Clustering Context in Off-Policy Evaluation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Off-policy evaluation can leverage logged data to estimate the effectiveness of new policies in e-commerce, search engines, media streaming services, or automatic diagnostic tools in healthcare. However, the performance of baseline off-policy estimators like IPS deteriorates when the logging policy significantly differs from the evaluation policy. Recent work proposes sharing information across similar actions to mitigate this problem. In this work, we propose an alternative estimator that shares information across similar contexts using clustering. We study the theoretical properties of the proposed estimator, characterizing its bias and variance under different conditions. We also compare the performance of the proposed estimator and existing approaches in various synthetic problems, as well as a real-world recommendation dataset. Our experimental results confirm that clustering contexts improves estimation accuracy, especially in deficient information settings.[1]

1 Introduction

The contextual bandit process models many real-world problems across industry and research, including healthcare, finance, and recommendation systems (Bouneffouf et al., 2020). In this setting, an agent observes a *context*, chooses an action according to a *policy*, and observes a *reward*. *Off-policy evaluation* (OPE) methods aim to estimate the effectiveness of a policy without empirically testing it, which can be particularly useful when A/B tests are costly, or if there is an inherent risk associated with poor policy performance, as is often the case in healthcare (Bastani & Bayati, 2019). Existing OPE methods can be broadly divided into parametric methods based on the *direct method* (DM), non-parametric methods based on *inverse propensity score* weighting (IPS, Horvitz & Thompson, 1952), and a combination of the two, such as the *doubly robust* method (DR, Dudík et al., 2011). When every action with non-zero probability under the evaluation policy also has a non-zero probability under the logging policy, IPS is unbiased. This condition is rarely satisfied in real-world problems, however, so IPS is typically biased in practice, especially for actions that violate the condition, or have close-to-zero probabilities in the logging policy (Sachdeva et al., 2020; Dudík et al., 2011; Saito & Joachims, 2022).

Recently proposed *Marginalized Inverse Propensity Score* estimator (MIPS, Saito & Joachims, 2022) improves upon IPS in large action spaces by pooling information across *action embeddings*. At the same time, MIPS suffers from the same problem as IPS for contexts in which a significant proportion of actions have low probability under the logging policy. In this case, MIPS lacks information about the actions to accurately estimate the importance weights, resulting in additional bias. In our work, we hypothesize that closeness at the context level should translate into similar behaviour for actions and rewards (for example, two movies of the same franchise in a recommendation system). Based on this hypothesis, we propose an estimator that *clusters* the context space, and pools information across

[1] The code for reproducing our experimental implementation is available at https://www.github.com/anonoymous/anonymous-repo

all the contexts within a cluster. Informally, the proposed method solves the problem of deficient action information for a particular context by leveraging the information from all other contexts within the same cluster.

We define and analyze the theoretical bandit setup with context clusters in Section 3, which leads to the formal derivation of the CHIPS estimator, for which we analyze bias and variance. In section 4, we compare the estimator's performance to the baselines on several synthetic and real-world datasets, verifying the theoretical findings, and demonstrating its effectiveness. Finally section 5 explores future lines of work and CHIPS' limitations.

2 Background on Off-Policy Evaluation and Related Work

The off-policy evaluation problem (OPE) is usually framed inside the general contextual bandit setup. Given an agent, determined by the policy $\pi : \mathcal{X} \times \mathcal{A} \to [0, 1]$, the bandit's data generation process is defined as iterative logging of the agent's behavior when presented with different contexts. In each iteration, a context $x \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$ is drawn i.i.d. from an unknown probability distribution $p(x)$ over the context space, an action $a \sim \pi(a|x)$ is selected from a finite action space $\mathcal{A}$, and a bounded reward $r \in [0, R_{\max}]$ is observed as a sample from an unknown conditional distribution $p(r|a, x)$. The off-policy evaluation problem has been extensively studied from both a theoretical (McNellis et al., 2017; Saito et al., 2021; Dumitrascu et al., 2018; Irpan et al., 2019; Wang et al., 2017) and a practical point of view given its applications in fields such as recommendation systems (Li et al., 2011; Bendada et al., 2020; Saito et al., 2020) or healthcare (Varatharajah & Berry, 2022).

We measure the performance of a policy π through its *value*, that we define as:

$$V(\pi) := \mathbb{E}_{p(x)\pi(a|x)p(r|a,x)}[r] = \mathbb{E}_{p(x)\pi(a|x)}[q(a, x)] \tag{1}$$

Here $q(a, x) = \mathbb{E}_{p(r|a,x)}[r]$ denotes the conditional expected reward given an action a and a context x.

In practice, we are interested in finding a policy maximizing the expected reward observed in the bandit process. A vital part of this process is the off-policy evaluation problem, in which we estimate the value of a policy π given a dataset $\mathcal{D} := \{(x_i, a_i, r_i)\}_{i=1}^N$ collected under a logging policy π_0 (i.e. $\mathcal{D} \sim \prod_{i=1}^N p(x)\pi_0(a|x)p(r|a, x)$). We use the mean squared error (MSE) to quantify how well the estimate $\hat{V}(\pi)$ approximates the real policy value $V(\pi)$:

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

A wide variety of approaches have been proposed in the literature to estimate $V(\pi)$. From them, three can be distinguished for being commonly used as starting points for developing new estimators. The first one is the Direct Method (DM), which tries to estimate $q(a, x)$ directly from Equation (1):

$$\hat{V}_{\text{DM}}(\pi; \mathcal{D}, \hat{q}) = \frac{1}{N} \sum_{i=1}^{N} \sum_{a \in \mathcal{A}} \hat{q}(a, x_i)$$

The bias of DM depends on the accuracy of the $\hat{q}(a, x) \approx q(a, x)$ approximation, but the variance is usually lower than in other approaches. Supervised learning in the DM's approach can be particularly useful when generalization of an agent's behaviour is needed due to limited information in the logging data (Sachdeva et al., 2020). However, when the reward function has a high variance, or the representation capacity is limited for the context-action pairs in the evaluation policy domain, $\hat{q}(a, x)$ could fail to accurately approximate $q(a, x)$ (Farajtabar et al., 2018; Beygelzimer & Langford, 2009; Kallus & Uehara, 2019). This problem, known as *reward misspecification*, can be quite difficult to detect in real-world examples (Farajtabar et al., 2018; Voloshin et al., 2021), and is the reason why DM is generally regarded as a highly biased estimator.

The second base approach is Inverse Propensity Scoring (IPS, Horvitz & Thompson, 1952), which approximates the policy value by reweighting the rewards to correct the shift in action probabilities between the logging and evaluation policies:

$$\hat{V}_{\text{IPS}}(\pi; \mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} \frac{\pi(a_i|x_i)}{\pi_0(a_i|x_i)} r_i = \frac{1}{N} \sum_{i=1}^{N} w(a_i, x_i) r_i$$

As per this definition, the context-action pairs selected by π in which $\pi_0(a|x) = 0$ could be problematic, which motivates the following assumption:

Assumption 2.1. (*Common Support*) Given an evaluation policy π and a logging policy π_0, the latest has common support for π if

$$\pi_0(a|x) > 0 \quad \forall a \in \mathcal{A}, x \in \mathcal{X} \ : \ \pi(a|x) > 0$$

The IPS estimator is unbiased under Assumption 2.1. However, even when assumption 2.1 holds, IPS can present excessive variance due to the weights $w(a_i, x_i)$ taking larger values (Dudík et al., 2011; Saito & Joachims, 2022). This case is especially notable when π_0 and π are significantly different or when trying to achieve universal support ($\pi_0(a|x) > 0 \ \forall a \in \mathcal{A}, x \in \mathcal{X}$) in large action spaces (Saito & Joachims, 2022; Peng et al., 2023; Saito et al., 2021). Controlling the scaling of the propensity scores has motivated many approaches based on IPS, using techniques such as weight clipping (Su et al., 2020a,b; Swaminathan & Joachims, 2015a) and self normalization (Swaminathan & Joachims, 2015b; Kuzborskij et al., 2020). The Doubly Robust (DR) estimator combines DM and IPS, aiming to obtain a low-bias, low-variance estimate:

$$V_{\text{DR}}(\pi; \mathcal{D}, \hat{q}) := V_{\text{DM}}(\pi; \hat{q}) + \frac{1}{N}\sum_{i=1}^{N} w(a_i, x_i)(r_i - \hat{q}(a_i, x_i))$$

The DR estimator has been the cornerstone of multiple approaches that modify the base estimator to address problems such as low overlap between π and π_0 (Wang et al., 2017; Metelli et al., 2021; Zhan et al., 2021; Guo et al., 2024), reward misspecification (Farajtabar et al., 2018), and limited samples in logging data (Su et al., 2020a; Felicioni et al., 2022). Unfortunately, the DR estimator can still inherit the large variance problem from IPS, for example, when dealing with large action spaces (Saito et al., 2023; Saito & Joachims, 2022; Shimizu & Forastiere, 2023; Sachdeva et al., 2023; Taufiq et al., 2023). The problem of dealing with large action spaces was recently studied, resulting in the *Marginalized Inverse Propensity Scoring* (MIPS) (Saito & Joachims, 2022) estimator, in which the authors pool information between similar actions given some embedding representation $e \in \mathcal{E} \subset \mathbb{R}_e^d$ of them to address deficient actions in the logging policy. For this purpose, they introduce an IPS-based estimator marginalizing the probability over the action space:

$$\hat{V}_{\text{MIPS}}(\pi; \mathcal{D}) := \frac{1}{n}\sum_{i=1}^{n} \frac{p\left(e_i \mid x_i, \pi\right)}{p\left(e_i \mid x_i, \pi_0\right)} r_i = \frac{1}{n}\sum_{i=1}^{n} w\left(x_i, e_i\right) r_i, \tag{2}$$

Where $p(e \mid x, \pi) := \sum_{a \in \mathcal{A}} p(e \mid x, a)\pi(a \mid x)$.

The idea of estimating deficient items' behaviour by *closely* observed ones inspired new approaches, like partitioning the action space in clusters (Peng et al., 2023; Saito et al., 2023), or an adaptive method for ranking policies by optimizing user classification into given behavioural models and estimating independently for each group (Kiyohara et al., 2023). The MR estimator (Taufiq et al., 2023) diverged from the action space transformations and proposed marginalization over the rewards density through a regression estimate of the importance weights:

$$\hat{V}_{\text{MR}}(\pi; \mathcal{D}) := \frac{1}{n}\sum_{i=1}^{n} w\left(r_i\right) r_i \tag{3}$$

Where $w(r)$ is defined as:

$$\begin{gathered} w(r) := f_{\phi^*}(r) := \operatorname{argmin} \mathbb{E}_{\phi}\left[\left(w(a,x) - f_{\phi}(r)\right)^2\right] \\ f_{\phi} \in \{f_{\phi} : \mathbb{R} \to \mathbb{R} \mid \phi \in \Phi\} \end{gathered} \tag{4}$$

Motivated by these approaches, as well as the fact that estimating from *similar* actions or make a regression over rewards could prove challenging if a significant proportion of these actions are missing for a given context, we propose the *Context-Huddling Inverse Propensity Score* (CHIPS) estimator that we introduce in the next section.

3 The CHIPS estimator

The CHIPS estimator is based on the idea of partitioning the context space into clusters to extrapolate the behaviour of an agent when presented with a previously unseen or underrepresented context x.

The assumption needed for this approximation to the OPE problem is that, given a policy, all contexts belonging to a cluster c should have a similar probability of observing an action a and will observe similar rewards when that action is chosen. Formally, we will consider a finite partition of the context space as the cluster space $\mathcal{C} := \{\mathcal{C}_i\}_{i=1}^{K}$ with $\mathcal{C}_i \subset \mathcal{X}$ and $c_i \cap \mathcal{C}_j = \varnothing$. We assume that we are given a $c \in \mathcal{C}$ for each context $x \in \mathcal{X}$, where we assume that c is drawn i.i.d from an unknown distribution $p(c|x)$. Thus, given a policy π, we can compute its value by refining Equation (1):

$$V(\pi) := \mathbb{E}_{p(x)p(c|x)\pi(a|x)p(r|a,c,x)}\left[r\right] = \mathbb{E}_{p(x)p(c|x)\pi(a|x)}\left[q(a,c,x)\right] \tag{5}$$

Where we denote $q(a,c,x) := \mathbb{E}_{p(r|a,c,x)}\left[r\right]$ and it is important to note that $\mathbb{E}_{p(c|x)\pi(a|x)}\left[q(a,c,x)\right] = \mathbb{E}_{\pi(a|x)}\left[q(a,x)\right]$, and therefore the refinement is consistent with Equation (1). Similar to the common support condition in IPS, we formulate the following property as the equivalent for the CHIPS estimator of Assumption 2.1.

Assumption 3.1. (*Common Cluster Support*) Given an evaluation policy π and a logging policy π_0, the latest has common cluster support for π if

$$p(a|c,\pi_0) > 0 \quad \forall a \in \mathcal{A}, c \in \mathcal{C} \;:\; p(a|c,\pi) > 0$$

Where we denote

$$p(a|c,\pi) = \int_{\mathcal{X}} \pi(a|x)p(x|c)dx$$

Assumption 3.1 is weaker than Assumption 2.1 since for a given triplet $(x,c,a) \in \mathcal{X} \times \mathcal{C} \times \mathcal{A}$, the fact that $\pi_0(a|x) = 0, \pi(a|x) > 0$ does not ensure the same holds for every context within c. The idea of a homogeneous behaviour for every context inside a given cluster would make the CHIPS estimator circumvent the bias increase when Assumption 2.1 is not met for the IPS estimator (if Assumption 3.1 holds). Regarding the reward, this concept is formalized in the following assumption.

Assumption 3.2. (*Reward Homogeneity*) We say that we observe reward homogeneity if the context x does not affect on the reward r given some action a and some context c (i.e., $r \perp x \mid c, a$).

The reward homogeneity assumption eliminates the dependency of the context on the reward when provided with the cluster and the action. Note that complying with Assumption 3.2 implies $q(a,c,x) = q(a,c,y) = q(a,c)$, where $x, y \in \mathcal{X}$, which together with Assumption 3.1 gives an alternative expression for the policy value in the following proposition:

Proposition 3.3. *Given a policy π, if Assumptions 3.1 and 3.2 hold, then we have that*

$$V(\pi) := \mathbb{E}_{p(c)p(a|c,\pi)}\left[q(a,c)\right] \tag{6}$$

Please refer to Appendix A.1 for a complete proof.

Considering the similarity of Equation (6) with the original policy value definition (Equation (1)), Proposition 3.3 naturally motivates the analytical expression of the CHIPS estimator:

$$\hat{V}_{\text{CHIPS}}(\pi;\mathcal{D}) := \frac{1}{N}\sum_{i=1}^{N}\frac{p(a_i|c_i,\pi)}{p(a_i|c_i,\pi_0)}r_i = \frac{1}{N}\sum_{i=1}^{N} w(a_i,c_i)r_i$$

3.1 Theoretical Analysis

First, we characterize the bias of the CHIPS estimator depending on the compliance with Assumptions 3.1 and 3.2.

Proposition 3.4. *Under the Common Cluster Assumption (3.1) and the Cluster Homogeneity Assumption (3.2), the CHIPS estimator is unbiased for any given policy π:*

$$\mathbb{E}_{\mathcal{D}}\left[\hat{V}_{CHIPS}(\pi;\mathcal{D})\right] = V(\pi)$$

Please refer to Appendix A.2 for a complete proof.

We note here that Proposition 3.4 implies that even when the Common Support Assumption (2.1) fails to ensure the unbiasedness of the IPS estimator, the CHIPS estimator can still use the more permissive Common Cluster Support (3.1), and the Reward Homogeneity (3.2) Assumption to ensure an unbiased estimate. Although Assumption 3.2 guarantees homogeneity at the reward level, a *completely* homogeneous behaviour would also eliminate the context dependency at the action level, implying a deterministic policy given cluster, i.e. $p(a|c,\pi) = \pi(a|x)\ \forall x \in c$. Both homogeneity conditions present a desirable scenario for the CHIPS estimator; however, they rarely occur when working in real-world data environments, which motivate the following assumption as a relaxation of the action-context independence:

Assumption 3.5. (δ-*Homogeneity*) Given a policy π, we say that the policy presents δ-homogeneity if there exist $\delta_\pi^- \leq 1$ and $\delta_\pi^+ \geq 1$ such that:

$$\delta_\pi^- \leq \frac{\pi(a|x)}{p(a|c,\pi)} \leq \delta_\pi^+ \quad \forall (x,c,a) \in \mathcal{D}$$

It is worth noting that if $p(a|c,\pi) \neq 0\ \forall (x,c,a) \in \mathcal{D}$ then it is always possible to find δ_π^-, δ_π^+ satisfying δ-Homogeneity. The following proposition gives an upper bound for the bias of the CHIPS estimator when Assumption 3.2 cannot be ensured:

Proposition 3.6. *Given the logging data $\{(x_i, a_i, r_i)\}_{i=1}^N$ observed under some logging policy π_0, and an evaluation policy π if the latest has common cluster support over the earliest, then we have that*

$$\left|Bias(\hat{V}_{CHIPS}(\pi;\mathcal{D}))\right| \leq \left|\mathbb{E}_{p(c)p(x|c)p(a|c,\pi)}\left[q(a,c,x)\cdot\Delta\right]\right|$$

Where by Assumption 3.5 we have bounds (δ_π^-, δ_π^+) for π, ($\delta_{\pi_0}^-$, $\delta_{\pi_0}^+$) for π_0, and we denote $\Delta = \max\{\delta_\pi^+, \delta_{\pi_0}^+\} - \min\{\delta_\pi^-, \delta_{\pi_0}^-\}$. Please refer to Appendix A.3 for a complete proof.

Proposition 3.6 formalizes the intuition on how the bias of the estimator under Assumption 3.1 depends on the extent to which the contexts inside a cluster behave homogeneously under a given policy. Formally, the gap $\delta_\pi^+ - \delta_\pi^-$ determines how close the CHIPS is to being unbiased, being the case $\delta_\pi^- = \delta_\pi^+ = 1$ the perfect scenario. In this case, we have that $\pi(a|x) = p(a|c,\pi)$, which means that the weights in IPS $w(a,x) = w(a,c)$, and we could in theory substitute any context for any other within the same cluster for calculations, mitigating the problems that arise when Assumption 2.1 does not hold. Additionally, we can also provide an expression for the difference in mean squared error with respect to IPS in the same conditions as Proposition 3.6:

Proposition 3.7. *Under the same conditions as in Proposition 3.6, the difference in mean squared error between CHIPS and MIPS can be expressed as*

$$\mathrm{MSE}\left(\hat{V}_{IPS}(\pi)\right) - \mathrm{MSE}\left(\hat{V}_{CHIPS}\right) = \mathbb{V}_D\left[\hat{V}_{IPS}(\pi)\right] - \mathbb{V}_D\left[V_{CHIPS}(\pi;D)\right] - \mathrm{Bias}\left(\hat{V}_{CHIPS}(\pi)\right)^2$$

Please refer to Appendix A.4 for a complete proof.

