased (DR) do not choose new actions at all, and thus their results do not appear in the figures of <u>policy value per new action</u>.**

How does PONA perform with varying sizes of training data? First, we compare the OPL methods using varying sizes of logged training data ($n \in \{500, 1000, 2000, 4000\}$), as shown in Figure 3. The figure shows that most methods, including PONA and LCPI, improve in overall policy value, policy value per

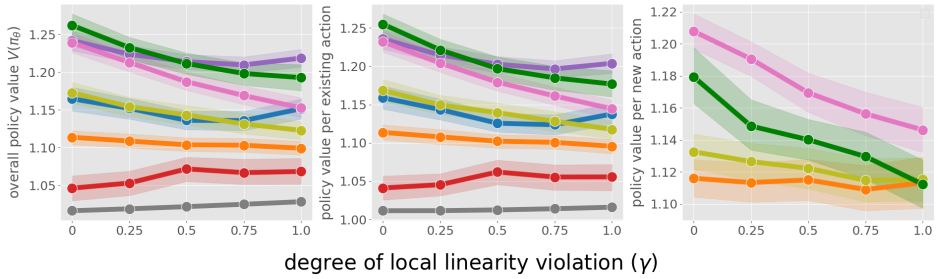

Figure 5: Comparisons of the overall policy value, policy value per existing action, and policy value per new action with varying degrees of local linearity violation (γ). Note that the metrics are normalized by those of the uniform random policy.

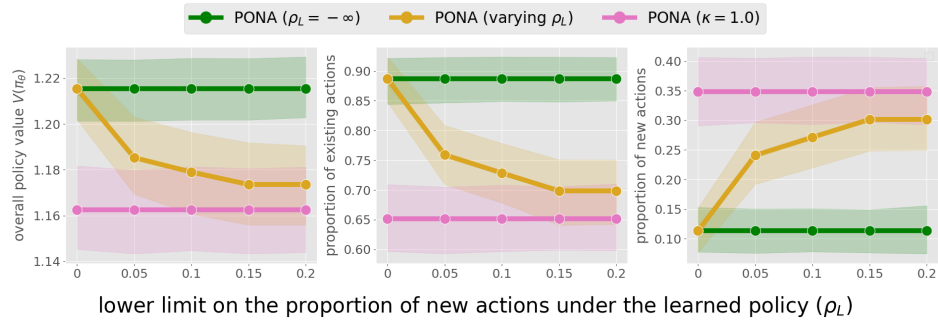

Figure 6: Comparisons of the overall policy value, proportion of existing actions, and proportion of new actions under the learned policy with varying lower limits on the proportion of new actions (ρ_L in Eq. 8).

existing action, and policy value per new action as the training data size increases, which aligns with expectations. Notably, LCPI performs particularly well in terms of the selection of new actions when the training data size is relatively large ($n = 2000, 4000$). However, it underperforms PolicyBased (DR) in terms of overall policy value. In contrast, PONA combines the strengths of LCPI and PolicyBased (DR), maintaining a policy value per new action comparable to RegBased (f) while matching PolicyBased (DR) in terms of overall policy value. Considering that existing methods, including PolicyBased (DR), cannot select new actions at all, it is noteworthy that **PONA not only selects new actions and gives fair opportunities to them as effectively as RegBased (f) but also achieves overall performance competitive with PolicyBased (DR).**

How does PONA perform with varying numbers of new actions? Next, Figure 4 illustrates how the OPL algorithms perform as the percentage of new actions, $100 \cdot \frac{|\mathcal{A}_{new}|}{|\mathcal{A}|}$, increases. As the percentage of new actions grows, the number of existing actions decreases. The figure shows that methods focusing exclusively on existing actions (RegBased (a), PolicyBased (IPS), PolicyBased (DR)) exhibit a slight improvement in policy value for existing actions as the percentage of new actions increases. This is because, with fewer existing actions, it becomes easier to identify the optimal action within the reduced set. Nevertheless, PONA consistently achieves the same or slightly better overall policy value compared to PolicyBased (DR), while also selecting new actions for any percentage of new actions. In comparison, PolicyBased (DR) and other traditional methods never select new actions, even as the percentage of new actions increases. PolicyBased (PI) is mostly worse than PONA and LCPI in every metric due to its reliance on linearity, which is severely violated in our experimental environment.

