Barycenter Policy Design for Multiple Policy Evaluation

Simon Weissmann[*]
Universität Mannheim
simon.weissmann@uni-mannheim.de

Till Freihaut[*]
University of Zurich
freihaut@ifi.uzh.ch

Claire Vernade
University of Tübingen
claire.vernade@uni-tuebingen.de

Giorgia Ramponi
University of Zurich
ramponi@ifi.uzh.ch

Leif Döring
Universität Mannheim
leif.doering@uni-mannheim.de

Abstract

A central challenge in reinforcement learning is designing data collection strategies that efficiently evaluate multiple target policies via importance sampling. When target policies are similar, exploiting those similarities in the behavior policy can substantially improve sample efficiency. This article introduces a behavior policy design, examining how different criteria for selecting a behavior policy influence importance sampling estimator properties. We evaluate the resulting behavior policies in downstream tasks, particularly in best policy selection problems. Additionally, we demonstrate how effectively leveraging similarities among target policies results in a more nuanced behavior policy design and enhances regret bounds for best policy selection. To facilitate rigorous analysis, the article is formulated within the stochastic bandit framework.

1 Introduction

Importance sampling (IS), see for instance [Owen, 2013], is a fundamental tool in Monte Carlo simulations, primarily used to estimate more efficiently expectations under distributions other than the sampling one. The IS literature traditionally addresses two key questions: (1) How to design an optimal importance sampling distribution for a given target distribution and (2) how to determine the required sample size for reliable estimation. In machine learning, IS is widely applied to evaluate new objectives using existing data. However, re-weighting data without ensuring proper alignment between the sampling and target distributions can result in high-variance estimators. As a result, research has focused on variance reduction techniques [e.g. Bottou et al., 2013, Kuzborskij et al., 2021, Sakhi et al., 2024] and coverage assumptions, often at the cost of introducing estimation bias or inefficient sampling distributions. An important application of IS in machine learning is off-policy evaluation (OPE) in (contextual) bandits [Wang et al., 2017, Agarwal et al., 2017, Gabbianelli et al., 2024], where multiple target policies $\pi_1, \dots, \pi_N$ are evaluated using data collected under a behavior policy. Recently, there has been renewed interest in IS's original principles, that is, directly addressing the problem of *constructing a behavior policy* to reduce the variance of the estimator [Hanna et al., 2017, Papini et al., 2024, Jain et al., 2024, Liu et al., 2025, Chen et al., 2024, Russo and Pacchiano, 2025]. While prior work focuses on designing behavior policies for single target distributions, more recent attention has been placed to simultaneously estimating multiple expectations [Demange-Chryst

18th European Workshop on Reinforcement Learning (EWRL 2025).

et al., 2023, Chen et al., 2024, Dann et al., 2023, Liu et al., 2025]. In Appendix A we provide a detailed review of related work. This setting, where data collection should efficiently identify the most valuable policy, remains underexplored and lacks theoretical guarantees. The present article addresses the following questions:

Q1: How should a behavior policy π_b be designed to effectively evaluate a given set of target policies? Can it be efficient to use several behavior policies $\pi_b^1, ..., \pi_b^M$?

To obtain a mathematically rigorous analysis, we study this question in the setting of stochastic bandits, defined by the tuple $(\mathcal{A}, P_a)$, where $\mathcal{A}$ is a finite action space and $P_a : \mathcal{A} \to \mathbb{R}$ is a probability kernel mapping actions to rewards. We leave the extension to the RL setting as future work. We assume that all reward distributions are R_*-subgaussian, and denote the expected reward under action $a \in \mathcal{A}$ as $Q(a) = \mathbb{E}[R(a)] := \int_{\mathbb{R}} x \, \mathrm{d}P_a(x)$. A policy $\pi = (\pi(a))_{a\in\mathcal{A}}$ is a probability distribution over actions and its *value* is defined as $v(\pi) = \sum_a \pi(a)Q(a)$. This article is driven by the best-policy selection problem, the identification from data of the best policy from a set Π_N using a carefully chosen behavior policy. The IS estimator of a policy π with respect to the sampling policy π_b is defined as

$$\widehat{v}_n(\pi) := \frac{1}{n}\sum_{t=1}^{n} \frac{\pi(A_t)}{\pi_b(A_t)} R_t, \tag{1}$$

where the pairs (A_t, R_t) are iid action-reward pairs obtained from the bandit model when playing the policy π_b. As mentioned previously, the choice of the behavior policy crucially impacts the IS estimator's performance. In this work, we consider different selection criteria for choosing a suitable behavior policy and discuss the different properties of these. Based on our findings we suggest a method that is both a theoretically feasible and an easily implementable way to find behavior policies to evaluate a set of target policies. Since this method provides a practical way to design behavior policies by using barycenters of a set of target policies, we call the strategy *barycenter-design-based policy evaluation (BD-PE)*. We analyze its theoretical limitations and show that if the target policies lack sufficient structure, the constructed behavior policy may degrade in performance as the number of policies increases. To address this issue, we propose an extension, *clustered barycenter design based policy evaluation (CBD-PE)*, which clusters target policies based on probabilistic similarity and designs multiple behavior policies accordingly. This theoretically grounded approach enables efficient scaling to large policy sets while maintaining strong performance guarantees.

Outline. In Section 2, we present an outline of the high-level idea behind the design of the behavior policies. Our main results on BD-PE and CBD-PE are presented in Sections 3 and 4 respectively. Our analysis focuses on the consequence of structural assumptions on the sample complexity of policy evaluation and selection. Finally, in Section 5, we present an empirical evaluation of our methods, and in Section 6, we discuss potential directions for future work..

2 Behavior policy design using barycentric projections

This section provides an overview of the proposed method, and the following sections are devoted to its analysis. Given a set of target policies $\Pi_N := \{\pi_1, \ldots, \pi_N\}$, our approach comprises the following steps:

1. We propose a sample-efficient behavior policy design, which yields one or more behavior policies. This strategy leverages barycentric projections: each behavior policy is constructed as the barycenter of a subset of target policies (where subsets can be determined by optional policy clustering).
2. These behavior policies are then employed to collect data and evaluate the target policies using importance sampling.
3. Finally, for best-policy selection, we choose the policy $\hat{\pi}_n$ with the largest estimated value.

The practical novelty of this work is Step 1, i.e. how to design the behavior policy so that the final choice in Step 2 is sample efficient. For Step 3 we provide $(\varepsilon$-$\delta)$-excess risk guarantees, i.e. we show that

$$\mathbb{P}(\mathcal{R}(\widehat{\pi}_n) < \varepsilon) \geq 1 - \delta,$$

where the excess-risk (or regret) is defined as

$$\mathcal{R}(\hat{\pi}_n) := v(\pi_*) - v(\hat{\pi}_n)\,, \qquad \pi_* \in \arg\max_{i=1,\dots,N} v(\pi_i). \tag{2}$$

At a high level, our behavior policy design proceeds in two stages. First, the set of target policies can be partitioned into clusters based on policy similarity. Second, barycentric projections are applied within each cluster to generate one behavior policy per cluster. This results in a number of behavior policies equal to the chosen number of clusters M. For initial clarity, we first discuss the barycenter approach assuming no clustering ($M = 1$).

Our approach utilizes barycenters of target policies, defined within the space of policies. To quantify relationships between policy distributions, we employ f-divergences:

Definition 2.1. *Let π, π_b be two discrete probability distributions and $f : (0, \infty) \to \mathbb{R}$ a convex function with $f(1) = 0$. Additionally, we use the standard convention $0f(\frac{0}{0}) = 0$. Then, an f-divergence is defined by* $\mathrm{D}_f(\pi\|\pi_b) := \sum_a \pi_b(a) f\left(\frac{\pi(a)}{\pi_b(a)}\right)$.