It is also worth studying the bias of the CHIPS estimator when the Common Cluster Support assumption does not hold, while the Assumption 3.2 holds. For this purpose, we acknowledge that the bias of the IPS estimator when Assumption 2.1 is not met can be given in terms of the actions violating such assumption (Sachdeva et al., 2020):

$$\left|\mathrm{Bias}(\hat{V}_{\mathrm{IPS}}(\pi;\mathcal{D}))\right| = \mathbb{E}_{p(x)}\left[\sum_{\mathcal{U}(x,\pi_0)} \pi(a|x)q(a,c,x)\right]$$

Where $\mathcal{U}(x,\pi_0) := \{a \in \mathcal{A} \mid \pi_0(a,x) = 0\}$ are known as the *deficient* actions. Following a similar approach we introduce the following proposition:

Proposition 3.8. *Given the logging policy π_0 and some evaluation policy π, the absolute bias of the CHIPS estimator when Assumption 3.2 holds can be expressed as*

$$\left|Bias(\hat{V}_{CHIPS}(\pi;\mathcal{D}))\right| = \mathbb{E}_{p(c)}\left[\sum_{\mathcal{U}(c,\pi_0)} p(a|\pi,c)q(a,c)\right]$$

Where $\mathcal{U}(c,\pi_0) := \{a \in \mathcal{A} \mid p(a|\pi_0,c) = 0\}$. Please refer to Appendix A.5 for a complete proof.

As a consequence of Proposition 3.8 we can find an analytical expression for the bias reduction of the CHIPS estimator with respect to IPS:

Corollary 3.9. *Under the conditions of Proposition 3.8, we have that*

$$\left|Bias(\hat{V}_{IPS}(\pi;\mathcal{D}))\right| - \left|Bias(\hat{V}_{CHIPS}(\pi;\mathcal{D}))\right| = \mathbb{E}_{p(c)}\left[\sum_{\mathcal{U}(c,x,\pi_0)\setminus\mathcal{U}(c,\pi_0)} p(a|\pi,c)q(a,c)\right]$$

Where $\mathcal{U}(c,x,\pi_0) := \{a \in \mathcal{A} \mid \pi_0(a|x) = 0\}$. Please refer to Appendix A.5 for a complete proof.

Note that in this case, the CHIPS' reduction in absolute bias depends directly on the number of actions that violate Assumption 2.1, but still comply with Assumption 3.2. Thus, the greater the number of deficient actions by Common Support condition covered by the Common Cluster Support, the more significant the bias reduction with respect to IPS. In this conditions, its also interesting to study the difference in bias with respect to the other two transformation-based methods (MR and MIPS), a result given by the next proposition:

Proposition 3.10. *Let f_{ϕ^*} be defined as in Equation (4) with $f_{\phi^*} = w(a,x) + \epsilon$ for some $\epsilon \in \mathbb{R}$ and $e \in \mathcal{E}$ give action embeddings. Under the conditions of the Proposition 3.8, we have that:*

$$\begin{aligned}
&\left|\text{Bias}\left(\hat{V}_{MR};\mathcal{D}\right)\right| - \left|\text{Bias}\left(\hat{V}_{CHIPS};\mathcal{D}\right)\right| \\
&= -\mathbb{E}_{p(c)}\left[\sum_{a\in(\mathcal{U}(x,c,\pi_0)\setminus\mathcal{U}(c,\pi_0))c} q(a,c)p(a \mid \pi,c)\right] + \epsilon\mathbb{E}_{p(c)}\left[\sum_{a\in\mathcal{A}} q(a,c)p\left(a \mid \pi_0,c\right)\right] \\
&\left|\text{Bias}\left(\hat{V}_{MIPS};\mathcal{D}\right)\right| - \left|\text{Bias}\left(\hat{V}_{CHIPS};\mathcal{D}\right)\right| \\
&= \mathbb{E}_{p(x)}\left[\sum_{e\in\mathcal{U}(e,\pi_0)} p(e \mid x,\pi)q(x,e)\right] - \mathbb{E}_{p(c)}\left[\sum_{a\in\mathcal{U}(c,\pi_0)} p(a \mid c,\pi)q(a,c)\right]
\end{aligned}$$

Please refer to Appendix A.6 for a complete proof.

Note that in the CHIPS case

When studying homogeneity at an action level, we have focused on the probability of observing an action for a particular context x within a cluster c (i.e., $\pi(a|x)$). Conversely, we can also study the *predictability* of a context given an action and a cluster under a policy π, which we denote as $p(x|a,c) = \pi(x|a,c)$. Ideally, we would have that the conditional probability distribution of the context given the action and the cluster is uniform (i.e., $\pi(x_i|a,c) = \pi(x_j|a,c)\ \forall\ x_i, x_j \in c$). Predictability is used in the following proposition, that characterizes the relation between the reduction in variance of the CHIPS estimator with respect to IPS:

Proposition 3.11. *Given a logging policy π_0, under the Common Support Assumption (2.1) and the Reward Homogeneity Assumption (3.2) we have that*

$$N\left(\mathbb{V}_{\mathcal{D}}\left[\hat{V}_{IPS}(\pi;\mathcal{D})\right] - \mathbb{V}_{\mathcal{D}}\left[\hat{V}_{CHIPS}(\pi;\mathcal{D})\right]\right) = \mathbb{E}_{p(c)p(a|c,\pi_0)}\left[\mathbb{V}_{\pi_0(x|a,c)}\left[w^2(a,x)\right]\mathbb{E}_{p(r|a,c)}\left[r^2\right]\right]$$

Note that this quantity is always positive, implying that CHIPS always reduces the variance of IPS. Please refer to Appendix A.7 for a complete proof.

Proposition 3.11 implies that when Assumptions 2.1 and 3.2 are met, CHIPS' reduction in variance with respect to IPS corresponds with the total reduction in the mean squared error when trying to approximate the real policy value $V(\pi)$ (since both estimators are unbiased under these conditions). This mean squared error gap depends on two factors. First, we have $\mathbb{E}_{p(r|a,c)}\left[r^2\right]$ which depends on how noisy the rewards are given an action inside the same cluster (directly related to Assumption 3.2). Second, the variance of IPS weights conditioned to the predictability $p(x|a,c)$, which becomes larger depending on either $w(a,x)$ having a wide range (for example, when logging and evaluation policies differ considerably) or $\pi(x|a,c)$ being uninformative (context behaving homogeneously given the cluster and action). This suggests that the variance reduction in CHIPS is particularly noticeable in the cases in which IPS has high variance and the contexts behave similarly within a cluster.

Furthermore, if MIPS and CHIPS are in the same space (considering contexts $c \in \mathcal{C}$ as described and action embeddings $e \in \mathcal{E}$), Proposition 3.11 can be extended to show that CHIPS has less variance than MIPS:

Proposition 3.12. *In context-action-embedding joint space ($\mathcal{X} \to \mathcal{C} \to \mathcal{A} \to \mathcal{E} \to [0, R_{max}]$), if Assumptions 3.1 and 3.2 hold, as well as their MIPS counterparts (Common Embedding Support and No Direct Effect), then we have that*

$$\mathbb{V}_{\mathcal{D}}\left[\hat{V}_{IPS}(\pi)\right] \geq \mathbb{V}_{\mathcal{D}}\left[\hat{V}_{MIPS}(\pi)\right] \geq \mathbb{V}_{\mathcal{D}}\left[\hat{V}_{CHIPS}(\pi)\right] \geq 0$$

Please refer to Appendix A.8 for a complete proof.

3.2 Empirical Calculations

The alternative analytical expression for the policy value given in Equation 6 eliminates the dependency on the original definition of policy value and motivates the CHIPS estimator under assumptions 3.1 and 3.2. However, in practice, assessing if such conditions hold is complicated, particularly if we have limited logging data. To mitigate this problem and justify using CHIPS in real-world settings, we need to make an approximation to context-homogeneous behavior on both action and reward levels within a cluster. In practice, we have a clustering method $\xi : \mathcal{X} \to \mathcal{C}$, and we use the transformation:

$$\begin{aligned} \tau\colon (\mathcal{X}, \mathcal{A}, [0, R_{max}]) &\to (\mathcal{X}, \mathcal{C}, \mathcal{A}, [0, R_{max}]) \\ (x, a, r) &\mapsto (x, \xi(x), a, r). \end{aligned}$$

Given a policy π and a cluster c, we use the definition to estimate $p(a|c,\pi)$:

$$p(a|c,\pi) = \int_{\mathcal{X}} \pi(a|x)p(x|c)dx = \int_{x \in c} \pi(a|x)p(x|c) \approx \frac{1}{|\mathcal{D}_c|}\sum_{\mathcal{D}_c} \pi(a|x) \tag{7}$$

Here, we denote $\mathcal{D}_c = \{(x, \tilde{c}, a, r) \in \tau(\mathcal{D}) \ : \ \tilde{c} = c\}$. In Equation 7, we used that $p(x|c) = 0$ if $c \neq \xi(x)$. Since this equation is essentially $\mathbb{E}_{p(x|c)}\left[\pi(a|x)\right]$, we approximate this value by averaging $\pi(a|x)$ over all contexts inside the given cluster.

The second approximation needed involves the reward being independent of the context given the action and the cluster, i.e., $q(a,c,x) = q(a,c)$. Following a similar approach than in the previous case, for a particular (given) action a and cluster c, we observe that $q(a,c) = \mathbb{E}_{p(x|c)}\left[\pi(r|a,c,x)\right]$, which motivates the idea of an *average reward* per cluster. In our synthetic experiments, the reward is binary, therefore we will assume that the observations inside a cluster are observations in a Bernoulli process (i.e., $R_c \sim \text{Ber}(\theta)$) and estimate this average reward using two different approaches:

- **Maximum Likelihood (ML)** In which we just average the rewards observed within a cluster c for each action a as $\hat{r}_{\text{mean}}(a,c) = \frac{1}{|R_c|}\sum_{R_c} r_k$ with $R_c \coloneqq \{r_k : (x_k, c_k, a_k, r_k) \in \mathcal{D}_c\}$.
- **Maximum A Posteriori (MAP)**. In this setting, estimating the average reward is equivalent to estimating the most probable θ using a beta prior, where we obtain:

$$\hat{r}_{\text{bayes}}(\alpha, \hat{\beta}; c) = \frac{(\alpha - 1) + \sum_{R_c} r_k}{\alpha + \hat{\beta} + |R_c| - 2}$$

  Where we denote $\alpha, \hat{\beta}$ as the parameters of the prior Beta distribution. In our experiments, we use non-informative priors ($\alpha = \hat{\beta}$) Tuyl et al. (2008); Kerman (2011) and we explore the choosing of this parameter for arbitrary problems in Appendix D.4. Please refer to Appendix B for the complete derivations of the MAP and ML estimations.

4 Experiments

4.1 Synthetic dataset

We compare CHIPS with other baseline estimators (IPS, DM, DR, SNIPS (Swaminathan & Joachims, 2015b), DRoS (Su et al., 2020a), SNDR (Thomas & Brunskill, 2016), MR (Taufiq et al., 2023)) in estimating the evaluation policy value in a cluster-based synthetic dataset in which we can control the difficulty of the OPE problem. A description of all hyperparameters used for generation (e.g., a_{num}, c_{exp} ...) can be found in Appendix C. We start by generating cluster centers $\mathcal{C} \coloneqq \{c_k\}_{k=1}^m$ inside a d_x-dimensional ball $B(0, \text{c}_{\text{exp}}) \coloneqq \{x \in \mathbb{R}^{d_x} \ : \ ||x||^2 < \text{c}_{\text{exp}}\}$ using a variation of the Box-Muller transformation (Box & Muller, 1958):

$$c_k = \frac{\text{c}_{\text{exp}} \cdot {u_k}^{-dx} \cdot z_k}{||z_k||},$$

where $U \coloneqq \{u_k\}_{k=1}^m \sim U[0,1]$ and $Z \coloneqq \{z_k\}_{k=1}^m \sim \mathcal{N}(0, \mathbb{1}_{d_x})$. We sample $S \coloneqq \{s_k\}_{k=1}^m \sim U[0,1]$, and use the softmax transformation $\phi(\mathcal{S})$ to define $p(c_i) = \phi(\mathcal{S})_i$. Then, we sample cluster centers according to this distribution $w = \{w_i\}_{i=1}^{x_{\text{num}}} \sim \phi(\mathcal{S})$, and, for each center c_i, we uniformly sample points belonging to the n-ball centered on c_i, using the same variation of the Box-Muller transform that we used previously:

$$\mathcal{X}_i = (x_i^1, ..., x_i^{h_i}) \sim U[B(c_i, c_{\text{rad}})]$$

Note here that $h_i = \sum_{i=1}^{x_{\text{num}}} \mathbb{1}_{\{c_i = w_i\}}$. We define the context space as the union of these generated points $\mathcal{X} = \bigcup_{i=1}^{m} \mathcal{X}_i = \{x_i\}_{i=1}^{x_{\text{num}}}$. We sample $\mathcal{V} = \{v_i\}_{i=1}^{x_{\text{num}}} \sim \mathcal{N}(0,1)$ and define $p(x_i) = \phi(\mathcal{V})_i$ using the ϕ softmax transformation again. We then use these probabilities to sample the logging ($\mathcal{X}_{\text{log}}$) and evaluation ($\mathcal{X}_{\text{eval}}$) data, with $|\mathcal{X}_{\text{eval}}| = e_{len}$ and $|\mathcal{X}_{\text{log}}| = b_{len}$. To generate the policies, we sample $y_i = \{y_i^j\}_{j=1}^{a_{\text{num}}} \sim \mathcal{N}(0,1)$ for every cluster c_i (where a_{num} is the number of actions) and $z = \{z_k\}_{i=1}^{x_{\text{num}}} \sim \mathcal{N}(0,1)$ to define the policies for every context in cluster c_i as:

$$\pi(a_j|c_i, x_k) = \frac{e^{y_i^j + \sigma z_k}}{\sum_{m=1}^{a_{\text{num}}} e^{y_i^m + \sigma z_k}} \qquad \pi_0(a_j|c_i, x_k) = \frac{e^{\beta(y_i^j + \sigma z_k)}}{\sum_{m=1}^{a_{\text{num}}} e^{\beta(y_i^m + \sigma z_k)}}, \quad -1 \le \beta \le 1$$

Given a context x_k, both policies are determined by a term that depends on the cluster and the action (u_i^j), and a term that depends on the context itself (x_k). Here $0 \le \sigma \le 1$ controls how independent a policy is from the context and β how close the logging and evaluation policies are. For obtaining the actions, we sample $\mathcal{A}_{\text{log}} \sim \pi_0$ and $\mathcal{A}_{\text{eval}} \sim \pi$. For generating the rewards, we create a misspecified reward setting by defining:

$$r(a_i, c_i, x_i) = \mathbb{1}\left\{u_i < \pi(a_i|c_i, x_i) \cdot \frac{||x_i||_1}{c_{\text{exp}} d_x}\right\},$$

where $u_i \sim U[0,1]$. The reward depends on two factors; the first one is the Manhattan norm of the context; the further from 0, the more likely it is to observe a positive reward. The second factor is the evaluation policy $\pi(a_i|c_i, x_i)$, which makes this a misspecified reward setting when the logging and evaluation policies are different enough. In this case, the (a_i, c_i, x_i) triplets having the highest probability of observation under the evaluation policy are more likely to observe positive rewards, resulting in a significant difference with respect to the observed rewards under the logging policy for such triplets. We sample rewards using this method for the logging ($\mathcal{R}_{\text{log}}$) and evaluation ($\mathcal{R}_{\text{eval}}$) data to obtain $\mathcal{D}_{\text{log}} := (\mathcal{X}, \mathcal{C}, \mathcal{A}_{\text{log}}, \mathcal{R}_{\text{log}})$ and $\mathcal{D}_{\text{eval}} := (\mathcal{X}, \mathcal{C}, \mathcal{A}_{\text{eval}}, \mathcal{R}_{\text{eval}})$. Finally, we select a subset for N samples from both sets. A representation of the generated structure can be found in Figure 20.

4.1.1 Synthetic results

In this section we analyze CHIPS performance while varying parameters of the synthetic dataset. In our experiments, the generation process for each parameter value is repeated 100 times with different random seeds. The final reported results are the average over all experiments, with the standard deviation corresponding to the lighter bands represented in all the figures. The basic configuration for the parameters used throughout the experiments can be found in Appendix C, along with the specifications of the hardware used. We use Random Forest (Breiman, 2001) to obtain $\hat{q}(x,a)$ in DM-based methods and mini-batch KMeans (Sculley, 2010) implementation in SciKit-Learn (Pedregosa et al., 2011) as the clustering method for CHIPS (alternative clustering methods are also discussed in Appendix D). Mini-batch KMeans is a variant of KMeans that slightly compromises accuracy to gain a drastic reduction in computational time. The main idea consists of sampling (a mini-batch) from the training samples and assigning the samples to the nearest centroid, then updating the centroids using the streaming average of the sample and all previous samples assigned to that centroid. We also use $\beta = -1$, maximizing the distributional shift between logging and evaluation policies.

Number of clusters. For this experiment, we vary the number of clusters the CHIPS estimator uses, with values ranging from 1 to 1000. Since $\beta = -1$, the implementation of CHIPS using ML reward estimation is unsuccessful (see Appendix D.3 for a further discussion). On the other hand, for the MAP case, we observe a v-shaped error graph (see Figure 1 (left)), suggesting that CHIPS performance is sensitive to effectiveness of clustering. In particular, we have a highly biased estimation when assuming insufficient or excessive clusters (see Figure 3). The reason for this bias in the first case might be an oversimplification of the structure of the cluster space. Conversely, we progressively gain bias when we select too many clusters according to Proposition 3.8 as CHIPS converges to IPS. In this case, CHIPS is also vulnerable to reward misspecification, which causes an increase in variance.

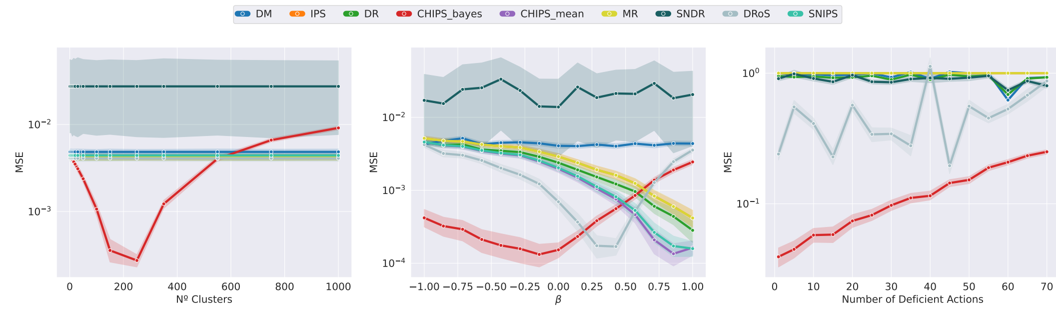


Figure 1: From left to right, the mean square error in the synthetic dataset experiments varying the number of clusters, the distributional shift between logging and evaluation policy (β), and the number of deficient actions in the logging data (normalized w.r.t. IPS).