How does PONA perform when Condition 4.4 is violated? We next evaluate the robustness of the proposed methods against violations of their key condition, namely, local linearity. To test this, we increase the value of γ in the definition of the synthetic reward function on the x-axis of Figure 5. The figure demonstrates that as γ increases, the violations of the local linearity condition become more pronounced,

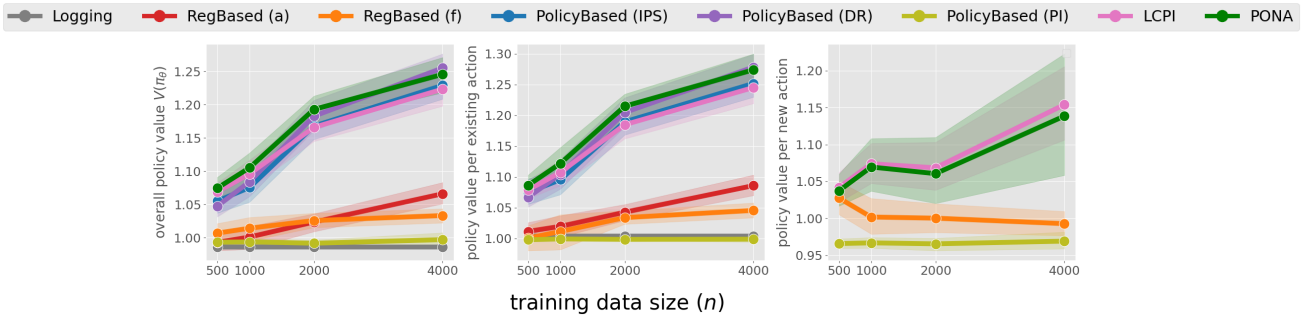

Figure 7: Comparisons of the overall policy value, policy value per existing action, and policy value per new action with varying training data sizes (n) on the KuaiRec dataset. Note that the metrics are normalized by those of the uniform random policy.

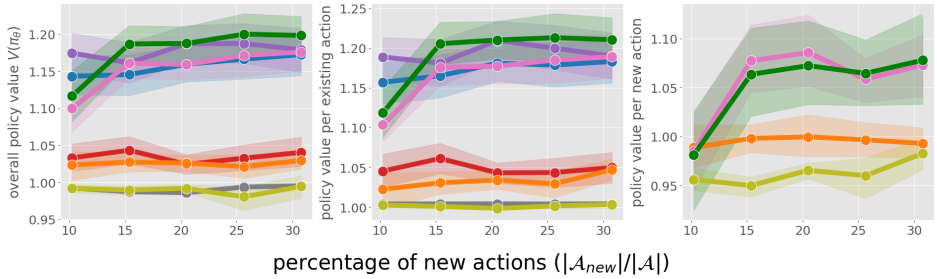

Figure 8: Comparisons of the overall policy value, policy value per existing action, and policy value per new action with varying percentages of new actions ($|\mathcal{A}_{new}|/|\mathcal{A}|$) on the KuaiRec dataset. Note that the metrics are normalized by those of the uniform random policy.

leading to a gradual decrease in the overall policy value of LCPI. In contrast, existing methods that do not rely on any assumptions about the q-function exhibit little change in policy value with varying values of γ, which is reasonable. Most interestingly, PONA, which also relies on local linearity, is more robust to the condition's violation than LCPI. This occurs because its data-driven tuning process in Eq. 8 adjusts its hyperparameter κ to assign greater weight to the DR component as the violation becomes more severe. This adjustment adaptively maintains the policy value for existing actions, thereby preserving a high overall policy value. Even in scenarios where the ability to learn new actions is at its weakest, PONA performs comparably to RegBased (f) for new actions, significantly outperforming PolicyBased (DR) and other baseline methods in selecting new actions. These results highlight PONA's ability to maintain a high policy value and effectively learn new actions, even when its key condition is severely violated.

How does PONA perform when varying the constraints on the proportions of new actions during parameter tuning? In the previous experiments, we did not impose any constraints on the behavior of the learned policy when tuning the weight parameter κ of PONA (i.e., always setting $\rho_L = -\infty$ and $\rho_U = \infty$ in Eq. 8). As a result, the tuning procedure selected the κ value that simply maximized the overall policy value on the validation data.

In Figure 6, we report the overall policy value and the proportions of existing and new actions under the learned policies for varying lower limits ρ_L in Eq. 8, while keeping $\rho_U = \infty$. Note that we include the metrics for PONA ($\rho_L = -\infty$) and PONA ($\kappa = 1.0$) as references in the figure. This experiment allows us to investigate how well PONA's behavior can be controlled through the tuning of κ. The right plot in Figure 6 demonstrates that, as the lower limit on the proportion of new actions becomes larger, the proportion of new actions under **PONA (with varying ρ_L)** increases. This result shows that we can control the aggressiveness of the learned policy by appropriately formulating the tuning process of the weight parameter κ. It is also reasonable that the overall policy value of PONA decreases with larger ρ_L, as this increases the reliance of PONA on the local linearity condition.