There are three popular examples in the RL literature that are particularly useful as they allow simple closed form computations.

- A popular divergence from the IS literature is the KL divergence D_{KL} with $f_{\mathrm{KL}}(x) = x\log(x)$, see [Chatterjee and Diaconis, 2018, Agapiou et al., 2017, Sanz-Alonso, 2018, Beh et al., 2023].
- The χ^2 divergence D_{χ^2} with $f_{\chi^2}(x) = (x-1)^2$, see [Gabbianelli et al., 2024]. This is equivalent to the minimization of variances between the behavior policy and the set of target policies, see [Jain et al., 2024, Chen et al., 2024, Liu et al., 2025].
- The Hellinger distance D_{Hel} with $f_{\mathrm{Hel}}(x) = \frac{1}{2}(\sqrt{x}-1)^2$, see [Foster et al., 2024, Rohatgi et al., 2025]. In contrast to the first two, the square-root of the Hellinger distance is even a metric.

For a fixed divergence we will study the behavior policies $\pi_{\mathrm{KL}}, \pi_{\mathrm{Hel}}$ and π_{χ^2} as the solutions of the problems

$$\pi_b \in \arg\min_{\pi\in\Delta_{\mathcal{A}}} \frac{1}{N}\sum_{i=1}^{N} D_\iota(\pi_i, \pi), \quad \iota \in \{\mathrm{KL}, \mathrm{Hel}, \chi^2\}. \tag{3}$$

The new policy π_b is called the barycenter of $\pi_1, ..., \pi_N$ with respect to the chosen divergence. It is important to note that the three divergences directly give closed-form solutions while other divergences (such as total variation) may not. This means that the proposed method does not suffer from any additional computational cost. Furthermore, we note that the mean distance in (3) can be replaced by the maximum distance. However, this modification generally does not yield a closed-form solution. Instead, it requires solving another convex program that, with additional computational cost, scales in the number of target policies N. Although this is possible in general, we will restrict the analysis to the setting where a closed-form solution exists.

In order to estimate the value of the target policies we use plain vanilla importance sampling with samples drawn from the designed behavior policy. A crucial factor for low estimation error in IS is a small maximal importance weight:

$$\sigma_{\mathrm{IS}} := \max_{\pi\in\Pi_N}\max_{a\in\mathcal{A}} \frac{\pi(a)}{\pi_b(a)}.$$

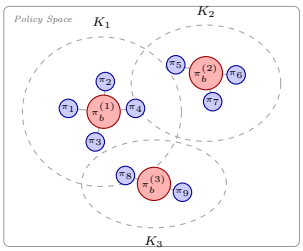
In fact, the excess-risk bounds scale in the sample complexity quadratically in σ_{IS}, see Proposition 3.2 below. We show that barycentric projections using mean divergences provide uniform control over the maximal importance weights for our design of behavior policies, see Table 1. For a more extensive discussion we refer the reader to Appendix B.
The importance weight bounds suggest that the χ^2-distance is the way to go, in particular for $N \gg K$. However, as these are only upper bounds, we will also consider the KL in the following sections. In Section 3, we show that the bounds are tight for Mean-KL and Mean-χ^2, in the sense that there are simple bandit models reaching such importance weights.

The performance of IS can degrade significantly when the behavior and target policies are substantially misaligned, leading to potentially large importance weights.

It is thus natural to combine similar policies into clusters that are treated separately to reduce the sample complexity caused by large importance weights, in particular for large-scale problems. We propose a refined behavior design algorithm, that uses clustering and scales effectively to many target policies, see Figure 1 for an illustration. The refined method works as follows:

1. Cluster (if needed) policies into M clusters.
2. Compute M barycentric policies.
3. Use cluster wise importance sampling.

Figure 1: An illustration of clustered barycenters



A general outline is provided in Algorithm 1, with full details in Sections 4 and 5.

Algorithm 1 Best policy selection using clustered barycenter design

Require: Target policies Π_N, number of clusters $M \geq 1$ ($M = 1$: no clustering), sample sizes n_j, divergence for barycenter projections.

Ensure: Selected best policy $\widehat{\pi}_n$

1: Cluster target policies Π_N into $M \geq 1$ clusters $K_1, \ldots, K_M$ (e.g. with Algorithm 2)
2: For each cluster K_j, compute barycenter behavior policies $\pi_b^{(j)}$ according to (3).
3: Collect datasets $D^{(j)} = (R_t^{(j)}, A_t^{(j)})_{t=1}^{n_j}$ of size n_j using all $\pi_b^{(j)}$
4: For each cluster K_j, compute $\widehat{v}_n^{(j)}(\pi_i)$ according to (5): $\widehat{v}_n^{(j)}(\pi_i) = \frac{1}{n_j} \sum_{t=1}^{n_j} \frac{\pi_i(A_t^{(j)})}{\pi_b^{(j)}(A_t^{(j)})} R_t^{(j)}$ for all $\pi_i \in K_j$;
5: Select best policy: $\hat{\pi}_n \leftarrow \arg\max_i \hat{v}_n^{(j)}(\pi_i)$.

This algorithm provides a theoretically grounded and implementable method for designing behavior policies to effectively evaluate a set of target policies and select the best one. Note that if target policies are already well-aligned, clustering may be unnecessary. Guidance for choosing M is provided in Section 4.1, and Figure 2 presents an experimental study corroborating our theoretical findings.

3 Policy evaluation with barycenter behavior policy design

In this section, we provide excess risk bounds and additionally show that, without clustering, certain policy sets can be hard to evaluate.

3.1 Regret upper bound for barycenter behavior policy design

As mentioned above, we will focus on the behavior policies $\pi_{\mathrm{KL}}, \pi_{\mathrm{Hel}}$ and π_{χ^2} as the solutions of (3). The corresponding IS estimator for arbitrary policies π is given by:

$$\widehat{v}_n(\pi) := \frac{1}{n} \sum_{t=1}^{n} w_b^{\pi}(A_t) R_t \,, \quad w_b^{\pi}(a) := \frac{\pi(a)}{\pi_b(a)}$$

A crucial simple fact about importance sampling with respect to any of these behavior policies is the boundedness of importance weights. It is important to note that the subsequent theoretical analysis

Table 1: Simple closed form solutions for barycentric projections and upper bounds on maximal importance weights. N are the number of target policies, K the number of arms.

Divergence	Closed Form Solution	Bound on σ_{IS}
Mean-KL	$\pi_b(a) = \frac{1}{N} \sum_{i=1}^{N} \pi_i(a)$	N
Mean-χ^2	$\pi_b(a) \propto \sqrt{\sum_{i=1}^{N} \pi_i(a)^2}$	$\min\{N, \sqrt{NK}\}$
Mean-Hellinger	$\pi_b \propto \left(\frac{1}{N} \sum_{i=1}^{N} \sqrt{\pi_i(a)} \right)^2$	N^2

relies solely on this boundedness and is applicable to any behavior policy with similarly bounded importance weights.

Lemma 3.1. *If $w_b^\pi(\cdot) := \frac{\pi(\cdot)}{\pi_b(\cdot)}$ are the importance sampling weights of $\pi \in \Pi_N$ with respect to $\pi_b \in \{\pi_{\mathrm{KL}}, \pi_{\mathrm{Hel}}, \pi_{\chi^2}\}$, then the maximal importance weight*

$$\sigma_{IS} := \max_{\pi \in \Pi_N} \max_{a \in \mathcal{A}} w_b(a) \tag{4}$$

is bounded by N for π_{KL}, $\min\{N, \sqrt{NK}\}$ for π_{χ^2} and N^2 for π_{Hel}.