Beta. In this experiment, we explore the effect of the distributional policy shift between π and π_0. When we take lower values in our range (i.e., $\pi_0 \longleftrightarrow \pi$), the shift in the policies is considerable and introduces bias in IPS estimates for large context-action spaces (Saito & Joachims, 2022; Sachdeva et al., 2020). The CHIPS estimator tries to mitigate this effect by considering all context-action samples inside a cluster as if all had the same context. However, when β takes lower values, these virtual *extra* samples might not be enough to make an accurate estimation since the most relevant (x, a) pairs ($\pi(a|x)$ closer to 1), are severely underrepresented and misspecified (see Appendix D.3). Therefore, in this case, the ML estimation in CHIPS is not effective, while the MAP estimation offers some *resistance* to this problem by pushing the reward estimate towards the posterior expectation, making it sensitive to the choice of the prior. However, this resistance might be counterproductive when the distributional shift is small (β closer to 1), and both the ML estimate and IPS converge faster to a better estimation (see Figure 1 (center)).

Deficient actions. In this setting we explicitly set the probability (π_0) of observing a variable number of actions in the action space to 0 and evaluate CHIPS' response in a space with 200 actions and $\beta = -1$. This setting is quite challenging as not only we have deficient actions but also a significant distributional shift between policies. The majority of baselines perform at a similar level than IPS with the exception of DRoS (Su et al. (2020a)), that performs slightly better but is still outperformed by CHIPS.

Additional experiments and discussions of results varying other parameters, different clustering methods, and a time complexity analysis can be found in Appendix G.

4.2 Real dataset

Following the literature, for assessing the capabilities of the CHIPS estimator in a real-world environment, we compare the performance in the Open Bandit Dataset (OBD) (Saito et al., 2020) of IPS, DM, DR, MRDR (Farajtabar et al., 2018) and MIPS (Saito & Joachims, 2022), with and without SLOPE (Su et al., 2020b). The OBD dataset was gathered using two different policies during an A/B test: uniform random, which we consider as logging (i.e., π_0), and Thompson sampling (Thompson, 1933, 1935), which we consider as evaluation (i.e., π). The dataset is based on a recommendation system for fashion e-commerce. We observe user data as contexts x, items to recommend $a \in \mathcal{A}$ (with $|\mathcal{A}| = 240$) and rewards $r \in \{0, 1\}$ representing user interactions.

Following the experimental protocol of Saito & Joachims (2022) (see Appendix F), we experiment with the real dataset varying the number of logging samples available for the estimation using 50 000, 100 000, and 500 000 samples to compute the Empirical Cumulative Distribution Function (ECDF) of the normalized mean squared error with respect to IPS. We increase the number of clusters for CHIPS as more logging samples are available to try to maximize performance, following the intuition from our earlier experiments on the synthetic dataset (see Figure 12 (right)). We use 8 clusters for 100 000 samples as a reference from our results for 240 actions in the synthetic dataset (see Figure 12 (left)). Regarding the clustering method, we use again mini-batch KMeans, this method ensures fairly homogeneous clusters for the expected rewards within a cluster as observed in Figure 21.

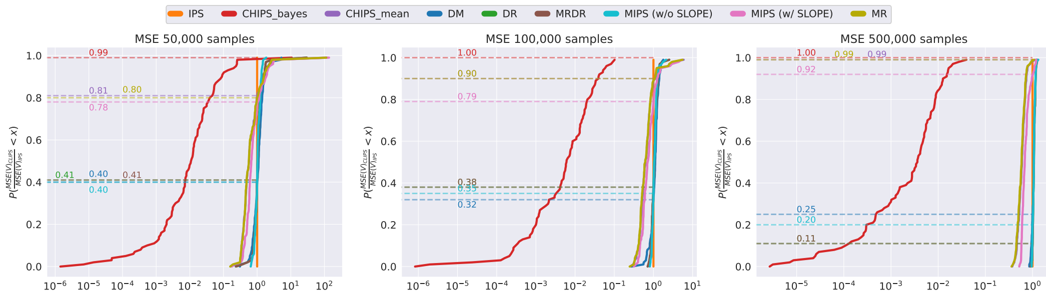


Figure 2: ECDF of the relative mean squared error with respect to IPS for the real dataset using 50000 (left), 100000 (center), and 500000 (right) logging samples.

We observe that the CHIPS estimator using the ML approximation is slightly better (+3%) than MIPS when few samples are available (see Figure 2, (left)). This performance gap widens (+11%) as the CHIPS estimator has more samples available (see Figure 2, (center)) and starts narrowing (+7%) as the number of samples is enough for MIPS to also start making more accurate estimations (see Figure 2, (right)).

Using the MAP reward estimation for CHIPS provides a considerable advantage in all experiments since the real dataset present severe reward misspecification, as discussed in Appendix D.3. Similarly to the synthetic dataset, the partition structure of the cluster space and the α parameter in MAP are sensitive parameters. In particular, for the number of clusters, we observe that using an insufficient or excessive number of clusters can negatively impact performance (see Figure 14 (left)) as we discussed in section Section 4.1.1. Regarding the value of α for the Beta prior, following the results from the synthetic experiment studying the effect of this parameter conjointly with the distributional shift between logging and evaluation policies (see discussion in Appendix D.4 and Figure 13), we used $\alpha = 20$ as the logging policy is uniform (the equivalent of $\beta = 0$ in the synthetic dataset). Figure 14 (right) shows how choosing a lower or higher value for α deteriorates the performance of the CHIPS estimator, reaffirming the results observed in the synthetic dataset (see Figure 13).

5 Conclusions, Limitations and Future Work

In this work we have explored an alternative approach to the OPE problem by clustering contexts instead of pooling information over actions to mitigate the problems arising in IPS when the Common Support condition does not hold. The proposed setup for the OPE problem using contexts led to the CHIPS estimator, which uses a similar approach to IPS applied over clusters instead of contexts. We have studied this estimator extensively from a theoretical and practical perspective, evaluating its performance for different configurations in a controlled synthetic dataset and a real-world example. The results obtained in the experiments for both cases demonstrate that the CHIPS estimator provides a significant improvement in estimation accuracy, outperforming existing estimators if the context space has a cluster structure. The accuracy of CHIPS is also influenced by the accuracy of the clustering method and the homogeneity behaviour of contexts inside the same cluster. Additionally, choosing a balanced number of clusters to avoid over- and under-simplification of the cluster structure is an important part of the estimation process and opens the possibility of exploring if it is possible to estimate the optimal value for hyperparameters beyond empirical estimation or even if combining CHIPS with pure action-embedding methods like MIPS can improve general performance.

References

Ankerst, M., Breunig, M. M., Kriegel, H.-P., and Sander, J. Optics: Ordering points to identify the clustering structure. *SIGMOD Rec.*, 28(2):49–60, jun 1999. ISSN 0163-5808. doi: 10.1145/304181.304187. URL https://doi.org/10.1145/304181.304187.

Attias, H. A variational baysian framework for graphical models. In Solla, S., Leen, T., and Müller, K. (eds.), *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999. URL https://proceedings.neurips.cc/paper_files/paper/1999/file/74563ba21a90da13dacf2a73e3ddefa7-Paper.pdf.

Bastani, H. and Bayati, M. Online decision making with high-dimensional covariates. *Operations Research*, 68, 11 2019. doi: 10.1287/opre.2019.1902.

Bendada, W., Salha, G., and Bontempelli, T. Carousel personalization in music streaming apps with contextual bandits. In *Proceedings of the 14th ACM Conference on Recommender Systems*, RecSys '20, pp. 420–425, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450375832. doi: 10.1145/3383313.3412217. URL https://doi.org/10.1145/3383313.3412217.

Beygelzimer, A. and Langford, J. 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.

Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

Blei, D. and Jordan, M. Variational inference for dirichlet process mixtures. *Bayesian Analysis*, 1, 03 2006. doi: 10.1214/06-BA104.

Bouneffouf, D., Rish, I., and Aggarwal, C. Survey on applications of multi-armed and contextual bandits. In *2020 IEEE Congress on Evolutionary Computation (CEC)*, pp. 1–8, 2020. doi: 10.1109/CEC48606.2020.9185782.

Box, G. E. P. and Muller, M. E. A Note on the Generation of Random Normal Deviates. *The Annals of Mathematical Statistics*, 29(2):610 – 611, 1958. doi: 10.1214/aoms/1177706645. URL https://doi.org/10.1214/aoms/1177706645.

Breiman, L. Random forests. *Machine Learning*, 45:5–32, 10 2001. doi: 10.1023/A:1010950718922.

Comaniciu, D. and Meer, P. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(5):603–619, may 2002. ISSN 0162-8828. doi: 10.1109/34.1000236. URL https://doi.org/10.1109/34.1000236.

Deng, L. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Dudík, M., Langford, J., and Li, L. Doubly robust policy evaluation and learning. In *International Conference on Machine Learning*, 2011. URL https://api.semanticscholar.org/CorpusID:7806620.

Dumitrascu, B., Feng, K., and Engelhardt, B. E. Pg-ts: Improved thompson sampling for logistic contextual bandits. In *Neural Information Processing Systems*, 2018. URL https://api.semanticscholar.org/CorpusID:29153062.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pp. 226–231. AAAI Press, 1996.

Everitt, B. An introduction to finite mixture distributions. *Statistical Methods in Medical Research*, 5 (2):107–127, 1996. doi: 10.1177/096228029600500202. URL https://doi.org/10.1177/096228029600500202. PMID: 8817794.

Farajtabar, M., Chow, Y., and Ghavamzadeh, M. More robust doubly robust off-policy evaluation. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1447–1456. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/farajtabar18a.html.

Felicioni, N., Dacrema, M. F., Restelli, M., and Cremonesi, P. Off-policy evaluation with deficient support using side information. In *Neural Information Processing Systems*, 2022. URL https://api.semanticscholar.org/CorpusID:258508965.

Frey, B. J. and Dueck, D. Clustering by passing messages between data points. *Science*, 315(5814): 972–976, 2007. doi: 10.1126/science.1136800. URL https://www.science.org/doi/abs/10.1126/science.1136800.

Guo, Y., Liu, H., Yue, Y., and Liu, A. Distributionally robust policy evaluation under general covariate shift in contextual bandits. *ArXiv*, abs/2401.11353, 2024. URL https://api.semanticscholar.org/CorpusID:267069353.

Horvitz, D. G. and Thompson, D. J. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685, 1952. URL https://api.semanticscholar.org/CorpusID:120274071.

Irpan, A., Rao, K., Bousmalis, K., Harris, C., Ibarz, J., and Levine, S. Off-policy evaluation via off-policy classification. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/b5b03f06271f8917685d14cea7c6c50a-Paper.pdf.

Kallus, N. and Uehara, M. *Intrinsically Efficient, Stable, and Bounded off-Policy Evaluation for Reinforcement Learning*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Kerman, J. Neutral noninformative and informative conjugate beta and gamma prior distributions. *Electronic Journal of Statistics*, 5, 01 2011. doi: 10.1214/11-EJS648.

Kiyohara, H., Uehara, M., Narita, Y., Shimizu, N., Yamamoto, Y., and Saito, Y. Off-policy evaluation of ranking policies under diverse user behavior. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, pp. 1154–1163, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599447. URL https://doi.org/10.1145/3580305.3599447.

Krizhevsky, A., Nair, V., and Hinton, G. Cifar-100 (canadian institute for advanced research). 2009. URL http://www.cs.toronto.edu/~kriz/cifar.html.

Kuzborskij, I., Vernade, C., Gyorgy, A., and Szepesvari, C. Confident off-policy evaluation and selection through self-normalized importance weighting. *ArXiv*, abs/2006.10460, 2020. URL https://api.semanticscholar.org/CorpusID:219792866.

Li, L., Chu, W., Bellevue, M., Langford, J., and Wang, X. An unbiased offline evaluation of contextual bandit algorithms with generalized linear models. *Journal of Machine Learning Research*, 1, 01 2011.

McNellis, R., Elmachtoub, A. N., Oh, S., and Petrik, M. A practical method for solving contextual bandit problems using decision trees. In Elidan, G., Kersting, K., and Ihler, A. T. (eds.), *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, August 11-15, 2017*. AUAI Press, 2017. URL http://auai.org/uai2017/proceedings/papers/171.pdf.

Metelli, A. M., Russo, A., and Restelli, M. Subgaussian and differentiable importance sampling for off-policy evaluation and learning. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 8119–8132. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/4476b929e30dd0c4e8bdbcc82c6ba23a-Paper.pdf.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Peng, J., Zou, H., Liu, J., Li, S., Jiang, Y., Pei, J., and Cui, P. Offline policy evaluation in large action spaces via outcome-oriented action grouping. In *Proceedings of the ACM Web Conference 2023*, WWW '23, pp. 1220–1230, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394161. doi: 10.1145/3543507.3583448. URL https://doi.org/10.1145/3543507.3583448.

Sachdeva, N., Su, Y.-H., and Joachims, T. Off-policy bandits with deficient support. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020. URL https://api.semanticscholar.org/CorpusID:214361356.

Sachdeva, N., Wang, L., Liang, D., Kallus, N., and McAuley, J. Off-policy evaluation for large action spaces via policy convolution, 2023.

Saito, Y. and Joachims, T. Off-policy evaluation for large action spaces via embeddings. In *International Conference on Machine Learning*, 2022. URL https://api.semanticscholar.org/CorpusID:246823434.

Saito, Y., Aihara, S., Matsutani, M., and Narita, Y. Open bandit dataset and pipeline: Towards realistic and reproducible off-policy evaluation. In *NeurIPS Datasets and Benchmarks*, 2020. URL https://api.semanticscholar.org/CorpusID:235435303.

Saito, Y., Udagawa, T., Kiyohara, H., Mogi, K., Narita, Y., and Tateno, K. 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, 2021. Association for Computing Machinery. ISBN 9781450384582. doi: 10.1145/3460231.3474245. URL https://doi.org/10.1145/3460231.3474245.

Saito, Y., Ren, Q., and Joachims, T. Off-policy evaluation for large action spaces via conjunct effect modeling. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29734–29759. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/saito23b.html.

Sculley, D. Web-scale k-means clustering. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pp. 1177–1178, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605587998. doi: 10.1145/1772690.1772862. URL https://doi.org/10.1145/1772690.1772862.

Shi, J. and Malik, J. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000. doi: 10.1109/34.868688.

Shimizu, T. and Forastiere, L. Doubly robust estimator for off-policy evaluation with large action spaces, 2023.

Su, Y., Dimakopoulou, M., Krishnamurthy, A., and Dudik, M. Doubly robust off-policy evaluation with shrinkage. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9167–9176. PMLR, 13–18 Jul 2020a. URL https://proceedings.mlr.press/v119/su20a.html.

Su, Y., Srinath, P., and Krishnamurthy, A. Adaptive estimator selection for off-policy evaluation. In III, H. D. and Singh, A. (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.

Swaminathan, A. and Joachims, T. Counterfactual risk minimization: Learning from logged bandit feedback. In Bach, F. and Blei, D. (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 814–823, Lille, France, 07–09 Jul 2015a. PMLR. URL https://proceedings.mlr.press/v37/swaminathan15.html.

Swaminathan, A. and Joachims, T. The self-normalized estimator for counterfactual learning. In *Advances in Neural Information Processing Systems, 28*, 2015b. URL https://api.semanticscholar.org/CorpusID:6359643.

Taufiq, M. F., Doucet, A., Cornish, R., and Ton, J.-F. Marginal density ratio for off-policy evaluation in contextual bandits. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=noyleECBam.

Thomas, P. S. and Brunskill, E. 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.

Thompson, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:285–294, 1933. URL https://api.semanticscholar.org/CorpusID:120462794.

Thompson, W. R. On the theory of apportionment. *American Journal of Mathematics*, 57(2):450–456, 1935. ISSN 00029327, 10806377. URL http://www.jstor.org/stable/2371219.

Tuyl, F., Gerlach, R., and Mengersen, K. A comparison of bayes–laplace, jeffreys, and other priors. *American Statistician - AMER STATIST*, 62:40–44, 02 2008. doi: 10.1198/000313008X267839.

Varatharajah, Y. and Berry, B. A contextual-bandit-based approach for informed decision-making in clinical trials. *Life*, 12(8), 2022. ISSN 2075-1729. doi: 10.3390/life12081277. URL https://www.mdpi.com/2075-1729/12/8/1277.

Voloshin, C., Le, H., Jiang, N., and Yue, Y. Empirical study of off-policy policy evaluation for reinforcement learning. In Vanschoren, J. and Yeung, S. (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/a5e00132373a7031000fd987a3c9f87b-Paper-round1.pdf.

Wang, Y.-X., Agarwal, A., and Dudík, M. 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.

Ward, J. H. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963. ISSN 01621459. URL http://www.jstor.org/stable/2282967.

Zhan, R., Hadad, V., Hirshberg, D. A., and Athey, S. Off-policy evaluation via adaptive weighting with data from contextual bandits. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021. URL https://api.semanticscholar.org/CorpusID:235313689.

Zhang, T., Ramakrishnan, R., and Livny, M. Birch: An efficient data clustering method for very large databases. In *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, SIGMOD '96, pp. 103–114, New York, NY, USA, 1996. Association for Computing Machinery. ISBN 0897917944. doi: 10.1145/233269.233324. URL https://doi.org/10.1145/233269.233324.

A Theoretical Result Proofs

A.1 Proposition 3.3

Given a policy π, if both Assumption 3.1 and 3.2 hold, from the refinement of the policy value definition in a cluster-based bandits process (introduced in Section 3), we have that:

$$
\begin{aligned}
V(\pi) &:= \mathbb{E}_{p(x)p(c|x)\pi(a|x)p(r|a,c,x)}\left[r\right] \\
&= \mathbb{E}_{p(c)p(x|c)\pi(a|x)}\left[q(a,c,x)\right] && (8) \\
&= \mathbb{E}_{p(c)}\left[\int_{\mathcal{X}} p(x|c) \sum_{a\in\mathcal{A}} p(a|c) q(a,c) dx\right] && (9) \\
&= \mathbb{E}_{p(c)}\left[\int_{\mathcal{X}} \sum_{a\in\mathcal{A}} p(x|c)\pi(a|x) q(a,c) dx\right] \\
&= \mathbb{E}_{p(c)}\left[\sum_{a\in\mathcal{A}} q(a,c) \int_{\mathcal{X}} p(x|c)\pi(a|x) dx\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbb{E}_{p(c)}\left[\sum_{a\in\mathcal{A}} p(a|c,\pi)q(a,c)\right] \quad (10)\\
&= \mathbb{E}_{p(c)p(a|c,\pi)}\left[q(a,c)\right]
\end{aligned}
$$

Where in Equation 8 we used the Bayes Theorem, in Equation 9 the fact that under Assumption 3.2 $q(a,c,x) = q(a,c)$, and the definition of $p(a|c,\pi)$ in Equation 10.