5.2 Real-World Data

To evaluate the real-world applicability of PONA, we assess its performance on the KuaiRec dataset (https://kuairec.com/) (Gao et al., 2022), which is derived from the recommendation logs of a video-sharing mobile app. This dataset includes a matrix with 1,411 users and 3,327 videos, where nearly 100% of the rewards are observable. Consequently, we can directly use the observed rewards to evaluate OPL methods on test data, which makes this dataset the only suitable real-world public dataset for our experiments. We use user information as the context x and video IDs as the actions a. We then rely on video tags and video categories to construct action features $f(a)$. More specifically, the video categories follow a three-level hierarchical structure. Combining the video tags and categories creates a four-dimensional discrete action-feature representation. To avoid an excessively large Cartesian action space, we first keep the top 15 values in each action-feature dimension and then retain only valid feature combinations that correspond to at least one video in the KuaiRec dataset. After this filtering, the resulting action set contains $|\mathcal{A}| = 117$ valid combinations. We use the watch ratio, defined as the play duration divided by the video duration, as the reward r. We define the logging policy π_0 by applying the softmax function to the expected reward function $q(x, a)$ within the set of existing actions as

$$\pi_0(a|x) := \begin{cases} \dfrac{\exp(\beta \cdot q(x,a))}{\sum_{a' \in \mathcal{A}_{existing}} \exp(\beta \cdot q(x,a'))} & \text{if} \quad a \in \mathcal{A}_{existing} \\ 0 & \text{if} \quad a \in \mathcal{A}_{new} \end{cases}$$

where β is a parameter that adjusts the trade-off between optimality and entropy in the logging policy. We set $\beta = 0.05$ throughout our experiments. Note that the above formulation of the logging policy ensures that the new actions in $\mathcal{A}_{new}$ have zero probability of being selected by the logging policy for any context x.

Compared Methods. We compare PONA and LCPI with the logging policy π_0, the regression-based method using action index (RegBased (a)), the regression-based method using action feature (RegBased (f)), policy-based method w/ IPS (PolicyBased (IPS)), policy-based method w/ DR (PolicyBased (DR)), and policy-based method w/ PI (PolicyBased (PI)). We report RegBased (f) as a baseline method that estimates the q-function using action features as input. It then picks the best action according to the estimated reward function $\hat{q}_\theta(x, f_{1:d}(a))$. RegBased ($f$) differs from RegBased (a) in that it may select new actions. To tune the hyperparameter κ for PONA, we perform a grid search over the range $[0, 0.25, 0.5, 0.75, 1.0]$, selecting the best value in terms of optimizing the policy value $V(\pi)$ on validation data.

Results. We compare the overall policy values learned by each OPL method across 50 simulations with different random seeds. All other settings match those of the synthetic experiments. Figures 7 and 8 present the results with varying training data sizes and percentages of new actions. The results demonstrate trends similar to those observed in the synthetic experiments. PONA achieves the overall policy value comparable to PolicyBased (DR). PONA also successfully identifies promising new actions and performs much better than RegBased (f) in terms of **policy value per new action**. Note that typical methods, including PolicyBased (DR), did not choose new actions at all on the real-world data as well. RegBased (f) and PolicyBased (PI) perform equally or worse compared to the uniform random policy in selecting new actions. This is because the real-world dataset contains a larger number of values for each action feature, making it more challenging to learn the rewards of new actions via mere regression or linearity-based methods.

6 Conclusion

In this paper, we tackled the challenge of off-policy learning with new actions. Existing methods cannot choose new actions simply because we do not observe any data about them. We first proposed Local Combination PseudoInverse (LCPI) for the policy gradient, which relaxes the restrictive conditions of the PseudoInverse method by accounting for interactions among action features, enabling the more effective selection of new actions. To balance the strengths of LCPI for new actions and Doubly Robust for existing actions, we also introduced the Policy Optimization for New Actions (PONA) algorithm. By integrating these estimators through a weight parameter, PONA achieves efficient action selection while maintaining strong overall performance.

References

Fernando Amat, Ashok Chandrashekar, Tony Jebara, and Justin Basilico. Artwork personalization at netflix. In *Proceedings of the 12th ACM conference on recommender systems*, pp. 487–488, 2018.

Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(11), 2013.

Shreyas Chaudhari, David Arbour, Georgios Theocharous, and Nikos Vlassis. Distributional off-policy evaluation for slate recommendations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 8265–8273, 2024.

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

Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. More robust doubly robust off-policy evaluation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 1447–1456. PMLR, 2018.

Nicolo Felicioni, Maurizio Ferrari Dacrema, Marcello Restelli, and Paolo Cremonesi. Off-policy evaluation with deficient support using side information. *Advances in Neural Information Processing Systems*, 35, 2022.

Chongming Gao, Shijun Li, Wenqiang Lei, Jiawei Chen, Biao Li, Peng Jiang, Xiangnan He, Jiaxin Mao, and Tat-Seng Chua. Kuairec: A fully-observed dataset and insights for evaluating recommender systems. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 540–550, 2022.