The proof can be found in Appendix B. It is important to note that for $K > N$ all of these bounds are not dependent on the number of arms. This indicates that choosing $\pi_{\mathrm{KL}}, \pi_{\chi^2}$ or π_{Hel} can be extended to the continuous armed setting. We discuss this extension in Appendix C.2. Furthermore, the boundedness of importance weights simplifies the analysis and corresponds to a uniform coverage assumption, a common requirement in the OPE literature, see, for instance, Wang et al. [2024]. In our considered behavior policy design framework, the crucial quantity is σ_{IS}, a constant that can be interpreted as a measure of how well the behavior policy aligns with the set of target policies. As such, controlling σ_{IS} is of central importance. We will come back to this topic in Section 4 where we suggest clustering methods to strongly decrease the importance weights.

Since the simple form of the barycenter behavior policy design implies bounded weights, the regret (as defined in Equation 2) of the IS estimator is easy to estimate. Assuming R_*-subgaussian rewards, Hoeffding's inequality immediately implies

$$\mathbb{P}(\widehat{v}_n(\pi) - v(\pi) \geq \varepsilon) \leq \exp\left(-\frac{2\varepsilon^2 n}{(R_* \sigma_{\mathrm{IS}})^2}\right).$$

Hence, given $\delta \in (0,1)$, with probability at least $1-\delta$ we have

$$v(\pi) \geq \widehat{v}_n(\pi) - \frac{1}{\sqrt{2n}} R_* \sigma_{\mathrm{IS}} \sqrt{\log\left(\frac{N}{\delta}\right)}, \; \widehat{v}_n(\pi) \geq v(\pi) - \frac{1}{\sqrt{2n}} R_* \sigma_{\mathrm{IS}} \sqrt{\log\left(\frac{N}{\delta}\right)}$$

for all $\pi \in \Pi_N$. These computations outline the key steps in proving the following proposition. The detailed proof is given in Appendix C.3.

Proposition 3.2. *For arbitrary $\delta \in (0,1)$ and $\varepsilon > 0$, let*

$$n \geq n(\varepsilon, \delta) = \frac{2R_*^2 \sigma_{\mathrm{IS}}^2 \log\left(\frac{N}{\delta}\right)}{\varepsilon^2}$$

many iid pairs $(A_t, R_t)_{t=1}^n$ of action and rewards be generated by a behavior policy with importance weights bounded by σ_{IS}, and $\widehat{\pi}_n := \arg\max_i \hat{v}_n(\pi_i)$. Then

$$\mathbb{P}(\mathcal{R}(\widehat{\pi}_n) < \varepsilon) \geq 1 - \delta.$$

From the above result, we observe that the sample size $n(\varepsilon, \delta)$ required for achieving a small regret with high probability scales with σ_{IS}, reaching its worst case when $\sigma_{\mathrm{IS}} \in \{N, \min\{\sqrt{KN}, N\}, N^2\}$, depending on the chosen behavior policy. This means that when the behavior policy is poorly aligned with the target policies, the sample complexity increases linearly or even quadratically in the number of target policies, making evaluation inefficient. To understand this inefficiency, consider the naive approach of evaluating each target policy individually. In this case, $\sigma_{\mathrm{IS}} = 1$, but we would need to sample separately for each of the policies, leading to a total sample complexity of $n \geq \frac{2NR_*^2 \log(\frac{N}{\delta})}{\varepsilon^2}$. Therefore, BD-PE is more sample efficient only if $\sigma_{\mathrm{IS}} \leq \sqrt{N}$. The following lower bound further underscores that this condition does not always hold, emphasizing the need for a more nuanced approach for the barycenter behavior policy design.

3.2 Lower bound

In this section, we provide lower bounds on the importance weights and the resulting excess risk for the behavior policies π_{KL} and π_{χ^2}. These bounds illustrate that the proposed behavior policies have limitations when the set of target policies have unpleasant structure and the amount of target policies increases. The proofs of both results can be found in Appendix C.4. We begin with the lower bound for the importance weights of π_{KL}.

Proposition 3.3. *For arbitrary $N \geq 3$ there exists a multi-armed bandit model $(\mathcal{A}^{(N)}, P_{R^{(N)}})$ and a set of target policies $\Pi_N^{(N)} = \{\pi_1^{(N)}, \ldots, \pi_N^{(N)}\}$ such that $\sigma_{\mathrm{IS}}^{(N)} \geq \sigma_N = \frac{N}{2}$ and*

$$\mathbb{P}(\mathcal{R}(\widehat{\pi}_n) > \varepsilon) \geq \frac{1}{\sqrt{2n}} \exp\left(-\frac{n}{2\sigma_N^2}\right)$$

for all $\varepsilon > 0$ sufficiently small.

Remark 3.4. *In Proposition 3.3, we stated that ε needs to be sufficiently small. Our proof is based on a specific construction of a two-armed stochastic bandit with deterministic rewards $r_1 > r_2$, and a set of target policies. By "sufficiently small" we mean that $\varepsilon \leq \Delta$ where Δ represents the gap between the best policy and the remaining policies, which all share the same value. In our construction, this gap can be explicitly expressed as*

$$\Delta = \left(\frac{1}{N} - \frac{1}{N(N-1)}\right)\left(-\frac{2}{N} + 1\right) r_1 > 0, \quad \forall N \geq 3.$$

Consequently, the gap scales with $r_1 > 0$ but remains independent of its specific value.

Next, we give the lower bound for the case that π_{χ^2} is chosen as the behavior policy.

Proposition 3.5. *For arbitrary $N \geq 3$ there exists a multi-armed bandit model $(\mathcal{A}^{(N)}, P_{R^{(N)}})$ and a set of target policies $\Pi_N^{(N)} = \{\pi_1^{(N)}, \ldots, \pi_N^{(N)}\}$ such that $\sigma_{\chi^2}^{(N)} \geq \sigma_N = \frac{N}{2}$ and*

$$\mathbb{P}(\mathcal{R}(\widehat{\pi}_n) > \varepsilon) \geq \frac{1}{\sqrt{2n}} \exp\left(-\frac{n}{2\sigma_N}\right)$$

for all $\varepsilon > 0$ sufficiently small.

With the last two results we have established lower bounds showing that for certain target policies in a bandit model, σ_{IS} can almost reach its maximum value of N.

This implies that the probability of incurring large is directly influenced by the largest importance weight, which in turn depends on the chosen behavior policy. The observed findings indicate that ensuring a reasonable size for σ_{IS} requires more careful selection of the behavior policy π_b. Otherwise, the sample complexity can grow unfavorably, highlighting the necessity of a more nuanced behavior policy design.

3.3 On the relationship between the barycenter and the target policies

Despite the barycenter behavior policy design's promise of bounded importance weights, our previous analysis revealed that the bound may be quite large. Potentially, this may lead to poor performance in policy evaluation unless certain structural assumptions are made about the set of target policies. In the following section, we show that the policy evaluation can be significantly improved by assuming proximity between each target policy $\pi \in \Pi_N$ and the barycenter π_b. Specifically, if each $\pi_i \in \Pi_N$ satisfies $D_\iota(\pi_i \mid \pi_b) \leq \eta$, where $\iota \in \{\mathrm{KL}, \chi^2\}$ for some $\eta > 0$, then a tighter upper bound on the weight function can be established, leading to improved sample complexity in our regret analysis. The following statement gives an upper bound on σ_{IS} in terms of η. The detailed proof is given in Appendix C.5.