A.2 Proposition 3.4

Given a policy π and under Assumptions 3.1 and 3.2 we have that:

$$
\begin{aligned}
\mathbb{E}_{\mathcal{D}}\left[\hat{V}_{\text{CHIPS}}(\pi;\mathcal{D})\right] &= \mathbb{E}_{\mathcal{D}}\left[w(a,c)r\right] \quad (11)\\
&= \mathbb{E}_{p(x)p(c|x)\pi_0(a|x)p(r|a,c,x)}\left[w(a,c)r\right]\\
&= \mathbb{E}_{p(c)p(x|c)\pi_0(a|x)}\left[w(a,c)q(a,c)\right] \quad (12)\\
&= \mathbb{E}_{p(c)}\left[\int_{\mathcal{X}} p(x|c)\sum_{a\in\mathcal{A}}\pi_0(a|x)w(a,c)q(a,c)dx\right]\\
&= \mathbb{E}_{p(c)}\left[\int_{\mathcal{X}}\sum_{a\in\mathcal{A}} p(x|c)\pi_0(a|x)w(a,c)q(a,c)dx\right]\\
&= \mathbb{E}_{p(c)}\left[\sum_{a\in\mathcal{A}} w(a,c)q(a,c)\left(\int_{\mathcal{X}} p(x|c)\pi_0(a|x)dx\right)\right]\\
&= \mathbb{E}_{p(c)}\left[\sum_{a\in\mathcal{A}}\frac{p(a|c,\pi)}{\cancel{p(a|c,\pi_0)}}q(a,c)\cancel{p(a|c,\pi_0)}\right] \quad (13)\\
&= \mathbb{E}_{p(c)p(a|c,\pi)}\left[q(a,c)\right]\\
&= \mathbb{E}_{p(x)p(c|x)\pi(a|x)p(r|a,c,x)}\left[r\right] \quad (14)\\
&= V(\pi)
\end{aligned}
$$

In Equation 11, we have used the linearity of expectation, in Equation 12 the definition of $q(a,c,x)$ and Assumption 3.2. Equation 13 is just using the definition of $p(a|c,\pi)$ while Equation 14 is a combination of Proposition 3.3 and the equivalence $q(a,c) = q(a,c,x)$ under the given assumptions.

A.3 Proposition 3.6

Given the logging data $\mathcal{D} = \{(x_i,a_i,r_i)\}$, a logging policy π_0, and an evaluation policy π having common cluster support over it, we have that:

$$
\begin{aligned}
\text{Bias}(\hat{V}_{\text{CHIPS}}(V;\mathcal{D})) &= \mathbb{E}_{\mathcal{D}}\left[w(c,a)r\right] - V(\pi)\\
&= \mathbb{E}_{p(x)p(c|x)\pi_0(a|x)p(r|a,c,x)}\left[w(a,c)r\right] - V(\pi)\\
&= \mathbb{E}_{p(x)p(c|x)\pi_0(a|x)}\left[w(a,c)q(a,c,x)\right] - V(\pi)\\
&= \mathbb{E}_{p(x)p(c|x)\pi_0(a|x)}\left[w(a,c)q(a,c,x)\right] - \mathbb{E}_{p(x)p(c|x)\pi(a|x)}\left[q(a,c,x)\right]\\
&= \mathbb{E}_{p(c)p(x|c)\pi_0(a|x)}\left[w(a,c)q(a,c,x)\right] - \mathbb{E}_{p(c)p(x|c)\pi(a|x)}\left[q(a,c,x)\right]\\
&= \mathbb{E}_{p(c)}\left[\int_{\mathcal{X}} p(x|c)\sum_{a\in\mathcal{A}}\pi_0(a|x)w(c,a)q(a,c,x)dx\right]\\
&\quad - \mathbb{E}_{p(c)}\left[\int_{\mathcal{X}} p(x|c)\sum_{a\in\mathcal{A}}\pi(a|x)q(a,c,x)dx\right] \quad (15)
\end{aligned}
$$

Under Assumption 3.5 we have that $\delta^-_\pi \le \frac{\pi(a|c,x)}{p(a|c,\pi)} \le \delta^+_\pi$, $\quad \delta^-_\pi \le \frac{\pi(a|c,x)}{p(a|c,\pi)} \le \delta^+_\pi \quad \forall(x,c,a) \in \mathcal{D}$. We denote then $\delta^+ = \max\{\delta^+_\pi, \delta^+_{\pi_0}\}, \delta^- = \min\{\delta^-_\pi, \delta^-_{\pi_0}\}, \Delta = \delta^+ - \delta^-$, and we can give an upper bound as follows:

$$
\mathbb{E}_{p(c)}\left[\int_{\mathcal{X}} p(x|c)\sum_{a\in\mathcal{A}}\pi_0(a|x)w(c,a)q(a,c,x)dx\right] \quad (16)
$$

$$
\begin{aligned}
&\quad - \mathbb{E}_{p(c)}\left[\int_{\mathcal{X}} p(x|c) \sum_{a\in\mathcal{A}} \pi(a|x) q(a,c,x) dx\right] \\
&\leq \mathbb{E}_{p(c)}\left[\int_{\mathcal{X}} p(x|c) \sum_{a\in\mathcal{A}} \delta^{+} p(a|c,\pi_0) w(c,a) q(a,c,x) dx\right] \qquad (17)\\
&\quad - \mathbb{E}_{p(c)}\left[\int_{\mathcal{X}} p(x|c) \sum_{a\in\mathcal{A}} \delta^{-} p(a|c,\pi) q(a,c,x) dx\right] \\
&= \mathbb{E}_{p(c)}\left[\sum_{a\in\mathcal{A}} p(a|c,\pi) \int_{\mathcal{X}} p(x|c)\delta^{+} \frac{p(a|c,\pi)}{\cancel{p(a|c,\pi_0)}} \cancel{p(a|c,\pi_0)} q(a,c,x) dx\right] \\
&\quad - \mathbb{E}_{p(c)}\left[\sum_{a\in\mathcal{A}} p(a|c,\pi) \int_{\mathcal{X}} p(x|c)\delta^{-} q(a,c,x) dx\right] \\
&= \mathbb{E}_{p(c)p(a|c,\pi)}\left[\delta^{+} \int_{\mathcal{X}} p(x|c) q(a,c,x) dx\right] - \mathbb{E}_{p(c)p(a|c,\pi)}\left[\delta^{-} \int_{\mathcal{X}} p(x|c) q(a,c,x) dx\right] \\
&= \mathbb{E}_{p(c)p(a|c,\pi)}\left[\mathbb{E}_{p(x|c)}\left[q(a,c,x)\right](\delta^{+} - \delta^{-})\right] \\
&= \mathbb{E}_{p(c)p(a|c,\pi)}\left[\mathbb{E}_{p(x|c)}\left[q(a,c,x)\right]\Delta\right]
\end{aligned}
$$

Note that in Equation 17 we can follow an analogous path to establish a lower bound:

$$
\begin{aligned}
&\mathbb{E}_{p(c)}\left[\int_{\mathcal{X}} p(x|c) \sum_{a\in\mathcal{A}} \pi_0(a|x) w(c,a) q(a,c,x) dx\right] - \mathbb{E}_{p(c)}\left[\int_{\mathcal{X}} p(x|c) \sum_{a\in\mathcal{A}} \pi(a|x) q(a,c,x) dx\right] \\
&\geq \mathbb{E}_{p(c)}\left[\int_{\mathcal{X}} p(x|c) \sum_{a\in\mathcal{A}} \delta^{-} p(a|c,\pi_0) w(c,a) q(a,c,x) dx\right] \\
&\quad - \mathbb{E}_{p(c)}\left[\int_{\mathcal{X}} p(x|c) \sum_{a\in\mathcal{A}} \delta^{+} p(a|c,\pi) q(a,c,x) dx\right] \\
&= -\mathbb{E}_{p(c)p(a|c,\pi)}\left[\mathbb{E}_{p(x|c)}\left[q(a,c,x)\right]\Delta\right]
\end{aligned}
$$

From which we have:

$$
\left|\text{Bias}(\hat{V}_{\text{CHIPS}}(\pi;\mathcal{D}))\right| \leq \left|\mathbb{E}_{p(c)p(x|c)p(a|c,\pi)}\left[q(a,c,x)\cdot\Delta\right]\right|
$$

A.4 Proposition 3.7

Since the observations are independent we have that

$$
\begin{aligned}
&N\left(\text{MSE}\left(\hat{V}_{\text{IPS}}(\pi)\right) - \text{MSE}\left(\hat{V}_{\text{CHIPS}}(\pi)\right)\right. \\
&= \mathbb{V}_{x,a,r}[\omega(x,a)r] - \mathbb{V}_{c,a,r}[\omega(a,c)r] - N\,\text{Bias}\left(\hat{V}_{\text{CHIPS}}(\pi)\right)^2
\end{aligned}
$$

We now analyze the difference in variance:

$$
\begin{aligned}
&V_{p(c)p(x|c)\pi_0(a|x)p(r|a,c,x)}[\omega(x,a)r] - V_{p(c)p(x|c)\pi_0(a|x)p(r|a,c,x)}[\omega(a,c)r] \\
&= \mathbb{E}_{p(c|p(x|c)\pi_0(a|x)p(r|a,c,x)}\left[\omega(x,a)r^2\right] - V(\pi)^2 \\
&\quad - \left(\mathbb{E}_{p(c)p(x|c)\pi_b(a|x)p(r|a,c,x)}\left[\omega(a,c)^2\cdot r^2\right] - \left(V(\pi) + \text{Bias}\left(\hat{V}_{\text{CHIPS}}(\pi)\right)\right)^2\right. \\
&= \mathbb{E}_{p(c)p(x|c)\pi_0(a|x)}\left[\left(\omega(x,a)^2 - \omega(a,c)^2\right)\mathbb{E}_{p(r|a,c,x)}\left[r^2\right]\right] \\
&\quad + 2V(\pi)\,\text{Bias}\left(\hat{V}_{\text{CHIPS}}(\pi)\right) + \text{Bias}\left(\hat{V}_{\text{CHIPS}}(\pi)\right)^2
\end{aligned}
$$

This implies that

$$
\begin{aligned}
&N\left(\text{MSE}\left(\hat{V}_{\text{IPS}}(\pi)\right) - \text{MSE}\left(\hat{V}_{\text{CHIPS}}(\pi)\right)\right) \\
&= \mathbb{E}_{p(c)p(x|c)\pi_0(a|x)}\left[\left(\omega(x,a)^2 - \omega(a,c)^2\right)\mathbb{E}_{p(r|a,c,x)}\left[r^2\right]\right] \\
&\quad + 2V(\pi)\,\text{Bias}\left(\hat{V}_{\text{CHIPS}}(\pi)\right) + (1-N)\,\text{Bias}\left(\hat{V}_{\text{CHIPS}}(\pi)\right)^2
\end{aligned}
$$

A.5 Proposition 3.8

Given the logging policy π_0 and some evaluation policy π, the absolute bias of the CHIPS estimator when Assumption 3.2, we have that:

$$
\begin{aligned}
\mathrm{Bias}(\hat{V}_{\mathrm{CHIPS}}(V;\mathcal{D})) &= \mathbb{E}_{\mathcal{D}}\left[w(c,a)r\right] - V(\pi) \\
&= \mathbb{E}_{p(x)p(c|x)\pi_0(a|x)p(r|a,c,x)}\left[w(a,c)r\right] - V(\pi) \\
&= \mathbb{E}_{p(c)p(x|c)\pi_0(a|x)}\left[w(a,c)q(a,c)\right] - \mathbb{E}_{p(c)p(x|c)\pi(a|x)}\left[w(a,c)q(a,c)\right] \\
&= \mathbb{E}_{p(c)p(x|c)\pi_0(a|x)}\left[w(a,c)q(a,c)\right] - \mathbb{E}_{p(c)p(x|c)\pi(a|x)}\left[w(a,c)q(a,c)\right] \\
&= \mathbb{E}_{p(c)}\left[\int_{\mathcal{X}} p(x|c)\sum_{a\in\mathcal{A}}\pi_0(a|x)w(c,a)q(a,c)dx\right] \\
&\qquad - \mathbb{E}_{p(c)}\left[\int_{\mathcal{X}} p(x|c)\sum_{a\in\mathcal{A}}\pi(a|x)q(a,c)dx\right] \\
&= \mathbb{E}_{p(c)}\left[\sum_{a\in\mathcal{A}} w(c,a)q(a,c)\int_{\mathcal{X}} p(x|c)\pi_0(a|x)dx\right] \\
&\qquad - \mathbb{E}_{p(c)}\left[\sum_{a\in\mathcal{A}} q(a,c)\int_{\mathcal{X}} p(x|c)\pi(a|x)dx\right] \\
&= \mathbb{E}_{p(c)}\left[\sum_{a\in\mathcal{A}} w(c,a)q(a,c)p(a|c,\pi_0)\right] - \mathbb{E}_{p(c)}\left[\sum_{a\in\mathcal{A}} q(a,c)p(a|c,\pi)\right] \\
&= \mathbb{E}_{p(c)}\left[\sum_{a\in\mathcal{U}(c,\pi_0)^c} w(c,a)q(a,c)p(a|c,\pi_0)\right] - \mathbb{E}_{p(c)}\left[\sum_{a\in\mathcal{A}} q(a,c)p(a|c,\pi)\right] \qquad (18) \\
&= \mathbb{E}_{p(c)}\left[\sum_{a\in\mathcal{U}(c,\pi_0)^c} \frac{p(a|c,\pi_0)}{\cancel{p(a|c,\pi_0)}}\cancel{p(a|c,\pi_0)}q(a,c)\right] \\
&\qquad - \mathbb{E}_{p(c)}\left[\sum_{a\in\mathcal{A}} q(a,c)p(a|c,\pi)\right] \\
&= \mathbb{E}_{p(c)}\left[-\sum_{a\in\mathcal{U}(c,\pi_0)} p(a|c,\pi)q(a,c)\right]
\end{aligned}
$$

Where in Equation 18 we note that $p(a|c,\pi_0) = 0$ if $a \in \mathcal{U}(c,\pi_0)$. Following an analogous procedure we can give an expression for the bias of IPS in a cluster bandits setup:

$$
\begin{aligned}
\mathrm{Bias}(\hat{V}_{\mathrm{IPS}}(V;\mathcal{D})) &= \mathbb{E}_{\mathcal{D}}\left[w(a,x)r\right] - V(\pi) \\
&= \mathbb{E}_{p(x)p(c|x)\pi_0(a|x)p(r|a,c,x)}\left[w(a,x)r\right] - V(\pi) \\
&= \mathbb{E}_{p(c)p(x|c)\pi_0(a|x)}\left[w(a,x)q(a,c)\right] - \mathbb{E}_{p(c)p(x|c)\pi(a|x)}\left[q(a,c)\right] \\
&= \mathbb{E}_{p(c)}\left[\int_{\mathcal{X}} p(x|c)\sum_{a\in\mathcal{U}(c,x,\pi_0)^c}\pi_0(a|x)w(a,x)q(a,c)dx\right] \\
&\qquad - \mathbb{E}_{p(c)}\left[\int_{\mathcal{X}} p(x|c)\sum_{a\in\mathcal{A}}\pi(a|x)q(a,c)dx\right] \\
&= \mathbb{E}_{p(c)}\left[\sum_{a\in\mathcal{U}(c,x,\pi_0)^c} q(a,c)\int_{\mathcal{X}} p(x|c)\frac{\pi(a|x)}{\cancel{\pi_0(a|x)}}\cancel{\pi_0(a|x)}dx\right] \\
&\qquad - \mathbb{E}_{p(c)}\left[\sum_{a\in\mathcal{A}} q(a,c)\int_{\mathcal{X}} p(x|c)\pi(a|x)dx\right] \\
&= \mathbb{E}_{p(c)}\left[\sum_{a\in\mathcal{U}(c,x,\pi_0)^c} q(a,c)p(a|c,\pi)\right] - \mathbb{E}_{p(c)}\left[\sum_{a\in\mathcal{A}} q(a,c)p(a|c,\pi)\right]
\end{aligned}
$$

$$= \mathbb{E}_{p(c)}\left[- \sum_{a \in \mathcal{U}(c,x,\pi_0)} p(a|c,\pi) q(a,c) \right]$$

Since $q(a,c) \geq 0$ in the binary reward setting, it follows that $\left|\text{Bias}(\hat{V}_{\text{CHIPS}}(\pi;\mathcal{D}))\right| = \mathbb{E}_{p(c)}\left[\sum_{\mathcal{U}(c,\pi_0)} p(a|\pi,c)q(a,c)\right]$ and $\left|\text{Bias}(\hat{V}_{\text{IPS}}(\pi;\mathcal{D}))\right| = \mathbb{E}_{p(c)}\left[\sum_{\mathcal{U}(c,x,\pi_0)} p(a|\pi,c)q(a,c)\right]$ and consequently we have that:

$$\begin{aligned} \left|\text{Bias}(\hat{V}_{\text{IPS}}(\pi;\mathcal{D}))\right| - \left|\text{Bias}(\hat{V}_{\text{CHIPS}}(\pi;\mathcal{D}))\right| &= \mathbb{E}_{p(c)}\left[\sum_{\mathcal{U}(c,x,\pi_0)} p(a|\pi,c)q(a,c)\right] \\ &\quad - \mathbb{E}_{p(c)}\left[\sum_{\mathcal{U}(c,\pi_0)} p(a|\pi,c)q(a,c)\right] \\ &= \mathbb{E}_{p(c)}\left[\sum_{\mathcal{U}(c,x,\pi_0)\setminus\mathcal{U}(c,\pi_0)} p(a|\pi,c)q(a,c)\right] \end{aligned}$$

A.6 Proposition 3.10

Assuming that we have a set of embeddings $e \in \mathcal{E} \subset \mathbb{R}^{d_e}$ associated with the actions $a \in \mathcal{A}$ and an approximation $f_{\phi^*}(r)$ to the importance weights $w(a,x)$:

$$\begin{aligned} f_{\phi^*}(r) &:= \operatorname{argmin} \mathbb{E}_\phi\left[(w(a,x) - f_\phi(r))^2\right] \\ f_\phi &\in \{f_\phi : \mathbb{R} \to \mathbb{R} \mid \phi \in \Phi\} \end{aligned} \tag{19}$$