Alexandre Gilotte, Clément Calauzènes, Thomas Nedelec, Alexandre Abraham, and Simon Dollé. Offline a/b testing for recommender systems. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*, pp. 198–206, 2018.

Olivier Jeunen and Bart Goethals. Pessimistic reward models for off-policy learning in recommendation. In *Proceedings of the 15th ACM Conference on Recommender Systems*, pp. 63–74, 2021.

Nathan Kallus and Angela Zhou. Policy evaluation and optimization with continuous treatments. In *International Conference on Artificial Intelligence and Statistics*, pp. 1243–1251. PMLR, 2018.

Nathan Kallus, Yuta Saito, and Masatoshi Uehara. Optimal off-policy evaluation from multiple logging policies. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 5247–5256. PMLR, 2021.

Haruka Kiyohara, Yuta Saito, Tatsuya Matsuhiro, Yusuke Narita, Nobuyuki Shimizu, and Yasuo Yamamoto. Doubly robust off-policy evaluation for ranking policies under the cascade behavior model. In *Proceedings of the 15th International Conference on Web Search and Data Mining*, 2022.

Haruka Kiyohara, Masatoshi Uehara, Yusuke Narita, Nobuyuki Shimizu, Yasuo Yamamoto, and Yuta Saito. Off-policy evaluation of ranking policies under diverse user behavior. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1154–1163, 2023.

Haruka Kiyohara, Masahiro Nomura, and Yuta Saito. Off-policy evaluation of slate bandit policies via optimizing abstraction. In *Proceedings of the ACM on Web Conference 2024*, pp. 3150–3161, 2024.

Lihong Li, Shunbao Chen, Jim Kleban, and Ankur Gupta. Counterfactual estimation and optimization of click metrics in search engines: A case study. In *Proceedings of the 24th International Conference on World Wide Web*, pp. 929–934, 2015.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: infinite-horizon off-policy estimation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 5361–5371, 2018.

Alberto Maria Metelli, Alessio Russo, and Marcello Restelli. Subgaussian and differentiable importance sampling for off-policy evaluation and learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Jie Peng, Hao Zou, Jiashuo Liu, Shaoming Li, Yibao Jiang, Jian Pei, and Peng Cui. Offline policy evaluation in large action spaces via outcome-oriented action grouping. In *Proceedings of the ACM Web Conference 2023*, pp. 1220–1230, 2023.

Doina Precup, Richard S. Sutton, and Satinder P. Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the 17th International Conference on Machine Learning*, pp. 759–766, 2000.

Noveen Sachdeva, Yi Su, and Thorsten Joachims. Off-policy bandits with deficient support. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 965–975, 2020.

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*, pp. 828–830, 2021.

Yuta Saito and Thorsten Joachims. Off-policy evaluation for large action spaces via embeddings. In *International Conference on Machine Learning*, pp. 19089–19122. PMLR, 2022.

Yuta Saito, Shunsuke Aihara, Megumi Matsutani, and Yusuke Narita. 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*, 2021a.

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*, pp. 114–123, 2021b.

Yuta Saito, Qingyang Ren, and Thorsten Joachims. Off-policy evaluation for large action spaces via conjunct effect modeling. *arXiv preprint arXiv:2305.08062*, 2023.

Yuta Saito, Himan Abdollahpouri, Jesse Anderton, Ben Carterette, and Mounia Lalmas. Long-term off-policy evaluation and learning. In *Proceedings of the ACM on Web Conference 2024*, pp. 3432–3443, 2024a.

Yuta Saito, Jihan Yao, and Thorsten Joachims. Potec: Off-policy learning for large action spaces via two-stage policy decomposition. *arXiv preprint arXiv:2402.06151*, 2024b.

Tatsuhiro Shimizu, Koichi Tanaka, Ren Kishimoto, Haruka Kiyohara, Masahiro Nomura, and Yuta Saito. Effective off-policy evaluation and learning in contextual combinatorial bandits. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pp. 733–741, 2024.

Alex Strehl, John Langford, Lihong Li, and Sham M Kakade. Learning from logged implicit exploration data. In *Advances in Neural Information Processing Systems*, volume 23, pp. 2217–2225, 2010.

Yi Su, Lequn Wang, Michele Santacatterina, and Thorsten Joachims. Cab: Continuous adaptive blending for policy evaluation and learning. In *International Conference on Machine Learning*, volume 84, pp. 6005–6014, 2019.

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, pp. 9167–9176. PMLR, 2020.

Adith Swaminathan and Thorsten Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *The Journal of Machine Learning Research*, 16(1):1731–1755, 2015a.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, pp. 814–823. PMLR, 2015b.

Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. *Advances in Neural Information Processing Systems*, 28, 2015c.