Proposition 3.6. *Suppose that $D_\iota(\pi_i \mid \pi_b) \leq \eta$ for $\iota \in \{\mathrm{KL}, \chi^2\}$ and assume $\pi_b(a) > 0$ for all $a \in \mathcal{A}$, then it holds that*

$$\sigma_{\mathrm{IS}} \leq \min\left(N, 1 + \frac{2\eta}{\min_a \pi_b(a)} + \frac{2\sqrt{2\eta}}{\sqrt{\min_a \pi_{\mathrm{b}}(a)}}\right).$$

As a result, we can explicitly characterize the sample complexity in terms of η. For $\delta \in (0,1)$ and $\varepsilon > 0$, a sample size

$$n \geq n(\varepsilon, \delta) = 2R_*^2 \log(N/\delta) \min\left(N, 1 + \frac{2\eta}{\min_a \pi_b(a)} + \frac{2\sqrt{2\eta}}{\sqrt{\min_a \pi_{\mathrm{b}}(a)}}\right)^2 \varepsilon^{-2}$$

ensures that $\mathbb{P}(\mathcal{R}(\widehat{\pi}_n) < \varepsilon) \geq 1 - \delta$. A drawback of the derived sample complexity is the fact that it can become large when $\min_a \pi_b(a)$ is very small. This issue can be effectively alleviated by introducing a *safe behavior policy*, directly analogous to *safe or defensive importance sampling* from classical IS literature [Owen and Zhou, 2000]. This policy incorporates regularization in the form of

$$\pi_{\text{safe}}^{\lambda}(a) := (1-\lambda)\pi_{\text{KL}}(a) + \lambda u(a), \quad a \in \mathcal{A},\ \lambda \in (0,1)$$

where u is a uniform policy over $\mathcal{A}$. The "safe" aspect stems from this regularization, which guarantees a minimum probability of selecting any action, thus preventing the importance weights from becoming unbounded and causing estimation failure when $\min_a \pi_b(a)$ is small. Choosing $\lambda = \lambda(\eta) = \eta$ ensures that the upper bound on the importance weights remain independent of $\min_a \pi_b(a)$ as η tends to zero. We provide a detailed analysis of the safe behavior policy in Appendix C.6.

4 Best policy selection using clustered behavior policy design

In the previous section, we showed that selecting the best policy based on barycenter behavior policy design requires certain similarity assumptions on the set of target policies. Specifically, we showed that the maximal importance weights can be upper bounded in terms of the f-divergence between each target policy and the barycenter of the set, when the f-divergence correspond to the KL or χ^2. Considering all possible behavior policies, ensuring a small f-divergence may be challenging in general, a natural approach is to partition the set of target policies into clusters with small f-divergences and then apply policy evaluation based on the barycenter behavior policy design to each cluster.

In the following, we introduce an improved method that employs multiple behavior policies, each constructed as the barycenter corresponding to a fixed divergence in (3) to a subset of Π_N. Our regret analysis can be extended to this setting, demonstrating that the required sample sizes for achieving a low regret scale with the number of clusters and the maximal weights within each cluster. The specific design of the clusters will be discussed in Section 5. Importantly, the proposed clustering strategy does not require any additional interactions with the bandit environment. Moreover, based on our upper bounds, we derive a decision rule (7) that evaluates the effectiveness of clustering and guides the selection of a suitable cluster size.

4.1 Clustered based policy evaluation

We begin the discussion by proposing an improved structured importance sampling approach using clustered sets of target policies. Decompose the set of target policies $\Pi_N = \{\pi_1, \dots, \pi_N\}$ into M disjoint clusters $K_j := \{\pi_i^{(j)},\ i = 1, \dots, N_j\}$ of sizes N_j [1] and define

$$\sigma_{\text{c}}^{(j)} := \max_{\ell=1,\dots,N_j} \max_{a \in \mathcal{A}} \frac{\pi_\ell^{(j)}(a)}{\frac{1}{N_j}\sum_{i=1}^{N_j} \pi_i^{(j)}(a)}$$

for all $j = 1, \dots, M$. In our subsequent analysis, we will consider the uniform maximal weight over all clusters defined as

$$\sigma_{\text{c}} := \max_{j=1,\dots,M} \sigma_{\text{c}}^{(j)}$$

For instance, following Lemma 3.1 the value σ_{c} can always be bounded depending on the chosen divergence. We will describe a specific clustering algorithm in Section 5 with the overall goal to achieve small values of $\sigma_{\text{c}}^{(j)}$. To evaluate the policies within each cluster we construct the clustered importance sampling estimators

$$\widehat{v}_n^{(j)}(\pi) := \frac{1}{n_j} \sum_{t=1}^{n_j} \frac{\pi(A_t^{(j)})}{\pi_b^{(j)}(A_t^{(j)})} R_t^{(j)}\,, \quad \pi \in K_j\,, \tag{5}$$

for $\pi_b^{(j)} \in \{\pi_{\text{KL}}^{(j)}, \pi_{\text{Hel}}^{(j)}, \pi_{\chi^2}^{(j)}\}$ of the corresponding clusters K_j, $j \in \{1, \dots, M\}$. Here, $(A_t^{(j)}, R_t^{(j)})_{t=1}^{n_j}$ are iid pairs of action and rewards generated by the behavior policy $\pi_b^{(j)}$ and the overall sample size is $n = \sum_{j=1}^{M} n_j$. The clustered best policy selection is then defined as

$$\widehat{\pi}_n := \operatorname*{arg\,max}_{j \in \{1,\dots,M\}} \{\max_{\pi \in K_j} \widehat{v}_n^{(j)}(\pi)\}\,. \tag{6}$$

[1] We can easily use cluster algorithms as K-means, see Section 5.

4.2 Clustered regret analysis

In this section, we derive regret bounds of the proposed clustered best-policy selection approach. For simplicity, we make the following (mainly notational) assumption on the value of the excess risk per cluster.

Assumption 4.1. *Let* $\pi_* = \arg\max_{\pi \in \Pi_N} v(\pi)$ *and* $\pi_*^{(j)} := \arg\max_{\pi \in K_j} v(\pi)$ *for* $j = 1, \dots, M$, *then (i)* $v(\pi_*) = v(\pi_*^{(1)})$ *and (ii)* $v(\pi_*^{(1)}) - v(\pi_*^{(j)}) > \Delta$ *for some* $\Delta > 0$ *and all* $j = 2, \dots, M$.

To quantify the upper regret bound, we need to control the probability that the best policy is selected within cluster K_1 that contains the best policy. The full proof is given in Appendix D.1

Proposition 4.2. *Suppose that Assumption 4.1 is in place, and* $n_1 = \cdots = n_M$. *Furthermore, let* $\delta \in (0, 1)$ *and* $\varepsilon \in (0, \Delta]$. *If*

$$n = n_1 \cdot M \geq n(\varepsilon, \delta) := \frac{2M R_*^2 \sigma_c^2 \log\left(\frac{(M-2)(N_1+1)+N+M}{\delta}\right)}{\varepsilon^2},$$

then $\mathbb{P}(\widehat{\pi}_n \notin K_1) \leq \delta$.

Having established the high probability guarantee for $\widehat{\pi}_n \in K_1$, we can derive the overall regret bound of the clustered best-policy selection.

Theorem 4.3. *Suppose that Assumption 4.1 is in place, and* $n_1 = \cdots = n_M$. *Furthermore, let* $\delta \in (0, 1)$, $\varepsilon \in (0, \Delta]$. *If*

$$n = n_1 \cdot M \geq n(\varepsilon, \delta) := \frac{2M R_*^2 \sigma_c^2 \log(\frac{2+N+M+(M-1)(N_1+1)}{\delta})}{\varepsilon^2}$$

then

$$\mathbb{P}(\mathcal{R}(\widehat{\pi}_n) < \varepsilon) \geq 1 - \delta.$$

We provide the full proof of Theorem 4.3 in Appendix D.2. Note that Theorem 4.3 depends on the problem-dependent minimal gap Δ assuming that $\epsilon \in (0, \Delta]$. We present a problem-independent version for the expected excess risk in the appendix, see Corollary D.2.