Then if we assume that $f_{\phi^*}(r) = w(a,x) + \epsilon$ for some $\epsilon \in \mathbb{R}$ we have that

$$\begin{aligned} &\left|\text{Bias}\left(\hat{V}_{\text{MR}};\mathcal{D}\right)\right| - \left|\text{Bias}\left(\hat{V}_{\text{CHIPS}};\mathcal{D}\right)\right| \\ &= -\mathbb{E}_{p(c)}\left[\sum_{a\in\mathcal{U}(c,\pi_0)} p(a \mid \pi, c)q(a,c)\right] + \text{Bias}\left(\hat{V}_{\text{IPS}}; D\right) + \mathbb{E}_{\mathcal{D}}\left[f_{\phi^*}(r)r\right] - \mathbb{E}_{\mathcal{D}}(w(a,x)) \\ &= -\mathbb{E}_{p(c)}\left[\sum_{a\in\mathcal{U}(c,\pi)} p(a \mid \pi, c)q(a,c)\right] - V(\pi) + \mathbb{E}_{\mathcal{D}}[f_{\phi^*}(r)r] \\ &= -\mathbb{E}_{p(c)}\left[\sum_{a\in\mathcal{U}(c,\pi_0)} p(a \mid \pi, c)q(a,c)\right] - V(\pi) + \mathbb{E}_{\mathcal{D}}[w(a,x)r] + \epsilon\mathbb{E}_{\mathcal{D}}[r] \\ &= -\mathbb{E}_{p(c)}\left[\sum_{a\in\mathcal{U}(c,\pi)} p(a \mid \pi, c)q(a,c)\right] + \mathbb{E}_{p(c)}\left[\sum_{a\in\mathcal{U}(x,c,\pi_0)} q(a,c) \underbrace{\int_{x\in x} p(x \mid c)\pi(a \mid x)dx}_{p(a|\pi,c)}\right] \\ &\qquad - \mathbb{E}_{p(c)}\left[\sum_{a\in\mathcal{A}} q(a,c) \underbrace{\int_{x\in x} p(x \mid c)\pi(a \mid x)dx}_{p(a|\pi,c)}\right] + \varepsilon\mathbb{E}_D[r] \\ &= \mathbb{E}_{p(c)}\left[\sum_{a\in\mathcal{U}(x,c,\pi_0)\setminus\mathcal{U}(c,\pi_0)} q(a,c)p(a \mid \pi, c)\right] - \mathbb{E}_{p(c)}\left[\sum_{a\in\mathcal{A}} q(a,c)p(a \mid \pi, c)\right] + \varepsilon\mathbb{E}_D[r] \\ &= -\mathbb{E}_{p(c)}\left[\sum_{a\in(\mathcal{U}(x,c,\pi_0)\setminus\mathcal{U}(c,\pi_0))^c} q(a,c)p(a \mid \pi, c)\right] + \epsilon\mathbb{E}_{p(c)}\left[\sum_{a\in\mathcal{A}} q(a,c)p(a \mid \pi_0, c)\right] \end{aligned}$$

for the MIPS case, we note that MIPS' bias can also be expressed similarly to CHIPS':

$$
\begin{aligned}
&\operatorname{Bias}\left(\hat{V}_{\mathrm{MIPS}} ; D\right) = \\
&= \mathbb{E}_D[w(x,e)r] - V(\pi) \\
&= \mathbb{E}_{p(x)\pi_0(a|x)p(e|x,a)p(r|x,a,e)}[w(x,e)r] - V(\pi) \\
&= \mathbb{E}_{p(x)}\left[\sum_{a\in\mathcal{A}} \pi_0(a \mid x) \sum_{e\in\mathcal{E}} p(e \mid x,a) w(x,e) q(x,e)\right] \\
&\qquad - \mathbb{E}_{p(x)}\left[\sum_{a\in\mathcal{A}} \pi(a \mid x) \sum_{e\in\mathcal{E}} p(e \mid x,a) w(x,e) q(x,e)\right] \\
&= \mathbb{E}_{p(x)}\left[\sum_{e\in\mathcal{E}} q(x,e)\left(\sum_{a\in\mathcal{A}} \pi_0(a \mid x), p(e \mid x,a)\right)\right] \\
&\qquad - \mathbb{E}_{p(x)}\left[\sum_{e\in\mathcal{E}} q(x,e)\left(\sum_{a\in\mathcal{A}} \pi(a \mid x) p(e \mid x,a)\right)\right] \\
&= E_{p(x)}\left[\sum_{e\in\mathcal{U}(e,\pi_0)^c} p\left(e \mid x,\pi_0\right) q(x,e) \frac{p(e \mid x,\pi)}{p\left(e \mid x,\pi_0\right)}\right] \\
&\qquad - \mathbb{E}_{p(x)}\left[\sum_{e\in\mathcal{E}} q(x,e) p(e \mid x,\pi)\right] \\
&= -\mathbb{E}_{p(x)}\left[\sum_{e\in\mathcal{U}(e,\pi_0)} p(e \mid x,\pi) q(x,e)\right]
\end{aligned}
$$

Therefore the difference in bias is:

$$
\begin{aligned}
&\left|\operatorname{Bias}\left(\hat{V}_{\mathrm{MIPS}} ; \mathcal{D}\right)\right| - \left|\operatorname{Bias}\left(\hat{V}_{\mathrm{CHIPS}} ; \mathcal{D}\right)\right| \\
&= \mathbb{E}_{p(x)}\left[\sum_{e\in\mathcal{U}(e,\pi_0)} p(e \mid x,\pi) q(x,e)\right] - \mathbb{E}_{p(c)}\left[\sum_{a\in\mathcal{U}(c,\pi_0)} p(a \mid c,\pi) q(a,c)\right]
\end{aligned}
$$

A.7 Proposition 3.11

Lemma A.1. *Given a policy π, under Assumption 3.1 we have the transformation:*

$$
w(a,c) = \mathbb{E}_{\pi_0(x|a,c)}\left[w(a,x)\right]
$$

Proof:

Given a logging policy π_0 and an evaluation policy π, in the cluster setting of the bandits problem we have that:

$$
\begin{aligned}
w(a,c) &= \frac{p(a|\pi,c)}{p(a|\pi_0,c)} \\
&= \frac{\int_{\mathcal{X}} \pi(a|x) p(x|c)}{p(a|\pi_0,c)} \qquad (20) \\
&= \frac{\cancel{p(a|c,\pi_0)} \int_{\mathcal{X}} \frac{\pi(a|x)}{\pi_0(a|x)} \pi_0(x|a,c)}{\cancel{p(a|\pi_0,c)}} \qquad (21) \\
&= \mathbb{E}_{\pi_0(x|a,c)}\left[w(a,x)\right]
\end{aligned}
$$

Where we have used the definition $p(a|\pi,c) = \int_{\mathcal{X}} \pi(a|x)p(x|c)$ in Equation 20, and that $\pi_0(x|a,c) = \frac{p(x|c)\pi_0(a|x)}{p(a|c,\pi_0)}$ in Equation 21.

Given a logging policy π_0 and an evaluation policy π, under Assumption 3.1 and Assumption 3.2 we have that

$$
N\left(\mathbb{V}_{\mathcal{D}}\left[\hat{V}_{\mathrm{IPS}}(\pi;\mathcal{D})\right] - \mathbb{V}_{\mathcal{D}}\left[\hat{V}_{\mathrm{CHIPS}}(\pi;\mathcal{D})\right]\right)
$$

$$
\begin{aligned}
&= N\left(\mathbb{V}_{\mathcal{D}}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\pi(a_i|x_i)}{\pi_0(a_i|x_i)}r_i\right] - \mathbb{V}_{\mathcal{D}}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{p(a_i|c_i,\pi)}{p(a_i|c_i,\pi_0)}r_i\right]\right)\\
&= \mathbb{V}_{\mathcal{D}}\left[\frac{\pi(a|x)}{\pi_0(a|x)}r\right] - \mathbb{V}_{\mathcal{D}}\left[\frac{p(a|c,\pi)}{p(a|c,\pi_0)}r\right] \qquad (22)\\
&= \left(\mathbb{E}_{\mathcal{D}}\left[w(a,x)^2r^2\right] - \underbrace{\cancelto{0}{\mathbb{E}_{\mathcal{D}}\left[w(a,x)r\right]^2}}_{V(\pi)}\right) - \left(\mathbb{E}_{\mathcal{D}}\left[w(a,c)^2r^2\right] - \underbrace{\cancelto{0}{\mathbb{E}_{\mathcal{D}}\left[w(a,c)r\right]^2}}_{V(\pi)}\right) \qquad (23)\\
&= \mathbb{E}_{p(x)p(c|x)\pi_0(a|x)}\left[w(a,x)^2\mathbb{E}_{p(r|a,c,x)}\left[r^2\right]\right] - \mathbb{E}_{p(x)p(c|x)\pi_0(a|x)}\left[w(a,c)^2\mathbb{E}_{p(r|a,c,x)}\left[r^2\right]\right]\\
&= \mathbb{E}_{p(x)p(c|x)\pi_0(a|x)}\left[\left(w(a,x)^2 - w(a,c)^2\right)\mathbb{E}_{p(r|a,c,x)}\left[r^2\right]\right]\\
&= \mathbb{E}_{p(c)p(x|c)\pi_0(a|x)}\left[\left(w(a,x)^2 - w(a,c)^2\right)\mathbb{E}_{p(r|a,c,x)}\left[r^2\right]\right]\\
&= \mathbb{E}_{p(c)}\left[\int_{\mathcal{X}} p(x|c)\sum_{a\in\mathcal{A}}\pi_0(a|x)\left(w(a,x)^2 - w(a,c)^2\right)\mathbb{E}_{p(r|a,c,x)}\left[r^2\right]dx\right]\\
&= \mathbb{E}_{p(c)}\left[\sum_{a\in\mathcal{A}}\int_{\mathcal{X}} p(x|c)\pi_0(a|x)\left(w(a,x)^2 - w(a,c)^2\right)\mathbb{E}_{p(r|a,c,x)}\left[r^2\right]dx\right]\\
&= \mathbb{E}_{p(c)}\left[\sum_{a\in\mathcal{A}}\int_{\mathcal{X}} \frac{\pi_0(x|a,c)p(a|c,\pi_0)}{\cancel{\pi_0(a|x)}}\cancel{\pi_0(a|x)}\left(w(a,x)^2 - w(a,c)^2\right)\mathbb{E}_{p(r|a,c,x)}\left[r^2\right]dx\right] \qquad (24)\\
&= \mathbb{E}_{p(c)}\left[\sum_{a\in\mathcal{A}} p(a|c,\pi_0)\int_{\mathcal{X}}\pi_0(x|a,c)\left(w(a,x)^2 - w(a,c)^2\right)\mathbb{E}_{p(r|a,c,x}\left[r^2\right]dx\right]\\
&= \mathbb{E}_{p(c)p(a|c,\pi_0)}\left[\left(\int_X \pi_0(x|a,c)w(a,x)^2dx - w(a,c)^2\cancelto{1}{\int_{\mathcal{X}}\pi_0(x|a,c)dx}\right)\mathbb{E}_{p(r|a,c,x}\left[r^2\right]\right]\\
&= \mathbb{E}_{p(c)p(a|c,\pi_0)}\left[\left(\mathbb{E}_{\pi_0(x|a,c)}\left[w(a,x)^2\right] - \mathbb{E}_{\pi_0(x|a,c)}\left[w(a,x)\right]^2\right)\mathbb{E}_{p(r|a,c,x}\left[r^2\right]\right] \qquad (25)\\
&= \mathbb{E}_{p(c)p(a|c,\pi_0)}\left[\mathbb{V}_{\pi_0(x|a,c)}\left[w(a,x)\right]\mathbb{E}_{p(r|a,c,x)}\left[r^2\right]\right] \geq 0
\end{aligned}
$$

Note in Equation 22 we used that the samples in $\mathcal{D}$ are i.i.d, in particular the linearity of variance under this condition. The cancellation of terms in Equation 23 results from IPS and CHIPS being unbiased under Assumptions 3.1 and 3.2. In Equation 24 we used that $\pi_0(x|a,c) = \frac{p(x|c)\pi_0(a|x)}{p(a|c,\pi_0)}$, while Equation 25 uses Lemma A.1.

A.8 Proposition 3.12

The first thing we need to note is that CHIPS and MIPS are in different spaces regarding the contextual bandits generating process. MIPS assumes the existence of an action embedding space $e \in \mathcal{E} \subseteq \mathbb{R}^{d_e}$ and CHIPS assumes the existence of a partition of the context space $\mathcal{C} := \{\mathcal{C}_i\}_{i=1}^{K}$ with $\mathcal{C}_i \subset \mathcal{X}$ and $c_i \cap c_j = \varnothing$. For joining this spaces, we assume that given a policy π, at every iteration of the data generation process, apart from the classical context ($x \in \mathcal{X}$), action ($a \in \mathcal{A}$) and reward ($r \in [0, r_{max}] \subset \mathbb{R}$), we observe a cluster $c \sim p(c \mid x)$ and an action embedding $e \sim p(e \mid a, c, x)$. Given a policy π the policy value $V(\pi)$ equation can be then refined to:

$$
\begin{aligned}
V(\pi) &:= \mathbb{E}_{p(x)p(c|x)\pi(a|x)p(e|a,c,x)p(r|e,a,c,x)[r]}\\
&= \mathbb{E}_{p(x)p(c|x)\pi(a|x)p(e|a,c,x)q(e,a,c,x)}
\end{aligned}
$$

Here $q(e, a, c, x) := \mathbb{E}_{p(r|e,a,c,x)}[r]$. Note that as in MIPS and CHIPS case, the refinement does not contradict the classical policy value definition.

We also need to refine $p(a \mid c, \pi)$ (from CHIPS) and $p(e \mid a, \pi)$ (from MIPS) in the joint space:

$$
\begin{aligned}
p(a \mid c, \pi) &= \sum_{e\in\varepsilon}\int_{\chi} p(e \mid a, c, x)p(x \mid c)\pi(a \mid x)\\
p(e \mid x, \pi) &:= \sum_{c\in\mathcal{C}}\sum_{a\in\mathcal{A}} p(e \mid a, c, x)p(c \mid x)\pi(a \mid x)
\end{aligned}
$$

It is important to note that after joining the context space, to make a fair comparison between MIPS and CHIPS, there are some dependencies that we want to eliminate to prevent information from passing between variables that were not originally in the definition of MIPS and CHIPS. In particular, for CHIPS, we eliminate the dependency of the cluster with respect to the embedding given the context and the action (i.e., $c \perp e \mid (x, a)$), and for MIPS, the dependency of the action with respect to the cluster given the embedding and the context (i.e., $a \perp c \mid (x, e)$). From Proposition 3.11 we know that $\mathbb{V}_{\mathcal{D}}\left[\hat{V}_{\text{IPS}}(\pi;\mathcal{D})\right] \geq \mathbb{V}_{\mathcal{D}}\left[\hat{V}_{\text{CHIPS}}(\pi;\mathcal{D})\right]$ and from MIPS Theorem 3.6 we know that $\mathbb{V}_{\mathcal{D}}\left[\hat{V}_{\text{IPS}}(\pi;\mathcal{D})\right] \geq \mathbb{V}_{\mathcal{D}}\left[\hat{V}_{\text{MIPS}}(\pi;\mathcal{D})\right]$. Therefore, we need to make a comparison between $\mathbb{V}_{\mathcal{D}}\left[\hat{V}_{\text{MIPS}}(\pi;\mathcal{D})\right]$ and $\mathbb{V}_{\mathcal{D}}\left[\hat{V}_{\text{CHIPS}}(\pi;\mathcal{D})\right]$.

To follow the structure of Proposition 3.11, we are going to assume that Assumptions 3.1 and 3.2 hold as well as their counterparts from MIPS. The following identities hold under these conditions:

$$
\begin{aligned}
p(x \mid c)\pi(a \mid x) &= \frac{p(e \mid x, \pi)p(c \mid x, a)\pi_0(a \mid x)p(x)}{p(e \mid x, \pi_0)p(c)} \\
p(e \mid x, a, c) &= \frac{p(x \mid e, a, c)p(e \mid a, c)p\,(a \mid c, \pi_0)}{p(c \mid x, a)\pi_0(a \mid x)p(x)}
\end{aligned}
$$

Now, under these conditions, we need a relation between the weights of MIPS and CHIPS:

$$
\begin{aligned}
\omega(a,c)^2 &= \frac{p(a \mid c, \pi)}{p\,(a \mid c, \pi_0)} \\
&= \frac{\int_{\mathcal{X}} p(x \mid c)\sum_{e\in\mathcal{E}} \pi(a \mid x)p(e \mid c, a, x)}{p\,(a \mid c, \pi_0)} \\
&= \frac{\int_x \sum_{e\in\mathcal{E}} w(e,x)p(x \mid e, a, c)p(e|a, c)p(a \mid c, \pi_0)}{p(a \mid c, \pi_0)} \\
&= \sum_{e\in\varepsilon} p(e|a,c)\int_x p(x \mid e, a, c)\omega(e,x) \\
&= \mathbb{E}_{p(e|a,c)p(x|e,a,c)}[\omega(e,x)]
\end{aligned}
$$

Therefore the scaled difference in variance can be expressed as:

$$
\begin{aligned}
&N\left(\mathbb{V}_{\mathcal{D}}\left[\hat{V}_{\text{MIPS}}(\pi;\mathcal{D})\right] - \mathbb{V}_{\mathcal{D}}\left[\hat{V}_{\text{CHIPS}}(\pi;\mathcal{D})\right]\right) \\
&= N\left(\mathbb{V}_{\mathcal{D}}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\pi\,(e_i \mid x_i)}{\pi_0\,(e_i \mid x_i)}r_i\right] - \mathbb{V}_{\mathcal{D}}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{p\,(a_i \mid c_i, \pi)}{p\,(a_i \mid c_i, \pi_0)}r_i\right]\right) \\
&= \mathbb{V}_{\mathcal{D}}\left[\frac{\pi(e \mid x)}{\pi_0(e \mid x)}r\right] - \mathbb{V}_{\mathcal{D}}\left[\frac{p(a \mid c, \pi)}{p\,(a \mid c, \pi_0)}r\right] \\
&= (\mathbb{E}_{\mathcal{D}}\left[\omega(e,x)^2 r^2\right] - \underbrace{\mathbb{E}_{\mathcal{D}}[\omega(e,x)r]^2)}_{V(\pi)} - (\mathbb{E}_{\mathcal{D}}\left[\omega(a,c)^2 r^2\right] - \underbrace{\mathbb{E}_{\mathcal{D}}[\omega(a,c)r]^2)}_{V(\pi)} \\
&= \mathbb{E}_{p(c)}\left[\int_{\mathcal{X}} p(x \mid c)\sum_{a\in\mathcal{A}}\pi_0(a \mid x)\sum_{e\in\mathcal{E}} p(e \mid a, c, x)\left(\omega(e,x)^2 - \omega(a,c)^2\right)\mathbb{E}_{p(r|a,c)}\left[r^2\right]dx\right] \\
&= \mathbb{E}_{p(c)}\left[\sum_{a\in A} p\,(a \mid c, \pi_0)\sum_{e\in\mathcal{E}} p(e \mid a, c)\int_{\mathcal{X}} p(x \mid e, a, c)\left(\omega(e,x)^2 - \omega^2(a,c)\right)\mathbb{E}_{p(r|a,c)}\left[r^2\right]dx\right] \\
&= \mathbb{E}_{p(c)p(a|c,\pi_0)}\left[\mathbb{E}_{p(r|a,c)}\left[r^2\right]\left(\mathbb{E}_{p(e|a,c)p(x|e,a,c)^2}\left[\omega(e,x)^2\right]\right)\right] \\
&\qquad - \mathbb{E}_{p(c)p(a|c,\pi_0)}\left[\mathbb{E}_{p(r|a,c)}\left[r^2\right]\left(\omega(a,c)^2\sum_{e\in\mathcal{E}} p(e \mid a, c)\int_{\mathcal{X}} p(x \mid e, a, c)dx\right)\right] \\
&= \mathbb{E}_{p(c)p(a|c,\pi_0)}\left[\mathbb{E}_{p(r|a,c)}\left[r^2\right]\left(\mathbb{E}_{p(e|a,c)p(x|e,a,c)}\left[\omega(e,x)^2\right] - \mathbb{E}_{p(e|a,c)p(x|e,a,c)}\left[\omega(e,x)\right]^2\right)\right] \\
&= \mathbb{E}_{p(c)p(a|c,\pi_0)}\left[\mathbb{E}_{p(r|a,c)}\left[r^2\right]\mathbb{V}_{p(e|a,c)p(x|e,a,c)}[\omega(e,x)]\right] \geq 0
\end{aligned}
$$

This implies that under Assumptions 3.1 and 3.2 (and their counterparts in MIPS), the variance of CHIPS is lower than the variance of MIPS, proving the proposition.