Adith Swaminathan, Akshay Krishnamurthy, Alekh Agarwal, Miro Dudik, John Langford, Damien Jose, and Imed Zitouni. Off-policy evaluation for slate recommendation. In *Advances in Neural Information Processing Systems*, volume 30, pp. 3632–3642, 2017.

Masatoshi Uehara, Chengchun Shi, and Nathan Kallus. A review of off-policy evaluation in reinforcement learning. *arXiv preprint arXiv:2212.06355*, 2022.

Nikos Vlassis, Ashok Chandrashekar, Fernando Amat Gil, and Nathan Kallus. Control variates for slate off-policy evaluation. *Advances in Neural Information Processing Systems*, 34, 2021.

Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudık. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, pp. 3589–3597. PMLR, 2017.

A Extended Discussion on Related Work

This section discusses the related work more comprehensively than Section 3.

Off-Policy Evaluation and Learning. Off-Policy Evaluation (OPE) and Learning (OPL) (Dudík et al., 2014; Wang et al., 2017; Liu et al., 2018; Farajtabar et al., 2018; Su et al., 2019; 2020; Kallus et al., 2021; Metelli et al., 2021; Saito & Joachims, 2022; Saito et al., 2023) aim to evaluate and optimize decision-making policies using only historical logged data, without the need for direct testing in an online environment. Various methods and estimators have been developed in OPE to control the bias and variance of the estimation of the performance of new policies (Su et al., 2020; 2019; Swaminathan & Joachims, 2015b;c; Dudík et al., 2014; Farajtabar et al., 2018; Kallus et al., 2021; Kiyohara et al., 2022; 2023; Metelli et al., 2021; Saito & Joachims, 2021; Wang et al., 2017; Kiyohara et al., 2024). These advancements in OPE methodology have led to highly accurate estimation of the policy value in the standard setting without new actions (Uehara et al., 2022; Saito et al., 2021b;a). In OPL, these estimation techniques enable accurate estimation of the policy gradient and enhance the resulting effectiveness of policy learning from only logged data (Saito et al., 2024b). However, most existing methods in OPL rely crucially on the full support condition (Condition 2.1) and focus solely on selecting the most effective actions within those already present in the logged data (Sachdeva et al., 2020). Therefore, they cannot deal with new actions that are potentially available in many decision-making problems and severely violate full support. In this paper, we propose a novel formulation and algorithm that leverage action features to address OPL with new actions. In particular, our algorithm can identify and choose effective new actions without sacrificing the overall policy performance, while existing methods cannot handle new actions even with the availability of the action features.

OPE for Slate Bandits. In this study, we address the challenge of incorporating new actions by considering OPL with action features. Specifically, we tackle realistic scenarios where each action is characterized by multi-dimensional action features, and a single reward is observed for each action. By mapping actions to slates and features of an action to slots of a slate, we can establish a connection between our problem with action features and OPE for slate bandits, initially explored by (Swaminathan et al., 2017).

Slate bandits consider selecting a combinatorial action called *slate* composed of multiple sub-actions. For example, in medical treatment scenarios, one may aim to select the optimal combination of drug doses to improve patient outcomes. While these systems often have access to extensive logged data, an accurate OPE remains challenging due to the exponential increase in variance associated with slate-wise importance weighting (Swaminathan et al., 2017; Su et al., 2020; Kiyohara et al., 2024). To address this challenge, several versions of *PseudoInverse* (PI) estimators have been proposed (Su et al., 2020; Swaminathan et al., 2017; Vlassis et al., 2021). They use slot-wise importance weights to relax the support condition and reduce the variance of typical estimators such as IPS. They have also been proven to enable unbiased OPE under the *linearity* assumption on the q-function. Linearity requires that the q-function be linearly decomposable, ignoring interaction effects among different slots. While PI often outperforms IPS under linear reward structures, it introduces significant bias when the linearity assumption does not hold (Kiyohara et al., 2024; Shimizu et al., 2024).

To effectively handle new actions in OPL by leveraging action features, we extend the idea of PI to estimate the policy gradient by relaxing the linearity assumption on the reward. This relaxation is crucial for dealing with the real-world non-linearities in reward functions during policy learning. It should be noted that the key contributions of our work are substantially different from those of existing studies for slate bandits (Swaminathan et al., 2017; Kiyohara et al., 2024; Vlassis et al., 2021; Chaudhari et al., 2024). We focus on the problem of policy *learning with new actions*. In contrast, the existing literature around slate OPE addresses only the problem of policy *evaluation without considering new actions*.