Remark 4.4. *For simplicity, we assumed that the number of samples is distributed equally across clusters. However, sample allocation could be optimized by considering the respective* σ_{c_j}, *as the effectiveness of IS estimators depends on* σ_{c_j}. *This means that more samples should be allocated to clusters of target policies that have a high* σ_{c_j}. *Note that* σ_{c_j} *can be calculated explicitly for each cluster without suffering from a high additional cost.*

Let us conclude this section with a comparison of sample complexity with and without clustering. First, note that for $M = 1$, i.e. no clustering is applied, the sample complexity (ignoring the logarithmic terms) simplifies to $\frac{2R_*^2\sigma_{\mathrm{IS}}^2}{\varepsilon^2}$. This means that CBD-PE yields a sample complexity improvement if

$$M\sigma_c^2 < \sigma_{\mathrm{IS}}^2. \tag{7}$$

Thus, the effectiveness of CBD-PE is directly tied to the reduction in σ_c relative to σ_{IS}. We emphasize that we have access to the set of target policies Π_N. Therefore, we can compute σ_c and σ_{IS} without interacting with the bandit environment. This implies that one can evaluate the effectiveness of the clustering without incurring additional sampling costs. In the next section, we will numerically quantify this effect.

5 Empirical evaluation

In order to implement Algorithm 1 we construct a specific clustering approach. Motivated by Proposition 3.6, we seek to achieve small values of $\sigma_c^{(j)}$ by minimizing the respective divergence of all policies within the cluster. Therefore, one can adapt standard cluster algorithms to create clusters with small f-divergences and then apply CBD-PE as discussed in the subsequent sub-sections.

Clustering with respect to f divergences Clustering algorithms are widely used to partition items into groups based on similarity. Typically, these methods operate within a metric space, where distances between elements satisfy standard metric properties. However, in probability spaces, distributions can also be grouped based on divergence measures rather than traditional metrics. Notably, the squared Hellinger distance has the additional property that its square root defines a proper metric, enabling the use of standard clustering techniques such as *KMeans*. Since all KL, χ^2 and Hellinger distances belong to the family of *f-divergences*, we can leverage clustering results obtained via the Hellinger distance to infer meaningful groupings. A detailed theoretical justification for this approach is provided in Chaudhuri and McGregor [2008]. The description of the Hellinger-based clustering approach can be found in Algorithm 2.

Numerical validations To clearly visualize the effects, we empirically evaluate the proposed clustering approach on an extreme toy example. Our experiment aims to demonstrate that the proposed regret bounds improve when the target policy set exhibits a structured form. To verify this, we test whether (7) holds. For this experiment, we generate sets of softmax target policies of size $N = 1000$ in a 100-armed bandit setting with gaussian-distributed rewards by sampling weights for each arm and rescaling to softmax policies with a temperature parameter of 1. The maximal mean reward is set to 3 for arm 1, linearly decaying by 0.05 until arm 100. A detailed discussion on the construction can be found in Appendix E. The optimal target policy achieves a value of 2.77. We compare BD-PE and CBD-PE for different numbers of clusters M applied in Algorithm 2 and for the barycenters with respect to the KL and the χ^2 divergence. Importantly, the total number of samples used remains the same in each approach. When clustering is applied, samples are uniformly distributed across all clusters. First, we record the resulting values of $M\sigma_c^2$ for different numbers of clusters M. Recall, for $M = 1$ we recover the case $\sigma_c = \sigma_{\mathrm{IS}}$. Afterwards, we observe that increasing the cluster size lowers the effect on σ_c. This underscores the discussion at the end of Section 4.1. For practitioners it is important to find a suitable cluster size. The results are shown in Figure 2 (a). They show clustering significantly improves the value of $M\sigma_c^2$ until a cluster size of $M = 10$. This suggests that an improved regret is expected by adjusting the number of clusters. Therefore, we evaluate the average regret of the proposed approach across varying sample sizes for BD-PE and CBD-PE with different numbers of clusters. Additionally, we include a Monte Carlo approach as a baseline, corresponding to the case where $M = 1000$. Note that, a minimum sample size of $n = 1000$ is required. The results are shown in Figure 2 (b), where we observe the best performance for $M = 10$. Conversely, too large or too small number of clusters result in higher regrets on average.

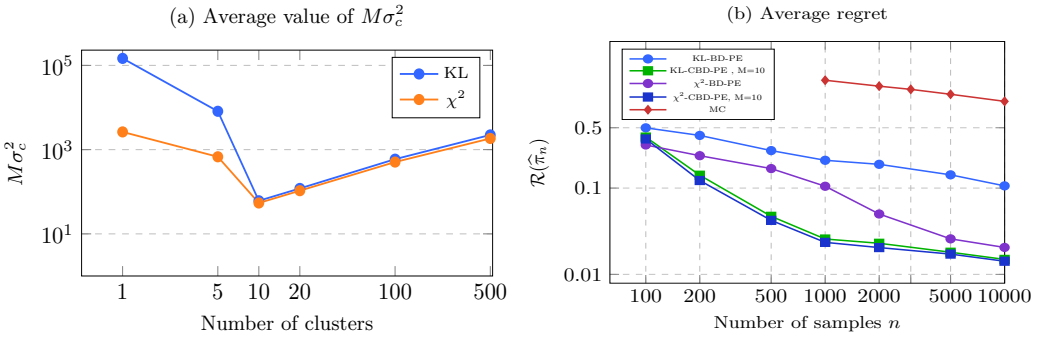


Figure 2: Comparison of (a) the values of $M\sigma_c^2$ and (b) the regret for different numbers of clusters M averaged over 1000 independent runs.

6 Conclusion

In this work, we addressed the question on how to effectively design one or more behavior policies to perform best-policy selection based on importance sampling estimators. Using barycenter behavior policies (with and without clustering of target policies) we provide theoretical and numerical validations on the performance. The results focus on the identification of the best policy rather than only obtaining accurate value estimates. By introducing clustering we also provide a simple practical solution to scale up the approach to many target policies, that does not require any additional

interaction with the underlying bandit environment. We observe that incorporating clustering significantly reduces regret, highlighting the advantage of cluster-based behavior policies and validating our theoretical findings, providing a positive answer to the questions posed in Question Q1.

Future research could explore extensions of these theoretical results to contextual bandits and MDPs, such as applications of IS in reinforcement learning [Jain et al., 2024, Papini et al., 2024]. Extensions would require a generalized clustering methodology that can effectively account for variations across different contexts and states. Another promising direction for future work is the sequential integration of policy evaluation and selection. Specifically, one could adaptively eliminate underperforming clusters throughout the clustered policy evaluation process, enhancing efficiency of decision-making.

References

S. Agapiou, O. Papaspiliopoulos, D. Sanz-Alonso, and A. M. Stuart. Importance sampling: Intrinsic dimension and computational cost. *Statistical Science*, 32(3):405–431, 2017. ISSN 08834237, 21688745. URL http://www.jstor.org/stable/26408299.

A. Agarwal, S. Basu, T. Schnabel, and T. Joachims. Effective evaluation using logged bandit feedback from multiple loggers. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17. ACM, Aug. 2017. doi: 10.1145/3097983.3098155. URL http://dx.doi.org/10.1145/3097983.3098155.

R. Ash. *Information Theory*. Dover books on advanced mathematics. Dover Publications, 1990. ISBN 9780486665214. URL https://books.google.ch/books?id=nJ3UmGvdUCoC.

J. Beh, Y. Shadmi, and F. Simatos. Insight from the kullback–leibler divergence into adaptive importance sampling schemes for rare event analysis in high dimension. *arXiv-Preprint*, arxiv:2309.16828, 2023. URL https://arxiv.org/abs/2309.16828.

L. Bottou, J. Peters, J. Quiñonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(101):3207–3260, 2013. URL http://jmlr.org/papers/v14/bottou13a.html.