B Reward Estimates Derivation

B.1 MAP

From the setting in Subsection 3.2 we denote $R_c \coloneqq \{r_i\}_{i=1}^{M}$ as the rewards observed in cluster c from the logging data[2]. We consider R_c as independent trials of a Bernoulli random variable with parameter θ (i.e., $R_c \overset{i.i.d.}{\sim} \text{Ber}(\theta)$). Therefore, we have that the likelihood can be expressed as:

$$\begin{aligned} p(R_c|\theta) &= \prod_{i=1}^{M} p(r_i|\theta) \\ &= \prod_{i=1}^{M} \theta^{r_i}(1-\theta)^{1-r_i} \\ &= \theta^{\sum_{i=1}^{M} r_i}(1-\theta)^{M-\sum_{i=1}^{M}} \end{aligned}$$

Using a Beta distribution as a prior we have that:

$$p(\theta) = \text{Beta}(\theta|\alpha,\hat{\beta}) = \frac{1}{\mathcal{B}(\alpha,\hat{\beta})}\theta^{\alpha-1}(1-\theta)^{\hat{\beta}-1}$$

Where $\mathcal{B}(\alpha,\hat{\beta}) = \frac{\Gamma(\alpha)\Gamma(\hat{\beta})}{\Gamma(\alpha+\hat{\beta})}$ and $\Gamma(\cdot)$ is the Gamma function. The posterior probability can then be expressed as:

$$\begin{aligned} p(\theta|R_c) &\propto p(R_c|\theta)p(\theta) \\ &\propto \theta^{\sum_{i=1}^{M} r_i}(1-\theta)^{M-\sum_{i=1}^{M}} \frac{1}{\mathcal{B}(\alpha,\hat{\beta})}\theta^{\alpha-1}(1-\theta)^{\hat{\beta}-1} \\ &\propto \theta^{\alpha-1+\sum_{i=1}^{M} r_i}(1-\theta)^{\hat{\beta}-1+M-\sum_{i=1}^{M} r_i} \\ &\propto \text{Beta}\left(\theta \,\middle|\, \alpha + \sum_{i=1}^{M} r_i,\ \hat{\beta} + M - \sum_{i=1}^{M} r_i\right) \end{aligned}$$

The MAP estimator of θ is the mode of the resulting Beta distribution, i.e.

$$\hat{\theta}_{\text{MAP}} = \frac{(\alpha-1)+\sum_{i=1}^{M} r_i}{\alpha+\hat{\beta}+M-2}$$

B.2 ML

Using the same setting as in the previous section ($R_c \overset{i.i.d.}{\sim} \text{Ber}(\theta)$) we have that the maximum likelihood estimation can be expressed as

$$\begin{aligned} \hat{\theta}_{\text{ML}} &= \underset{\theta\in\Theta}{\arg\max}\left\{\prod_{i=1}^{M} \theta^{r_i}(1-\theta)^{1-r_i}\right\} \\ &= \underset{\theta\in\Theta}{\arg\max}\left\{\underbrace{\log(\theta)\cdot\sum_{i=1}^{M} r_i + \log((1-\theta))\cdot\sum_{i=1}^{M}(1-r_i)}_{l(\theta)}\right\} \end{aligned}$$

We now search for local maxima by setting the differential to 0:

$$\frac{\partial l(\theta)}{\partial\theta} = 0 \implies \frac{\sum_{i=1}^{M} r_i}{\theta} + \frac{\sum_{i=1}^{M}(1-r_i)}{(1-\theta)} = 0$$

[2]Here we refer to the already transformed version using clusters. See definition of τ in Subsection 3.2

$$\implies \sum_{i=1}^{M} r_i - \theta \sum_{i=1}^{M} r_i = \theta \sum_{i=1}^{M} (1 - r_i)$$

$$\implies \hat{\theta}_{\text{ML}} = \frac{1}{M} \sum_{i=1}^{M} r_i$$

C Experimental Parameters and Hardware

Parameter	Value	Description
c_{exp}	10	Radius of the n-dimensional ball for context space generation.
c_{rad}	1	Cluster generation radius.
d_x	2	Dimension of context vectors.
x_{num}	1.000	No. of different context vectors in the experiment.
a_{num}	10	No. of actions in the experiment.
c_{num}	10	No. of clusters in the experiment.
$n_{samples}$	50.000	No. of logged samples to use in the experiment.
emp_{c_num}	100	No. of clusters to use empirically by the clustering method.
e_{len}	1.000.000	No. of samples extracted from the dataset for the evaluation policy
b_{len}	1.000.000	No. of samples extracted from the dataset for the evaluation policy
σ	0.2	Context-specific behaviour deviation from cluster behaviour.
β	-1	Deviation between evaluation and logging policies.
α	20	Parameter from beta distribution in Bayesian inference
$\hat{\beta}$	20	Parameter from beta distribution in Bayesian inference

Table 1: Parameters used in the basic configuration for experiments for generation and estimation.

CPU	AMD Ryzen Threadripper PRO 3975WX
RAM	256 GB
Cores	64
GPU	2x Nvidia A100 160GB

Table 2: Specifications of the machine in which the experiments were executed.

D Additional Experiments

D.1 Synthetic Experiments

Number of actions. From the fixed basic configuration that uses 100 clusters for CHIPS' estimates, we observe a progressive deterioration in the estimator capabilities when increasing the number of actions (see Figure 4). We theorize that this behaviour might be a consequence of the violation of Assumption 3.1 when trying to group contexts using an excessive number of clusters in a large action space, resulting in deficient actions inside the clusters. This problem can be mitigated by decreasing the number of clusters used in the clustering method for the CHIPS estimation (see Figure 12 (left)).

Number of samples. We observe an approximation to the performance of IPS as we increase the number of samples in the logged data that we identify as an effect of reducing the number of observed deficient action-context pairs in IPS, converging to an unbiased estimator under Assumption 2.1 (see Figure 1 (right)). In this case, the clustering effects under CHIPS become less noticeable according to Corollary 3.9 since $\mathcal{U}(c, x, \pi_0) \setminus \mathcal{U}(c, \pi_0) \to \varnothing$. It is worth mentioning that increasing the number of clusters when enough samples are available, as well as reducing it in the opposite case, can improve the performance of the CHIPS estimates, as shown in Figure 12 (right).

Cluster radius. Increasing the cluster radius in the generation process affects the separability of the cluster space and complicates the partitioning in clusters complying with Assumption 3.2. In this case, we could find significant differences in context behaviour for both actions and rewards within a cluster, resulting in increased bias from the empirical approximations. Therefore, we observe a convergence to IPS' performance as cluster radius increases since the context space becomes less separable (see Figure 6).

Sigma. Increasing context-specific noise in the generation process produces a similar effect as in the cluster radius case. In particular, the larger the noise, the more common it is to observe inconsistent behaviour in actions and rewards for contexts within a cluster, complicating the approximation of a homogeneous cluster-wise behaviour and resulting again in a bias increase (see Figure 8).

Alpha (prior). In this experiment, we vary the alpha parameter of the Beta prior maintaining all other settings fixed. Like in the number of clusters case, we observe a similar v-shaped graph indicating that, as expected from the previous β analysis (see Section 4.1.1), the CHIPS (MAP) estimator is sensitive to the prior. In particular, lower values push the expected reward of each cluster to the ML's estimate, while higher values push it to the prior's expected value, decreasing performance in both cases (see Figure 1 center). For different values of distributional shift (β), the optimal value will depend on the *resistance* MAP offers to converge to the ML estimate, favouring lower values as β becomes larger (see Figure 13).

Clustering Method. In this experiment, we evaluate the performance of the CHIPS (MAP) estimator using different clustering methods while varying the clustering radius in the synthetic generation process. In Figure 10, we observe that using Mean Shift (Comaniciu & Meer, 2002) or Bayesian Gaussian Mixture (Bishop, 2006; Attias, 1999; Blei & Jordan, 2006) fails to separate the context space resulting in the same performance as IPS. DBSCAN (Ester et al., 1996) mitigates IPS' increase in mean squared error when the context space is easier to separate (i.e., lower radii values) but converges to IPS when the context is complicated to separate (i.e., higher radii values). Affinity Propagation (Frey & Dueck, 2007) follows a similar behaviour to DBSCAN but still offers some improvement with respect to IPS when the space is difficult to separate. OPTICS Ankerst et al. (1999) makes a general improvement to the Affinity Propagation performance, especially noticeable when the context space is separable. MiniBatch K-Means (Sculley, 2010), Gaussian Mixture (Everitt, 1996), Birch (Zhang et al., 1996), Spectral Clustering (Shi & Malik, 2000), and Agglomerative Clustering (Ward, 1963) have similar performance, outperforming Affinity Propagation for the separable case. We also note a general upward tendency in mean squared error for every clustering method as the space becomes more complicated to separate.

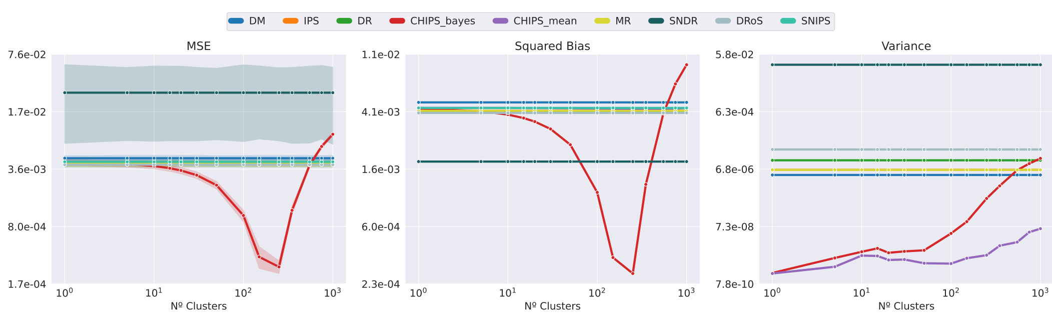


Figure 3: From left to right, MSE, Bias, and Variance of the CHIPS estimator compared to baselines while varying the number of clusters.

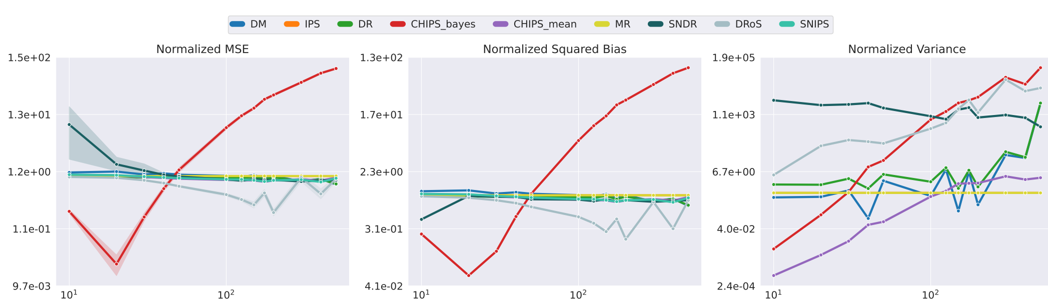


Figure 4: From left to right, MSE, Bias, and Variance of the CHIPS estimator compared to baselines while varying the number of actions.

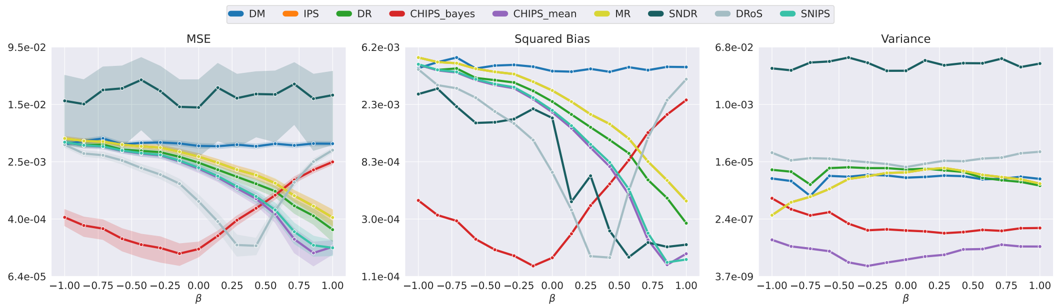


Figure 5: From left to right, MSE, Bias, and Variance of the CHIPS estimator compared to baselines while varying β values.[3]

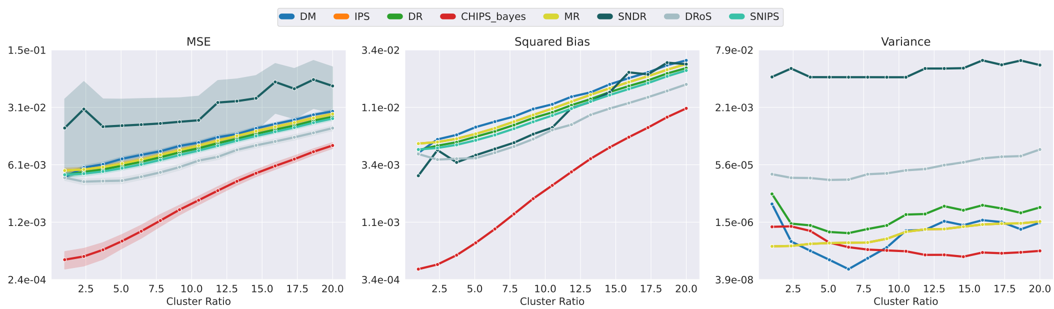


Figure 6: From left to right, MSE, Bias, and Variance of the CHIPS estimator compared to baselines while varying the radius of the clusters generated.

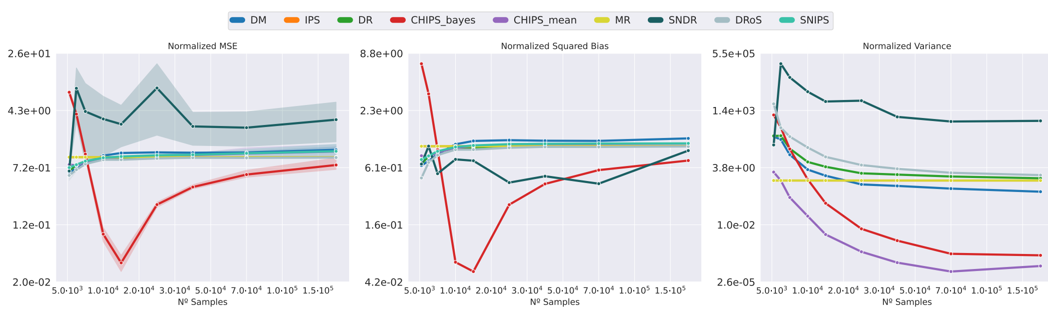


Figure 7: From left to right, MSE, Bias, and Variance of the CHIPS estimator compared to baselines while varying the number of samples provided from the logging policy.

D.1.1 Bi-parametric variations

The experiments varying single parameters described in the previous section indicate that increasing the number of actions in a fixed configuration progressively deteriorates CHIPS' performance. This behaviour is expected since the larger the action space, the more likely it is to incur in a situation in which Assumption 3.1 does not hold with a fixed number of clusters. In this situation, we found that reducing the number of clusters can mitigate the performance decay by pooling information from broader contexts clusters while increasing it could be beneficial in reduced action spaces (see Figure 12 (a)). Similarly, the number of samples from the logging policy also conditions how significant the performance gap between CHIPS and IPS is. In particular, the higher the number of samples, the more beneficial it is to use a higher number of clusters to try to obtain a more detailed

[3]In this case we used a slightly different version of the configuration settings to make a more challenging environment in which we use 10.000 samples and consequently reduce the number of empirical cluster estimation to 30 to easily assess the role that similarity of logging and evaluation policies play in CHIPS capabilities.

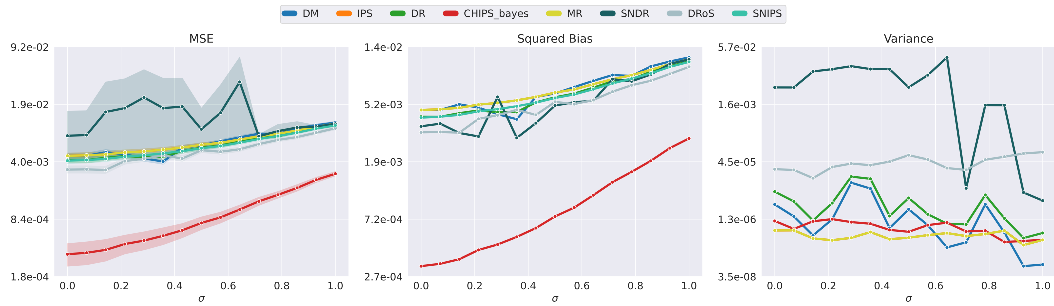


Figure 8: From left to right, MSE, Bias, and Variance of the CHIPS estimator compared to baselines while varying the context-specific noise σ.