OPL with Deficient Actions. Another setting similar to new actions is OPL with deficient actions (Sachdeva et al., 2020; Felicioni et al., 2022), which refer to the actions that have zero probability of being selected under the logging policy for *certain* contexts x. Rigorously, the set of deficient actions is written as $\{a \in \mathcal{A} \mid \exists x \in \mathcal{X}, \pi_0(a \mid x) = 0\}$. Deficient action is a broader concept that encompasses the new actions we consider as a special case. Specifically, a deficient action a may still be observed for some contexts

x under the logging policy π_0, whereas new actions are never observed for any context x. This distinction becomes significant when estimating the q-function, $\hat{q}(x, a)$. The existence of deficient actions still allows for learning the q-function $q(x, a)$ because we observe every action somewhere in the logged data. However, with the presence of new actions, it becomes extremely challenging to estimate the q-function because we do not observe even a single data point for such actions. Due to this distinction, existing OPL approaches to deal with deficient actions (Sachdeva et al., 2020; Felicioni et al., 2022) cannot handle new actions. This motivates us to develop new methods specifically dealing with new actions without sacrificing the overall quality of policy learning.

B Omitted Proofs

Here, we provide the derivations and proofs that are omitted in the main text.

C Proof of Unbiasedness of LCPI

We provide a detailed proof that LCPI is unbiased under Conditions 4.4 and 4.3, mirroring the structure of Proposition 1 in (Swaminathan et al., 2017) while incorporating our own notation. We begin with a key lemma and then use it to establish the final result. To prove this, let us define the following action sets:

$$\mathcal{A}_{\pi_0}(x) := \{a \in \mathcal{A} \mid \pi_0(a|x) > 0\},$$
$$\mathcal{A}_{\mathrm{LCS}}(x) := \{a \in \mathcal{A} \mid \pi_0(f_{1:s}(a)|x) > 0, \pi_0(f_l(a)|x) > 0, \forall l \in [1, \ldots, d]\}.$$

Lemma C.1. *If Condition 4.4 holds and $a \in \mathcal{A}_{LCS}(x)$, then $q(x, a) = \mathbb{I}_a^T \Gamma_{\pi_0,x}^{\dagger} \theta_{\pi_0,x}$, where $\theta_{\pi_0,x} := \mathbb{E}_{\pi_0(a|x)} [\mathbb{I}_a q(x, a)]$.*

Proof. By Condition 4.4, there is a fixed vector ϕ_x such that for each action a,

$$q(x, a) = \mathbb{I}_a^T \phi_x.$$

Define a binary matrix $M \in \{0, 1\}^{|\mathcal{A}_{\mathrm{LCS}}(x)| \times (dm + m^s)}$, whose rows are the row vectors $\mathbb{I}_a^T$ for each $a \in \mathcal{A}_{\mathrm{LCS}}(x)$. The vector $M\phi_x$ enumerates $q(x, a)$ for all $a \in \mathcal{A}_{\mathrm{LCS}}(x)$. Let $\mathrm{Null}(M)$ denote the null space of M, and let Π denote the orthogonal projection onto $\mathrm{Null}(M)$. Define $\phi_x^\star := (I - \Pi)\phi_x$. Since $M\phi_x = M\phi_x^\star$, it follows that

$$q(x, a) = \mathbb{I}_a^T \phi_x = \mathbb{I}_a^T \phi_x^\star \quad \text{for all } a \in \mathcal{A}_{\mathrm{LCS}}(x).$$

From the definition in Section 4.2, we have

$$\theta_{\pi_0,x} = \mathbb{E}_{\pi_0(a|x)} [\mathbb{I}_a q(x, a)] = \mathbb{E}_{\pi_0(a|x)} \left[\mathbb{I}_a \mathbb{I}_a^T \phi_x\right] = \Gamma_{\pi_0,x} \phi_x. \tag{9}$$

Observe that

$$\begin{aligned}
\mathrm{Null}(\Gamma_{\pi_0,x}) &= \left\{v \mid v^T \Gamma_{\pi_0,x} v = 0\right\} \\
&= \left\{v \mid \mathbb{E}_{\pi_0(a|x)} \left[v^T \mathbb{I}_a \mathbb{I}_a^T v\right] = 0\right\} \\
&= \left\{v \mid \mathbb{I}_a^T v = 0, \forall a \in \mathcal{A}_{\pi_0}(x)\right\} \\
&= \left\{v \mid \mathbb{I}_a^T v = 0, \forall a \in \mathcal{A}_{\mathrm{LCS}}(x)\right\} \\
&= \mathrm{Null}(M),
\end{aligned}$$

where the first step follows from the positive semi-definiteness of $\Gamma_{\pi_0,x}$. Since $\mathrm{Null}(\Gamma_{\pi_0,x}) = \mathrm{Null}(M)$, it follows from Eq. 9 that