S. Chatterjee and P. Diaconis. The sample size required in importance sampling. *The Annals of Applied Probability*, 28(2):1099 – 1135, 2018. doi: 10.1214/17-AAP1326. URL https://doi.org/10.1214/17-AAP1326.

K. Chaudhuri and A. McGregor. Finding metric structure in information theoretic clustering. In *COLT*, pages 391–402, 2008.

Y. Chen, A. Pacchiano, and I. C. Paschalidis. Multiple-policy evaluation via density estimation, 2024. URL https://arxiv.org/abs/2404.00195.

C. Dann, M. Ghavamzadeh, and T. V. Marinov. Multiple-policy high-confidence policy evaluation. In F. Ruiz, J. Dy, and J.-W. van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 9470–9487. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/dann23a.html.

J. Demange-Chryst, F. Bachoc, and J. Morio. Efficient estimation of multiple expectations with the same sample by adaptive importance sampling and control variates. *Statistics and Computing*, 33(5):103, July 2023. ISSN 1573-1375. doi: 10.1007/s11222-023-10270-y. URL https://doi.org/10.1007/s11222-023-10270-y.

D. J. Foster, A. Block, and D. Misra. Is behavior cloning all you need? understanding horizon in imitation learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=8KPyJm4gt5.

G. Gabbianelli, G. Neu, and M. Papini. Importance-weighted offline learning done right. In C. Vernade and D. Hsu, editors, *Proceedings of The 35th International Conference on Algorithmic Learning Theory*, volume 237 of *Proceedings of Machine Learning Research*, pages 614–634. PMLR, 25–28 Feb 2024. URL https://proceedings.mlr.press/v237/gabbianelli24a.html.

J. P. Hanna, P. S. Thomas, P. Stone, and S. Niekum. Data-efficient policy evaluation through behavior policy search. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1394–1403. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/hanna17a.html.

T. Hesterberg. Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37(2):185–194, 1995. doi: 10.1080/00401706.1995.10484303. URL https://www.tandfonline.com/doi/abs/10.1080/00401706.1995.10484303.

E. L. Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008. ISSN 10618600. URL http://www.jstor.org/stable/27594308.

A. Jain, J. Hanna, and D. Precup. Adaptive exploration for data-efficient general value function evaluations. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 65912–65943. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/794a72b1a9d5fc4c040eb3110d94c8a1-Paper-Conference.pdf.

Y. Jin, Z. Yang, and Z. Wang. Is pessimism provably efficient for offline RL? In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5084–5096. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/jin21e.html.

Y. Jin, Z. Ren, Z. Yang, and Z. Wang. Policy learning "without" overlap: Pessimism and generalized empirical bernstein's inequality, 2022. URL https://arxiv.org/abs/2212.09900.

I. Kuzborskij, C. Vernade, A. Gyorgy, and C. Szepesvari. Confident off-policy evaluation and selection through self-normalized importance weighting. In A. Banerjee and K. Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 640–648. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/kuzborskij21a.html.

L. Li, W. Chu, J. Langford, and X. Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM'11, page 297–306. ACM, Feb. 2011. doi: 10.1145/1935826.1935878. URL http://dx.doi.org/10.1145/1935826.1935878.

S. Liu, C. Chen, and S. Zhang. Efficient multi-policy evaluation for reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39:18951–18959, 04 2025. doi: 10.1609/aaai.v39i18.34086.

B. London and T. Sandler. Bayesian counterfactual risk minimization. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4125–4133. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/london19a.html.

A. Owen and Y. Zhou. Safe and effective importance sampling. *Journal of the American Statistical Association*, 95(449):135–143, 2000. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2669533.

A. B. Owen. *Monte Carlo theory, methods and examples*. https://artowen.su.domains/mc/, 2013.

M. Papini, G. Manganini, A. M. Metelli, and M. Restelli. Policy gradient with active importance sampling, 2024. URL https://arxiv.org/abs/2405.05630.

P. Rashidinejad, B. Zhu, C. Ma, J. Jiao, and S. Russell. Bridging offline reinforcement learning and imitation learning: a tale of pessimism. NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

D. Rohatgi, A. Block, A. Huang, A. Krishnamurthy, and D. J. Foster. Computational-statistical tradeoffs at the next-token prediction barrier: Autoregressive and imitation learning under misspecification, 2025. URL https://arxiv.org/abs/2502.12465.

A. Russo and A. Pacchiano. Adaptive exploration for multi-reward multi-policy evaluation, 2025. URL https://arxiv.org/abs/2502.02516.

O. Sakhi, I. Aouali, P. Alquier, and N. Chopin. Logarithmic smoothing for pessimistic off-policy evaluation, selection and learning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 80706–80755. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/9379ea6ba7a61a402c7750833848b99f-Paper-Conference.pdf.

D. Sanz-Alonso. Importance sampling and necessary sample size: An information theory approach. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2):867–879, 2018. doi: 10.1137/16M1093549. URL https://doi.org/10.1137/16M1093549.

A. Swaminathan and T. Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research*, 16(52):1731–1755, 2015. URL http://jmlr.org/papers/v16/swaminathan15a.html.

P. Thomas, G. Theocharous, and M. Ghavamzadeh. High confidence policy improvement. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2380–2388, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/thomas15.html.

L. Wang, A. Krishnamurthy, and A. Slivkins. Oracle-efficient pessimism: Offline policy optimization in contextual bandits. In S. Dasgupta, S. Mandt, and Y. Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 766–774. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/wang24a.html.

Y.-X. Wang, A. Agarwal, and M. Dudík. Optimal and adaptive off-policy evaluation in contextual bandits. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3589–3597. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/wang17a.html.

A Related work

Multiple policy evaluation and behavior policy optimization. Our work is closely related to the emerging field of multiple policy evaluation and the closely connected behavioral policy optimization. The literature can be divided into two parts, evaluating the target sampling directly from the underlying target policies [Dann et al., 2023] or evaluating them using data from a different policy, the so called behavior policy [Liu et al., 2025, Chen et al., 2024, Jain et al., 2024]. Regarding the first setting, the simplest approach is to use Monte Carlo simulation directly. As this requires a large amount of samples, Dann et al. [2023] have proposed to reuse samples across target policies based on a dissimilarity measure in the context of Reinforcement Learning. In favorable cases, this leads to a sample complexity that is sub-linear in the number of target policies. Similarity is measured on a trajectory level, meaning if the same action is played in the same state across target policies to reuse the samples. This is different from our information-theoretic similarity measure suitable for probability distributions. In the latter case, the goal is to collect suitable data with a behavior policy and use techniques like Importance Sampling to evaluate the set of target policies. In Chen et al. [2024] they first compute a coarse estimation of the visitation distribution and then minimize the variance of visitation distribution of the behavior policy and the estimated ones of the set of target policies. Similarly, Jain et al. [2024] take a sequential approach by designing a behavior policy that minimizes the average Mean-Squared Error(MSE) between the behavior policy and a set of target policies. Liu et al. [2025] also minimize trajectory-level variance to guide behavior policy design using a similarity measure based on variance. More recently, Russo and Pacchiano [2025] have developed (ϵ, δ)-PAC guarantees in an RL setting where different policies have to be evaluated for multiple reward functions. Additionally, in another line of research, Papini et al. [2024] introduce the behavioral policy optimization problem in the context of policy gradient methods, aiming to collect data that minimizes the variance of policy gradient estimates. In contrast, our work gives a general approach and considers different multiple target policies and explores similarities with respect to f divergences.