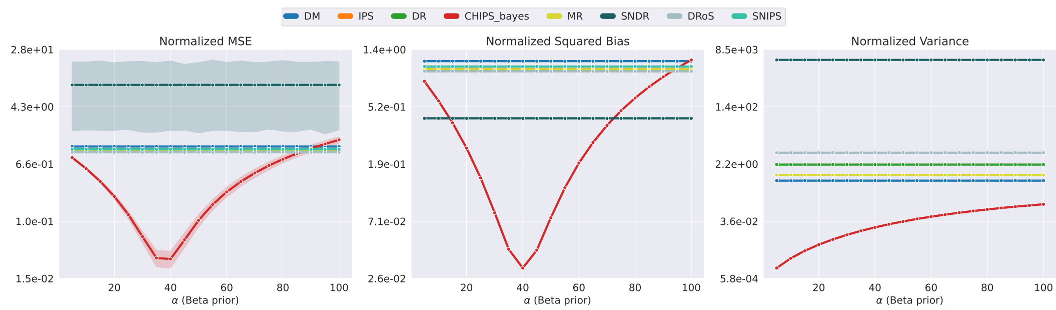


Figure 9: From left to right, MSE, Bias, and Variance of the CHIPS estimator compared to baselines while varying the α parameter.

partition structure of the context space, while a reduced number of clusters has an edge on few-sample cases (see Figure 12 (b)).

We also study the effect of varying the α parameter in CHIPS' (MAP) Beta prior, using different values of the distributional shift between policies (β). In Figure 13, we observe that mid values of α (30-50) offer better performance when there is a considerable distributional shift between logging and evaluation policy (i.e., $\beta \approx -1$) since the expected reward per cluster is pushed towards the prior's expectation, creating some resistance from converging to the average observed rewards (i.e., mitigating the reward misspecification existing under this conditions). As the distributional gap closes, lower values of α are more favourable since the samples observed per cluster are better representatives of the real expected reward. However, higher values for α (80-100) result in excessive resistance that deteriorates CHIPS' performance. It is also worth mentioning that as the distributional gap closes, CHIPS (MAP) loses its advantage with respect to IPS since the logging and evaluation policies are closer, and the ML estimates would offer better results, as previously shown in Figure 1.

D.2 Real Experiments

D.3 MAP vs ML

In this section, we analyze the reason behind the jump in performance using the CHIPS estimator with the MAP estimate for the expected reward per cluster. For this purpose, we have conducted two experiments, one in the synthetic dataset and the other in the real dataset. For the synthetic experiment, given a distributional shift value β, we select the most relevant context-action pair (x^*, a^*) under the evaluation policy π (i.e., $(x^*, a^*) = \arg\max_{(x,a)\in\mathcal{X}\times\mathcal{A}} \pi(a|x)$). Then we analyze the mean squared error of the expected reward (given x^*) estimations made by CHIPS (MAP) (i.e., $w(a^*, c^*)\hat{r}_{\text{bayes}}(a^*, c^*)$) and CHIPS (ML) (i.e., $w(a^*, c^*)\hat{r}_{\text{mean}}(a^*, c^*)$) w.r.t IPS (where c^* is the cluster associated with x^*). We also compute the number of observations in c^* in which action a^* was selected. This process is repeated 100 times with different policies generated under different random seeds, and the results for the number of samples per cluster and squared errors are averaged. We repeat this for ten different values of β ranging from -1 to 1 and represent the moving averages for

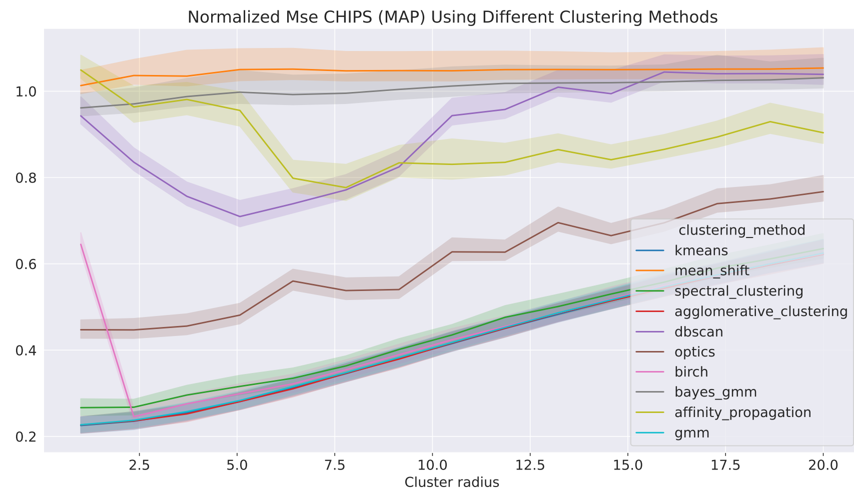


Figure 10: Normalized MSE of CHIPS (MAP) using different clustering methods with respect to IPS.

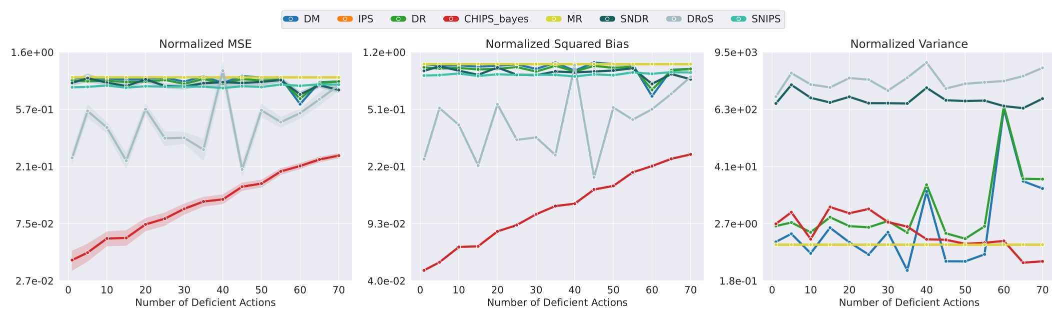
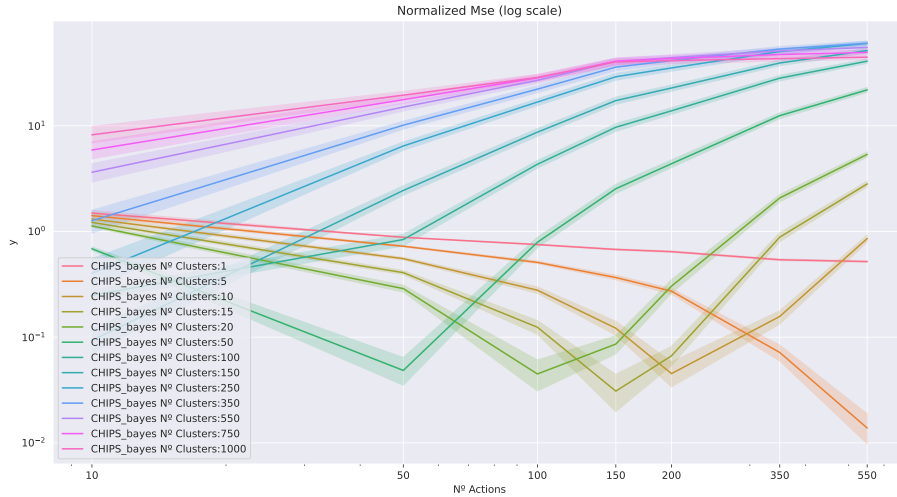


Figure 11: From left to right, MSE, Bias, and Variance of the CHIPS estimator compared to baselines while varying the number of deficient actions.

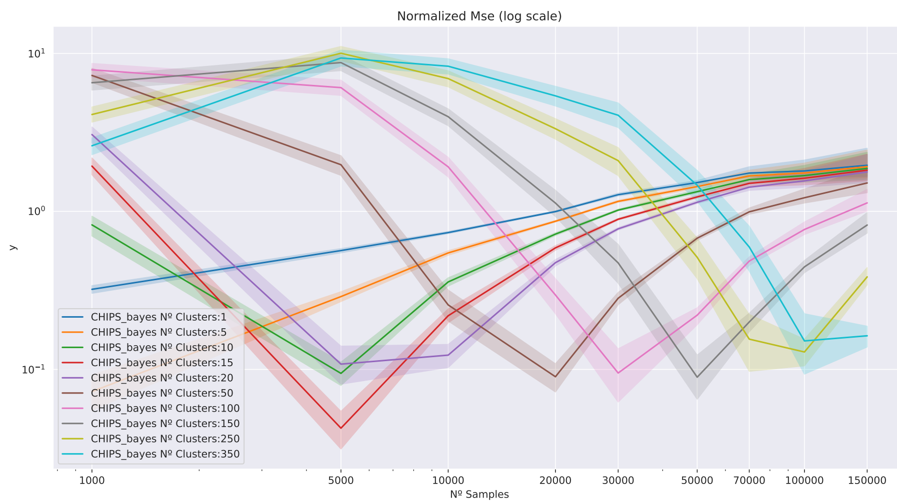
relative squared errors and samples in Figure 16. We observe that the number of samples per cluster increases with β as both policies become closer. This increase in the number of samples makes the ML estimates progressively more accurate since the extra samples push the estimated expected value to the real expected value. For lower values of β, when the gap between policies is more significant, although some samples are available in the cluster, the values for the rewards observed on them are non-informative of the real expected value (hence the difficulty of ML to make an accurate estimation and the difference between MAP and ML for misspecified reward settings as depicted in Figure 1). For the real dataset, we follow a similar procedure, but instead of the most relevant context-action pair, we select the top 15 and compare the conditional expected reward estimates MSE with respect to IPS' (see Figure 15). Since the logging policy for this dataset is uniform, the distributional shift between the logging and evaluation policies is not as significant as the one presented in the base configuration of the synthetic dataset ($\beta = -1$). In practice, this means that the CHIPS estimation of the expected reward per cluster using ML is more accurate than in the synthetic dataset but still far from the performance jump of the CHIPS estimate using MAP, as we would expect from the results in Figure 2.

D.4 Choosing α in Arbitrary Problems

In Figure 14 (b), we observe that the hyperparameters of the MAP estimation process can heavily impact the performance of the method. As previously discussed in Appendix D.1.1, MAP hyperparameters control the resistance with which the expected reward per cluster is *pulled* towards the prior's expectation. This resistance is particularly noticeable in smaller size clusters, in which estimating a reward based on observations alone is much more challenging. Since in these clusters the partitioning method cannot ensure high homogeneity at reward level, in our experimentation we decided to use a non-informative prior (i.e., $\alpha = \hat{\beta}$), to mitigate possible violations of Assumption 3.2 and reward



(a) Normalized performance of CHIPS (MAP) with respect to IPS using different number of clusters and actions.



(b) Normalized performance of CHIPS (MAP) with respect to IPS using different number of clusters and logging samples.

Figure 12: Bi parametric experiments results using different number of clusters for analyzing CHIPS capabilities when increasing actions (a) and logging samples (b).

misspecification. Intuitively, an optimal value for α under these conditions needs to balance the prior's resistance to prevent reward misspecification without incurring into creating a quasi-uniform reward estimation (excessively large values of α). In Figure 17 we explore the optimal value of α for a given average number of datapoints per cluster-action. As expected, for small size clusters, lower values of $alpha$ are favoured since the pull towards the prior's expectation is soft, while on bigger clusters, the value of α (and consequently the resistance) needs to grow to effectively control reward misspecification (otherwise the expected reward value would be pulled towards the value of the observed samples).

To choose the value of α in an arbitrary problem, we propose the following selection process:

1. Determine the number of clusters to use depending on the number of clusters (reference in Figure 12 (a).
2. Partition the context space $\mathcal{X}$ in clusters $c_1, c_2, ..., c_n$.
3. Generate synthetic data $\hat{X_{\text{ev}}}$ using $\mathcal{X}_{\text{train}}$ and π_e.

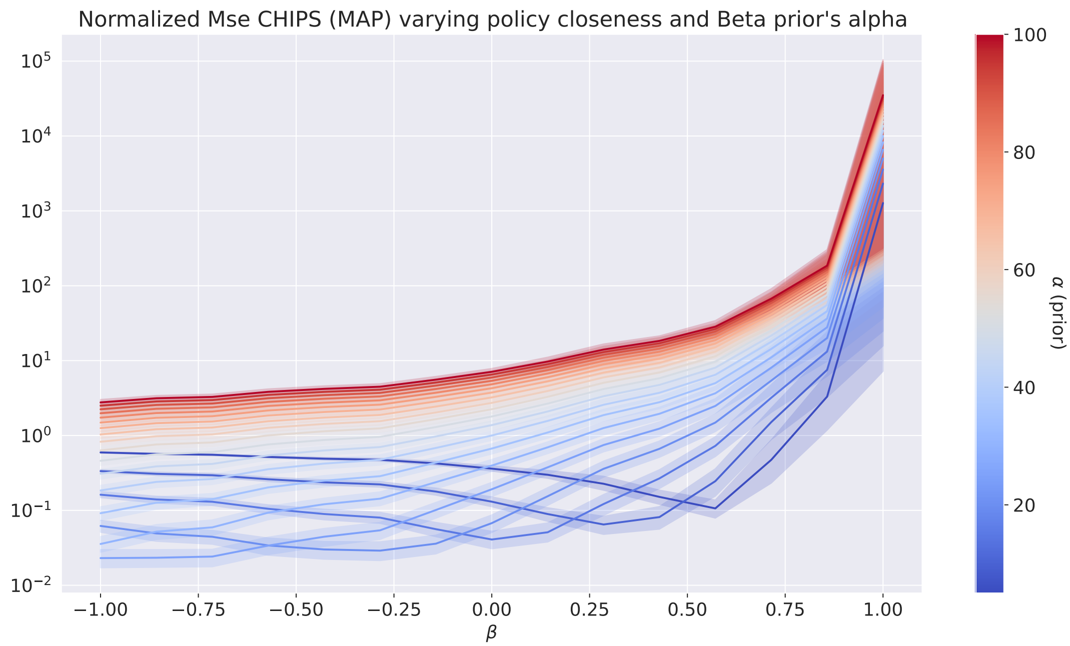


Figure 13: Normalized performance of CHIPS (MAP) with respect to IPS using different values for the α parameter in the Beta prior and distributional shift between logging and evaluation policies (β).

4. Estimate number of average data points per cluster-action from $\hat{X}_{\text{ev}}$.
5. Choose α from the reference in Figure 17.

For testing this selection process, following the experimental protocol of Taufiq et al. (2023), we transform five UCI datasets Dua & Graff (2017), MNIST Deng (2012) , and CIFAR-100 Krizhevsky et al. (2009) from multi-class classification problems into contextual bandits data Dudík et al. (2011). The results (averaged 50 times) in Figure 18 show a consistent improvement with respect to existing methods, empirically proving the effectiveness of the α selection process.

Additionally, we perform an alternative experiment using the real dataset, in which instead of fixing α and vary the number of clusters according to the reference in Figure 12 (b) with 50000, 100000 and 500000 samples (see Figure 2, we follow the α selection process, fix the number of clusters and increase the value of α according to Figure 17. In Figure 19 we observe equivalent results as in our previous experiment confirming the equivalence of using a reference for the number of samples and varying the number of clusters with a fixed value for α, or varying α with a fixed number of clusters obtained by using a reference for the number of actions.

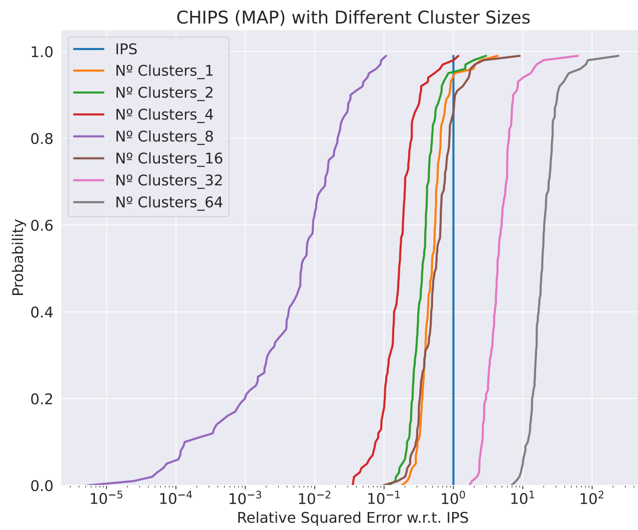
E Cluster Structure in Datasets

The generated synthetic dataset ensures that the expected reward inside a cluster is similar and that the best possible action is usually the same for all the context within the cluster (see Figure 20), mimicking real-world settings like e-commerce in which we can expect similar behaviour for close contexts (see Figure 21)

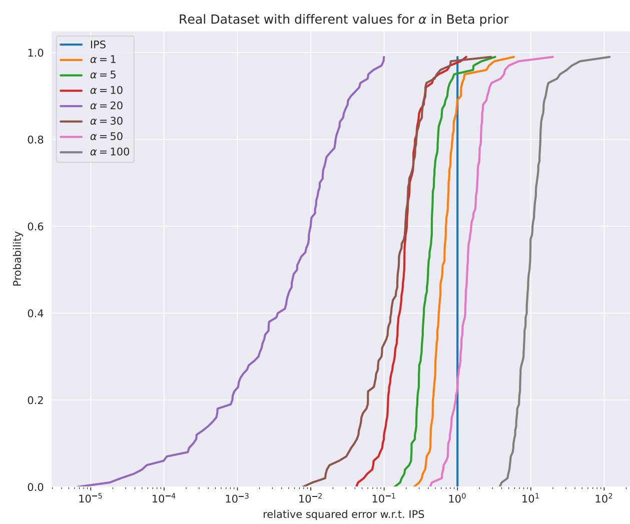
F Experimental Protocol

For evaluation in the real dataset, we follow Saito & Joachims (2022) protocol to evaluate estimators' accuracy given two sources of data. Given a logging policy π, a dataset collected under it $\mathcal{D}$, a logging policy π_0, and the dataset collected under it $\mathcal{D}_0$, we follow the following procedure:

1. Extract n independent bootstrap samples with replacement from the logging dataset $\mathcal{D}_0^* := \{(x_i, a_i, r_i)\}_{i=1}^n$.



(a) ECDFs of CHIPS (MAP) using different number of clusters in the real dataset.



(b) ECDFs of CHIPS (MAP) using different values of α for the Beta prior in the real dataset.

Figure 14: Additional experiments varying the number of clusters and the α parameter in the Beta prior for CHIPS (MAP) in the real dataset (using 100000 samples).