$$\theta_{\pi_0,x} = \Gamma_{\pi_0,x} \phi_x^\star.$$

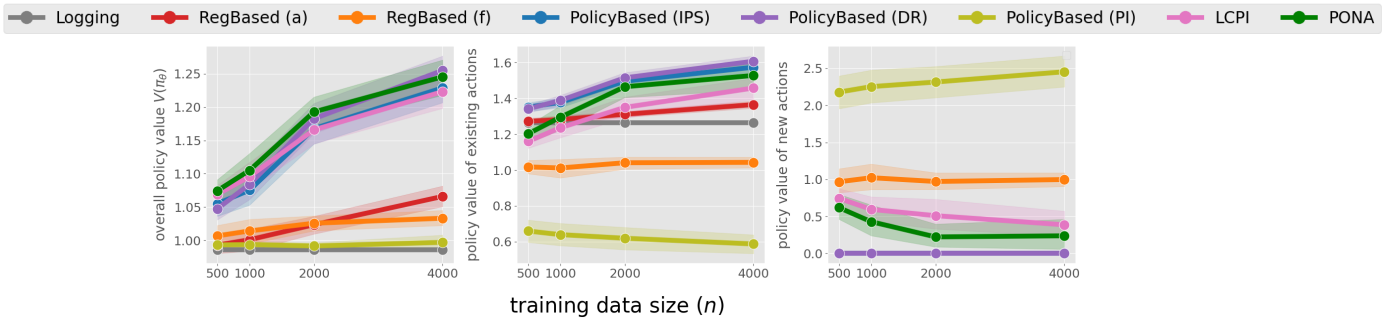

Figure 9: Comparisons of the overall policy value, policy value per existing action, and policy value per new action with varying training data sizes (n) on the KuaiRec dataset. Note that the metrics are normalized by those of the uniform random policy.

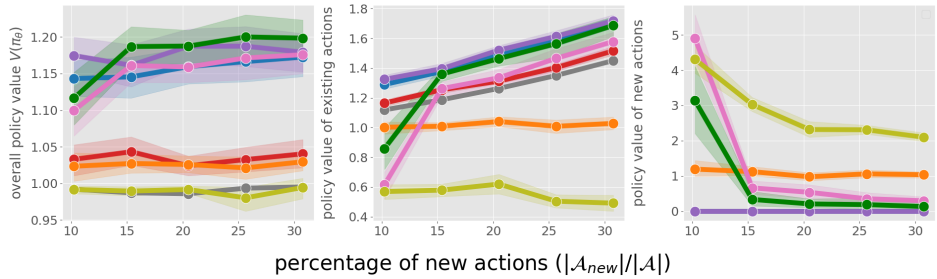

Figure 10: Comparisons of the overall policy value, policy value per existing action, and policy value per new action with varying percentages of new actions ($|\mathcal{A}_{new}|/|\mathcal{A}|$) on the KuaiRec dataset. Note that the metrics are normalized by those of the uniform random policy.

Moreover, from the definition of $\phi_x^\star$, we know that $\phi_x^\star \perp \mathrm{Null}(\Gamma_{\pi_0,x})$. Thus, by the definition of the pseudoinverse,

$$\Gamma_{\pi_0,x}^\dagger \theta_{\pi_0,x} = \phi_x^\star.$$

Putting this together, for $a \in \mathcal{A}_{\mathrm{LCS}}(x)$:

$$q(x,a) = \mathbb{I}_a^T \phi_x^\star = \mathbb{I}_a^T \Gamma_{\pi_0,x}^\dagger \theta_{\pi_0,x}.$$

This completes the proof of Lemma C.1. □

We now use the above Lemma to prove the unbiasedness of LCPI.

Proof.

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\left[\nabla_\theta \widehat{V}_{\mathrm{LCPI}}(\pi_\theta;\mathcal{D})\right] &= \mathbb{E}_{\mathcal{D}}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{a\in\mathcal{A}} \pi_\theta(a|x_i)\nabla_\theta \log \pi_\theta(a|x_i)\mathbb{I}_a^T\right)\Gamma_{\pi_0,x_i}^\dagger \mathbb{I}_{a_i} r_i\right] \\
&= \mathbb{E}_{p(x)\pi_0(a'|x)p(r|x,a')}\left[\left(\sum_{a\in\mathcal{A}} \pi_\theta(a|x)\nabla_\theta \log \pi_\theta(a|x)\mathbb{I}_a^T\right)\Gamma_{\pi_0,x}^\dagger \mathbb{I}_{a'} r\right] \\
&= \mathbb{E}_{p(x)}\left[\left(\sum_{a\in\mathcal{A}} \pi_\theta(a|x)\nabla_\theta \log \pi_\theta(a|x)\mathbb{I}_a^T\right)\Gamma_{\pi_0,x}^\dagger \mathbb{E}_{\pi_0(a'|x)}\left[\mathbb{I}_{a'} q(x,a')\right]\right] \\
&= \mathbb{E}_{p(x)}\left[\left(\sum_{a\in\mathcal{A}} \pi_\theta(a|x)\nabla_\theta \log \pi_\theta(a|x)\mathbb{I}_a^T\right)\Gamma_{\pi_0,x}^\dagger \theta_{\pi_0,x}\right]
\end{aligned}$$