Off-policy evaluation. Our work is somewhat orthogonal to the well-studied problem of off-policy evaluation (OPE) [Li et al., 2011, Bottou et al., 2013, Swaminathan and Joachims, 2015]. In OPE, data has already been collected using a fixed and often known behavior policy. IS estimators are then used to evaluate the set of target policies. However, as the behavior policy is fixed and can be rather arbitrarily, there can be a mismatch between the target and behavior policies. The focus of these works is to refine the estimator instead of selecting a suitable behavior policy. Several techniques have been proposed to mitigate variance due to policy mismatch. Clipping IS weights is a widely used approach [Ionides, 2008, Thomas et al., 2015, Bottou et al., 2013], as is introducing a pessimistic bias into the estimator [Swaminathan and Joachims, 2015, London and Sandler, 2019, Jin et al., 2021, Rashidinejad et al., 2021, Jin et al., 2022]. Another variance-reduction method is Self-Normalized IS, which stabilizes estimates while maintaining practical effectiveness [Kuzborskij et al., 2021, Hesterberg, 1995]. Many of these approaches assume uniformly bounded importance weights, an assumption recently relaxed by Gabbianelli et al. [2024] through the introduction of an exploration parameter γ, which implicitly constrains importance weights. This perspective aligns with the idea of selecting a "safe" behavior policy, which is formalized in this work.

B Proofs of Section 2

In this section we provide the calculations for the divergences summarized in Table 1. We start with the computation of the mean-KL divergence.

Lemma B.1. *Let $\Pi_N = \{\pi_1, \dots, \pi_N\}$ be a set of target policies. We define the arithmetic mean $\overline{\pi}$ as*

$$\overline{\pi}(a_k) := \frac{1}{N} \sum_{i=1}^{N} \pi_i(a_k) \quad \textit{for every} \quad k \in K.$$

Furthermore, let $\mathrm{D}_{\mathrm{KL}}(\Pi_N|\pi_b) := \frac{1}{N} \sum_{i=1}^{N} \mathrm{D}_{\mathrm{KL}}(\pi_i, \pi_b)$ be the average right KL divergence. Then it holds true that

$$D_{\mathrm{KL}}(\Pi_N|\pi_b) = \left(H(\overline{\pi}) - \frac{1}{N} \sum_{i=1}^{N} H(\pi_i) \right) + D_{\mathrm{KL}}(\overline{\pi}|\pi_b),$$

where $H(\pi) = -\sum_{a\in\mathcal{A}} \pi(a)\log\pi(a)$ is the entropy. Furthermore, choosing $\pi_b = \overline{\pi}$ minimizes the average right KL divergence.

Proof. We can compute the average KL divergence as follows:

$$\begin{aligned}
D_{\mathrm{KL}}(\Pi_N|p) &= \frac{1}{N}\sum_{i=1}^{N} D_{\mathrm{KL}}(\pi_i,\pi_b) \\
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{k\in K}\pi_i(a_k)\log\frac{\pi_i(a_k)}{\pi_b(a_k)} \\
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{k\in K}\pi_i(a_k)\log\pi_i(a_k) - \frac{1}{N}\sum_{i=1}^{N}\sum_{k\in K}\pi_i(a_k)\log\pi_b(a_k) \\
&= -\frac{1}{N}\sum_{i=1}^{N} H(\pi_i) - \sum_{k\in K}\log\pi_b(a_k)\frac{1}{N}\sum_{i=1}^{N}\pi_i(a_k) \\
&= \left(H(\overline{\pi}) - \frac{1}{N}\sum_{i=1}^{N} H(\pi_i)\right) - \sum_{k\in K}\log\pi_b(a_k)\overline{\pi}(a_k) - H(\overline{\pi}) \\
&= \left(H(\overline{\pi}) - \frac{1}{N}\sum_{i=1}^{N} H(\pi_i)\right) - \sum_{k\in K}\log\pi_b(a_k)\overline{\pi}(a_k) + \sum_{k\in K}\overline{\pi}(a_k)\log\overline{\pi}(a_k) \\
&= \left(H(\overline{\pi}) - \frac{1}{N}\sum_{i=1}^{N} H(\pi_i)\right) + D_{\mathrm{KL}}(\overline{\pi}|\pi_b)
\end{aligned}$$

With this form of the average right KL divergence, we now want to choose p, such that this term is minimized. We can see that we can split the formula into two parts

$$D_{\mathrm{KL}}(\Pi_N|\pi_b) = \left(H(\overline{\pi}) - \frac{1}{N}\sum_{i=1}^{N} H(\pi_i)\right) + D_{\mathrm{KL}}(\overline{\pi}|\pi_b).$$

We immediately see that the first part is independent of the choice of π_b. This implies that we only have to minimize the second term. The KL divergence of two distributions is minimized, if both distributions are the same, because then $\log\frac{\overline{\pi}}{\overline{\pi}} = 0$. Overall this says that $D_{\mathrm{KL}}(\Pi_N|\pi_b)$ is minimized for the choice of $\pi_b = \overline{\pi}$. □

Another common measure from the class of f-divergences is the Hellinger distance. The Hellinger distance has the property, that the square root is also a metric.

Lemma B.2 (Mean-Hellinger Distance)**.** *The minimizer of the mean-Hellinger divergence,*

$$\min_{\pi\in\Delta_{\mathcal{A}}} \frac{1}{N}\sum_{i=1}^{N} D_{\mathrm{Hel}}(\pi_i \,\|\, \pi),$$

is given in closed form by

$$\pi^*(a) \;=\; \frac{\left(\frac{1}{N}\sum_{i=1}^{N}\sqrt{\pi_i(a)}\right)^2}{\sum_{b=1}^{K}\left(\frac{1}{N}\sum_{i=1}^{N}\sqrt{\pi_i(b)}\right)^2}\,.$$

Proof. First note that

$$\frac{1}{N}\sum_{i=1}^{N} D_{\mathrm{Hel}}(\pi_i \,\|\, \pi) = \frac{1}{2N}\sum_{i=1}^{N}\sum_{a=1}^{K}\Big(\pi_i(a) + \pi(a) - 2\sqrt{\pi_i(a)\,\pi(a)}\Big).$$

Up to an additive constant $\frac{1}{2N}\sum_{i,a}\pi_i(a)$, we equivalently minimize

$$\sum_{a=1}^{K}\Big[\tfrac{1}{2}\,\pi(a)\;-\;\frac{1}{N}\sum_{i=1}^{N}\sqrt{\pi_i(a)\,\pi(a)}\Big].$$

That is,

$$\min_{\pi\in\Delta}\;\sum_{a}\Big[\tfrac{1}{2}\,\pi(a)-m(a)\sqrt{\pi(a)}\Big],\quad\text{where } m(a)=\frac{1}{N}\sum_{i=1}^{N}\sqrt{\pi_i(a)}.$$

Form the Lagrangian with multiplier λ for the simplex constraint:

$$\mathcal{L}(\pi,\lambda)=\sum_{a}\big[\tfrac{1}{2}\,\pi(a)-m(a)\sqrt{\pi(a)}\big]+\lambda\big(\sum_{a}\pi(a)-1\big).$$

Differentiating with respect to $\pi(a)$ and setting to zero gives

$$\tfrac{1}{2}-\frac{m(a)}{2\sqrt{\pi(a)}}+\lambda=0\;\;\Longrightarrow\;\;\sqrt{\pi(a)}\;=\;\frac{m(a)}{1+2\lambda}.$$

Since $1+2\lambda$ is constant across a, we deduce

$$\sqrt{\pi(a)}\propto m(a)\quad\Longrightarrow\quad\pi(a)\propto m(a)^2\;=\;\Big(\tfrac{1}{N}\sum_{i=1}^{N}\sqrt{\pi_i(a)}\Big)^2.$$

Normalizing over $a=1,\dots,K$ yields the claimed form. □

Note, that the χ^2 divergence is closely related to the variance, therefore the following result is also important for the variance optimization problem.