2. Estimate the policy value of π using the sample $\mathcal{D}_0^*$. We denote this estimate as $\hat{V}(\pi; \mathcal{D}_0^*)$.
3. Compute the relative mean squared error with respect to IPS:

$$\mathcal{Z}(\hat{V}, \mathcal{D}_0^*) = \frac{\left(V(\pi) - \hat{V}(\pi; \mathcal{D}_0^*)\right)^2}{\left(V(\pi) - \hat{V}_{\text{IPS}}(\pi; \mathcal{D}_0^*)\right)^2}$$

   Where $V(\pi) := \frac{1}{|\mathcal{D}|} \sum_{(\cdot,\cdot,r_i) \in \mathcal{D}} r_i$.
4. Repeat steps 1,2, and 3 $T = 100$ times and compute the Empirical Cumulative Distribution Function (ECDF) as:

$$\hat{F}_{\mathcal{Z}}(x) := \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\{\mathcal{Z}_t(\hat{V}, \mathcal{D}_0^*) \leq x\}$$

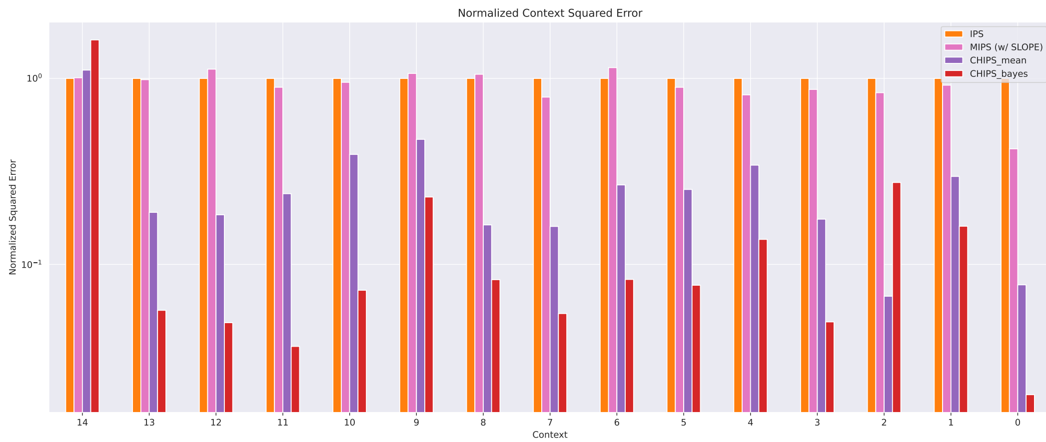


Figure 15: Normalized MSE with respect to IPS of the expected rewards for the 15 most common context-action pairs in the real logging dataset.

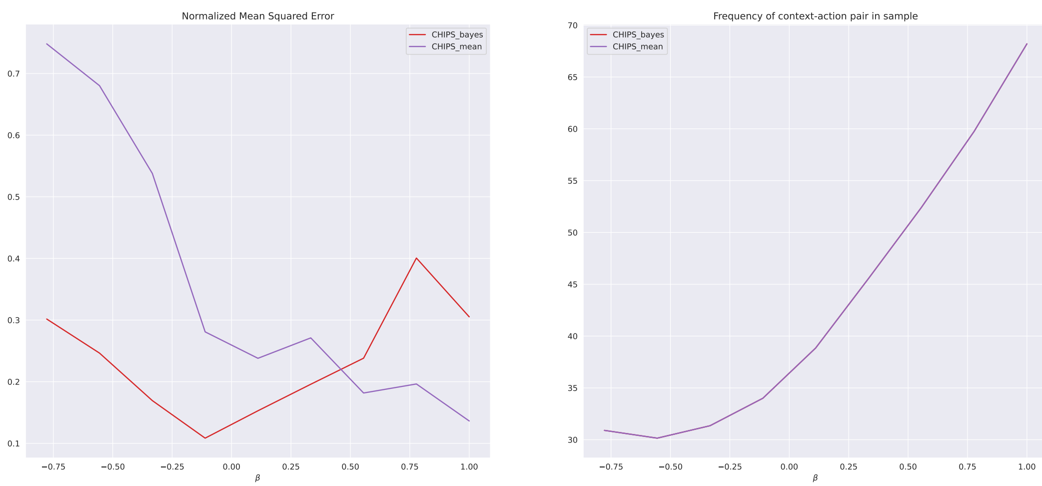


Figure 16: Normalized MSE of CHIPS with respect to IPS (left) and samples in the associated cluster (right) for the most common context-action pair in the evaluation policy while varying the distributional shift (β) in the synthetic dataset.

G Time Complexity

Algorithmically, since the CHIPS estimator can be regarded as performing the same procedure as IPS with different weights and rewards,the time complexity given n logging samples, and clustering method ξ can be expressed as:

$$\text{complexity}(\text{CHIPS}(n;\xi)) = \text{complexity}(\text{IPS}(n)) + \text{complexity}(\xi(n))$$

For example, since the time complexity of IPS is $\mathcal{O}(n)$, using DBSCAN ($\mathcal{O}(n\log n)$) as a clustering method, we would get a time complexity for CHIPS of $\mathcal{O}(n\log n)$. In our experiments, we used batch-Kmeans (Sculley, 2010) as clustering method, that has a time complexity of $\mathcal{O}(mkd_x t)$ where m is the batch size, k is the number of clusters, d_x is the dimension of the features and t is the number of iterations. In the implementation used, we fixed $m = 1024$ and $t = 100$, therefore, in this case, the time complexity of the CHIPS method is $\mathcal{O}(kd_x) + \mathcal{O}(n)$. The time complexity of the MIPS estimator can be estimated similarly as $\mathcal{O}(nd_e) + \mathcal{O}(n) = \mathcal{O}(nd_e)$, where the $\mathcal{O}(nd_e)$ term comes from the logistic regression used to estimate $\pi_0(a|x,e)$ (being e an action embedding) and d_e is the action embedding dimension. The methods using a supervised classifier (DM, DR, and MRDR) get their dominant term in time complexity from the training process of the classifier, in our case $\mathcal{O}(nds\log n)$ with s being the number of trees. In practice, this means that DM, , and MRDR will

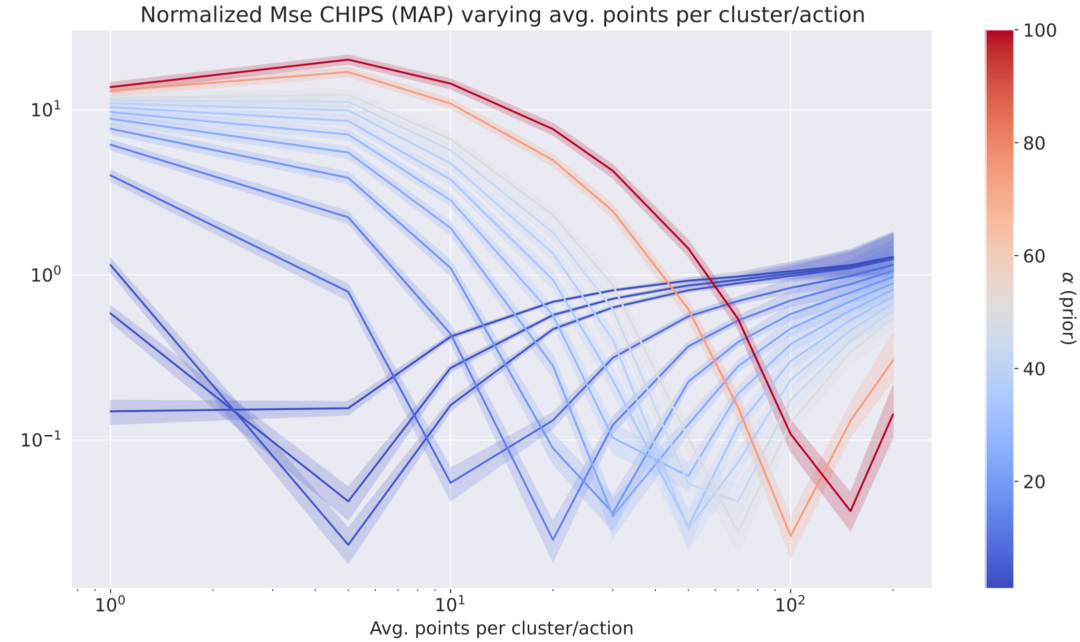


Figure 17: Normalized MSE of CHIPS (MAP) with respect to IPS using different values of α and number of expected data points per cluster-action.

Figure 18: MSE of CHIPS (MAP) using α selection policy with respect to IPS, DR, DM, MR Taufiq et al. (2023), DRosSu et al. (2020a) and SwitchDR Wang et al. (2017).

have significantly higher execution times (see Figure 22 (a)), and CHIPS will generally be faster than MIPS since $k << n$ to leverage the cluster structure, as we can appreciate in Figure 22 (b).

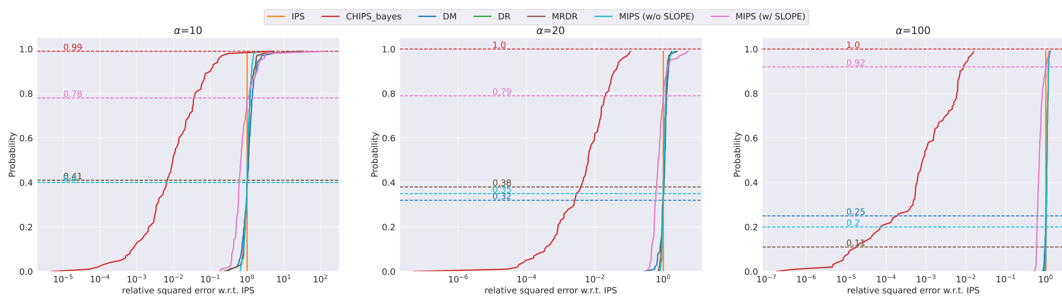


Figure 19: ECDF of the relative mean squared error with respect to IPS for the real dataset using 50000 (left), 100000 (center), and 500000 (right) logging samples and the α selection process.

(a) Expected reward per context for a sepcific action in the synthetic dataset.

(b) Action maximizing the expected reward per context in the synthetic dataset.

Figure 20: Representation of the synthetic dataset using 2-dimensional contexts.

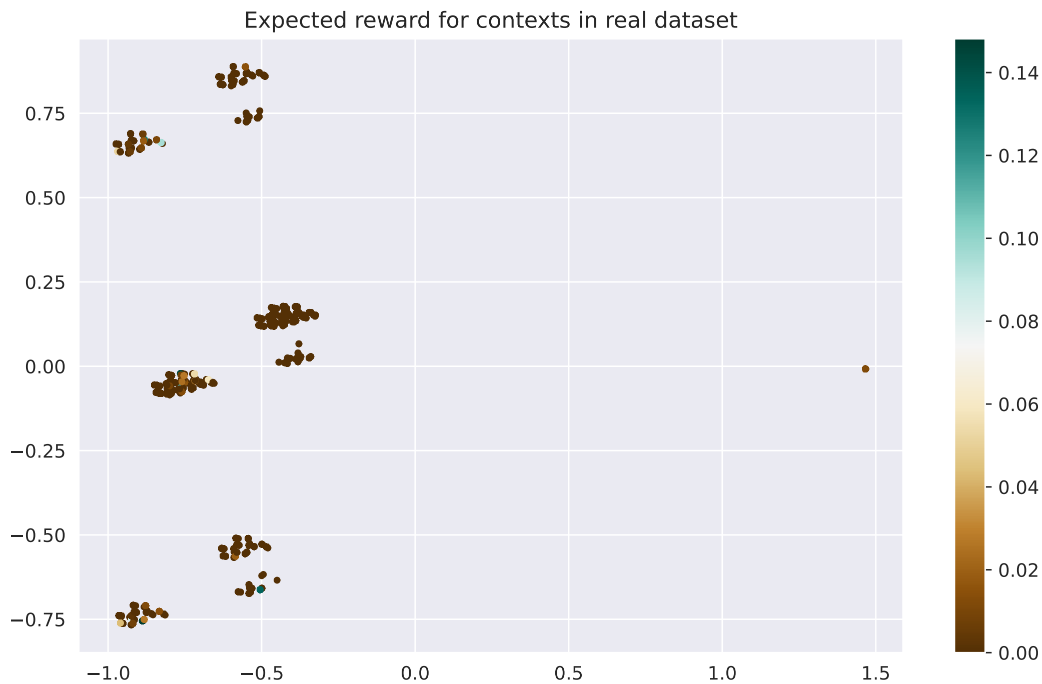
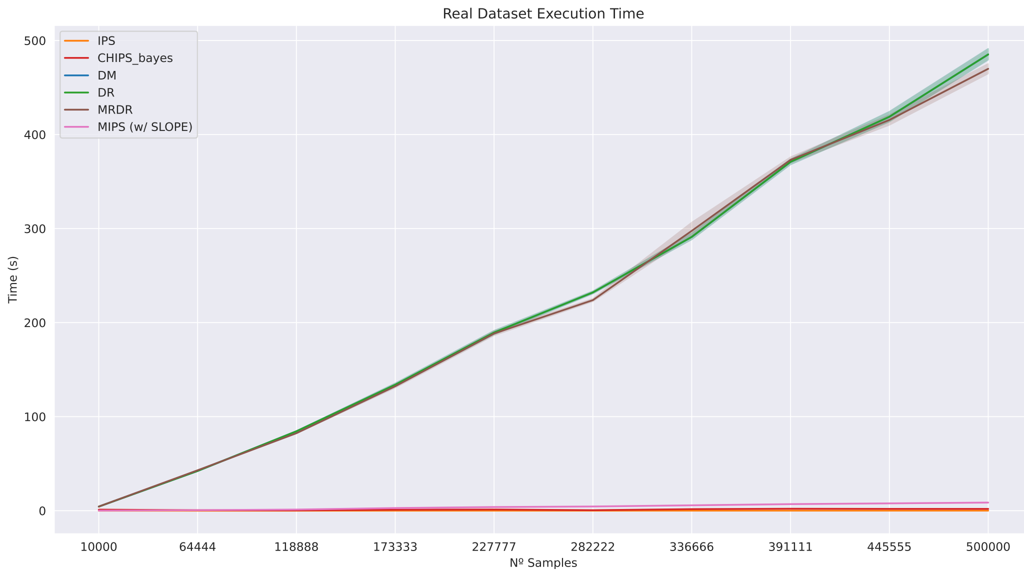
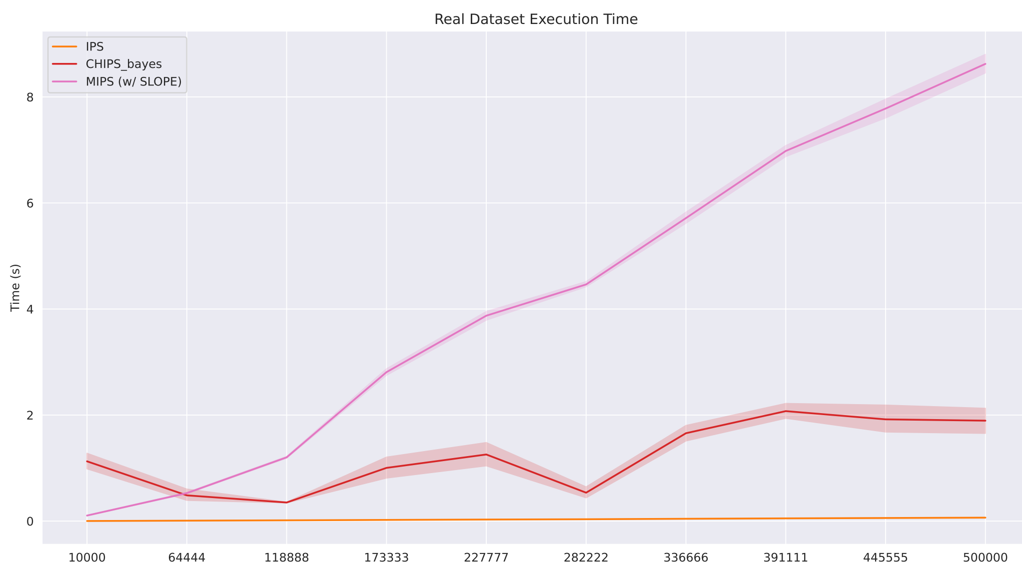


Figure 21: Expected reward in real dataset clusters projected in a 2-dimensional space using PCA.



(a) Execution times for the real dataset including DM, DR, and MRDR.



(b) Execution times for the real dataset for IPS, CHIPS, and MIPS.

Figure 22: Average execution times increasing the sample size (100 executions per sample size).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All the main claims made in the abstract and introduction are introduced in 1 where we also introduce the hypothesis about the context space that motivates our method. The scope and state of the art of the off-policy evaluation problem is analyzed in 2. The theoretical analysis of the properties of the estimator is detailed in Section 3. Finally, the improvement with respect to other estimators is backed by the evidence in 4 where we discuss the experimental protocol and results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations of the system can be found in Section 5. In 3, we analyze the estimator's properties theoretically, clearly detailing the assumptions made for this analysis, and study the bias and variance for the estimator when this assumptions does not hold. Our whole experimental protocol is detailed on Section 4, for ensuring robustness. We analyzed the system varying each parameter of the generation process individually and execute each configuration 100 times, reporting average and standard deviation bands in the graphs. Computational efficiency of our method compared with the rest can be found in Appendix G

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All the proofs for each proposition made in 3 can be found in Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The code for reproducing the experimental protocol is provided as additional material. Additionally, all the hyperparameters for the priors of the model and the generation process of the dataset can be found in Appendix C, a detailed experimental protocol for the synthetic and real datasets are detailed in Section 4 and Appendix F respectively, and an extensive study on the hyperparameters choosing for our method can be found in Appendix D.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper provides the code as supplemental material to completely reproduce the experiments detailed on the main section and appendices. A README.md file is provided with step-by-step instructions on how to configure the necessary environment through a Poetry file, download the OBD dataset, and execute the experiments.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The method for generating different logging and evaluation datasets under their respective policies can be found in Section 4. The hyperparameters chosen for the method can be found in Appendix C. The choosing of the hyperparameters is studied in depth in Appendix D.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.

- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All the 1-sigma standard deviation bands are reported in the graphs in Section 4) as well as the extra experiments in Appendix D.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer:[Yes]

   Justification: The computer resources used for executing all the experiments are detailed on Appendix C. We also provide warnings for heavy computation experiments (estimated time surpassing 2) hours in the code provided in the supplemental materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper has been carefully reviewed to comply with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The positive impact of the off-policy evaluation is discussed in Section 1 as well as successful cases of application to different fields. As a statistical method, off-policy evaluation does not pose a significant risk to have a negative societal impact when the technology is used correctly, as the methodology itself is not inherently applied to sensitive data (e.g., in the medical domain). In our particular case, and as stated in sections 2, 3, and 5, if the assumption about the homogeneous behaviour of context within a cluster does not hold or partially holds, the accuracy of the method is not guaranteed, potentially leading to incorrect decisions.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA] .

Justification: The developed method does not present any kind of direct risk for misuse. All datasets are in public domain and no sensitive material of any kind has been generated.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes] .

    Justification: All public datasets, code and methods used in the paper have been properly credited and their associated licenses respected.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes] .

    Justification: The experimentation process and model proposed are described in sections 3 and 4. The released code has been properly documented along with the generation process for reproducing the experimentation. A license with the term of use is also added in the supplemental material containing the code.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer:[NA] .

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.