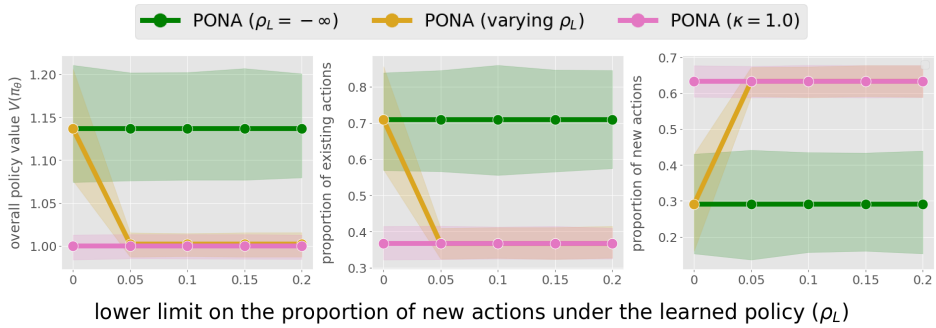

Figure 11: Comparisons of the overall policy value, proportion of existing actions, and proportion of new actions under the learned policy with varying lower limits on the proportion of new actions (ρ_L) on the KuaiRec dataset.

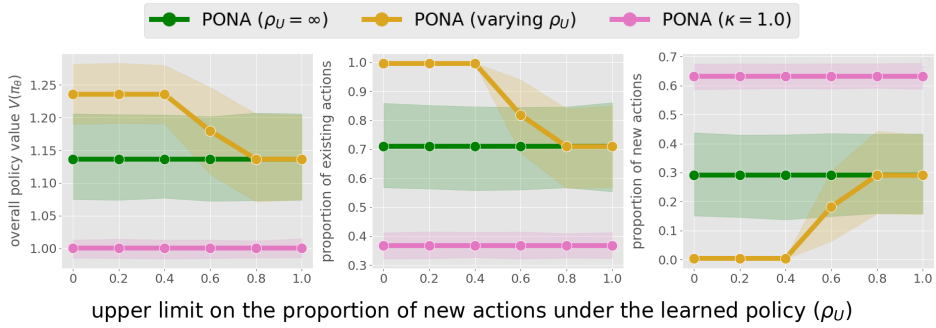

Figure 12: Comparisons of the overall policy value, proportion of existing actions, and proportion of new actions under the learned policy with varying upper limits on the proportion of new actions (ρ_U) on the KuaiRec dataset.

$$
\begin{aligned}
&= \mathbb{E}_{p(x)} \left[\sum_{a \in \mathcal{A}} \pi_\theta(a|x) \nabla_\theta \log \pi_\theta(a|x) q(x,a) \right] \quad \text{Condition 4.3, Lemma C.1} \\
&= \nabla_\theta V(\pi_\theta).
\end{aligned}
$$

□

D Experiment Details and Additional Results

This section describes the detailed experiment settings and reports additional results.

Detailed Setup. We describe the real-world experiment settings for KuaiRec (Gao et al., 2022) in detail. We use user information as the context x and reduce the feature dimensions using PCA implemented in scikit-learn (Pedregosa et al., 2011). We define the reward as the user watch ratio, clipped at the 99th percentile and normalized.

There are four types of action features: video tags, first-level categories, second-level categories, and third-level categories. For the combination features used by LCPI and PONA, which range from 1 to s dimensions, we use three types: video tags, first-level categories, and third-level categories. For existing actions, we prioritize those with overlapping features in the 1- to s-dimensional combination features to ensure that Local Combination Support is satisfied as much as possible.

Additional Results. This section presents additional results from real-world experiments. We report the results for varying training data sizes (n) in Figure 9, percentages of new actions ($|\mathcal{A}_{new}|/|\mathcal{A}|$) in Figure 10,

varying lower limits on the proportion of new actions (ρ_L) in Figure 11, and varying upper limits on the proportion of new actions (ρ_U) in Figure 12.

Limitations and Broader Impact. Our method relies on structural assumptions on action features and their local combinations. When these assumptions are severely violated, LCPI may overestimate the value of unsupported new actions, which could lead to suboptimal or potentially harmful decisions in high-stakes applications. Therefore, in practical deployments, we recommend restricting the target policy to actions satisfying feature-span support, monitoring the proportion and outcomes of newly selected actions, and using conservative choices of the mixing parameter when the validity of the assumptions is uncertain. This work is methodological and does not introduce a new dataset containing sensitive personal information.