Lemma B.3 (Mean-χ^2 Divergence)**.** *The solution to*

$$\min_{\pi\in\Delta_{\mathcal{A}}}\frac{1}{N}\sum_{i=1}^{N}D_{\chi^2}(\pi_i\|\pi),\quad\textit{with } D_{\chi^2}(\pi_i\|\pi):=\sum_{a}\frac{(\pi_i(a)-\pi(a))^2}{\pi(a)},$$

is given in closed form by

$$\pi^*(a)=\frac{\sqrt{\sum_{i=1}^{N}\pi_i(a)^2}}{\sum_b\sqrt{\sum_{i=1}^{N}\pi_i(b)^2}}.$$

Proof. We rewrite the objective as

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{a}\frac{\pi_i(a)^2}{\pi(a)}-1.$$

Ignoring constants, we solve

$$\min_{\pi\in\Delta_{\mathcal{A}}}\sum_{a}\frac{\sum_i\pi_i(a)^2}{\pi(a)}.$$

Forming the Lagrangian with multiplier λ:

$$\mathcal{L}(\pi,\lambda)=\sum_{a}\frac{\sum_i\pi_i(a)^2}{\pi(a)}+\lambda\left(\sum_{a}\pi(a)-1\right).$$

Setting derivative with respect to $\pi(a)$ to zero:

$$-\frac{\sum_i\pi_i(a)^2}{\pi(a)^2}+\lambda=0\quad\Longrightarrow\quad\pi(a)^2\propto\sum_i\pi_i(a)^2.$$

Normalizing gives the claimed result. □

Note that optimizing the mean variance has the objective as minimizing the mean χ^2, i.e.

$$\sum_a \frac{\sum_i \pi_i(a)^2}{\pi(a)}.$$

Therefore, the same Lagrangian argument as in Lemma B.3 applies, yielding the same closed-form solution.

Last, we consider the max variance, which is a popular optimization goal in the literature, but as it turns out, does not have a closed-form solution.

In the table we did not list the objectives, if one tries to minimize the maximum distance of for example the KL or the variance. Next, we will state for both of these cases another computation in form of solving an additional program is required.

Lemma B.4 (Max-KL Divergence)**.** *The problem*

$$\min_{\pi \in \Delta_{\mathcal{A}}} \max_{i \in \{1,\dots,N\}} D_{\mathrm{KL}}(\pi_i \| \pi)$$

can be written as the convex program

$$\min_{\pi,\, t} \; t \quad \textit{s.t. } D_{\mathrm{KL}}(\pi_i \| \pi) \le t, \forall i, \quad \sum_a \pi(a) = 1,\ \pi(a) > 0.$$

Proof. Introduce auxiliary variable t and rewrite:

$$\min_{\pi,t} t \quad \text{s.t. } \sum_a \pi_i(a) \log \frac{\pi_i(a)}{\pi(a)} \le t,\ \forall i.$$

Each KL constraint is convex in π, so the overall program is convex. This has no closed-form solution and instead requires to solve an additional convex program. □

Lemma B.5 (Max-Variance of Importance Weights)**.** *The problem*

$$\min_{\pi \in \Delta_{\mathcal{A}}} \max_{i \in \{1,\dots,N\}} \sum_a \frac{\pi_i(a)^2}{\pi(a)}$$

can be formulated as the convex program

$$\min_{\pi,\, t} \; t \quad \textit{s.t.} \sum_a \frac{\pi_i(a)^2}{\pi(a)} \le t, \forall i, \quad \sum_a \pi(a) = 1,\ \pi(a) > 0.$$

Proof. Introduce auxiliary t and rewrite the problem:

$$\min_{\pi,t} t \quad \text{s.t. } \sum_a \frac{\pi_i(a)^2}{\pi(a)} \le t,\ \forall i.$$

Each constraint is convex in $\pi(a)$ since $1/\pi(a)$ is convex. This does not have a closed-form solution and instead requires to solve a convex program. □

C Proofs and additional details of Section 3

In this section, we give the omitted proofs and additional details of Section 3.

C.1 Bounding the Importance weights

Next, we provide the proof for the bounds on σ_{IS} depending on the chosen behavior policy.

Lemma C.1 (Upper Bound on IS Weight under χ^2–Barycenter)**.** *Let $\mathcal{A}$ be an action set with $|\mathcal{A}| = K$, and let $\{\pi_i\}_{i=1}^N \subset \Delta_{\mathcal{A}}$ be N probability distributions over $\mathcal{A}$. Define their χ^2-barycenter by*

$$\pi_{\chi^2}(a) \;=\; \frac{\sqrt{\sum_{i=1}^N \pi_i(a)^2}}{\sum_{b \in \mathcal{A}} \sqrt{\sum_{i=1}^N \pi_i(b)^2}}.$$

Then the maximum importance-sampling weight

$$\sigma_{\mathrm{IS}} \;=\; \max_{i\in[N],\, a\in\mathcal{A}} \frac{\pi_i(a)}{\pi_{\chi^2}(a)}$$

satisfies the bound

$$\sigma_{\mathrm{IS}} \;\le\; \min\{\, N,\; \sqrt{NK}\}.$$

Proof. First, let us consider the case $K \geq N$, then we get

$$\frac{\pi_i(a)}{\pi_{\chi^2}(a)} = \pi_i(a)\,\frac{\sum_b \sqrt{\sum_j \pi_j(b)^2}}{\sqrt{\sum_j \pi_j(a)^2}} \leq \sum_b \sqrt{\sum_j \pi_j(b)^2} \leq \sum_b \sum_j \pi_j(b) = N\,,$$

where we applied $\sqrt{\sum_j \pi_j(b)^2} \leq \sum_j \pi_j(b)$ and used that the policies are probability distributions. Next, let us consider the case $K < N$. By definition,

$$\frac{\pi_i(a)}{\pi_{\chi^2}(a)} = \pi_i(a)\,\frac{\sum_b \sqrt{\sum_j \pi_j(b)^2}}{\sqrt{\sum_j \pi_j(a)^2}}\,.$$

Since $\max_i \pi_i(a) \leq \sqrt{\sum_j \pi_j(a)^2}$, it follows that for all a,

$$\max_i \frac{\pi_i(a)}{\pi_{\chi^2}(a)} \;\le\; \sum_{b\in\mathcal{A}} \sqrt{\sum_{j=1}^N \pi_j(b)^2} =: Z.$$

Now, we apply the Cauchy Schwarz inequality to obtain

$$Z = \sum_{b\in\mathcal{A}} \sqrt{\sum_{j=1}^N \pi_j(b)^2} \leq \sqrt{K \sum_b \sum_{j=1}^N \pi_j(b)^2} = \sqrt{KN.}$$

This completes the proof. □

We can summarize our results and add the Hellinger distance in the following lemma.

Lemma C.2. *If $w_b^\pi(\cdot) := \frac{\pi(\cdot)}{\pi_b(\cdot)}$ are the importance sampling weights of $\pi \in \Pi_N$ with respect to $\pi_b \in \{\pi_{\mathrm{KL}}, \pi_{\mathrm{Hel}}, \pi_{\chi^2}\}$, then the maximal weight*

$$\sigma_{IS} := \max_{\pi\in\Pi_N} \max_{a\in\mathcal{A}} w_b(a) \tag{8}$$

is bounded by N for π_{KL}, $\min\{\sqrt{KN}, N\}$ for π_{χ^2} and N^2 for π_{Hel}.

Proof. The proof for π_{KL} follows directly from the definition. For π_{χ^2} we can apply Lemma C.1. Last, consider π_{Hel}, it follows by the definition

$$\max_i \frac{\pi_i(a)}{\pi_{\mathrm{Hel}}(a)} = \max_i \frac{\pi_i(a)}{\left(\frac{1}{N}\sum_{i=1}^N \sqrt{\pi_i(a)}\right)^2} \sum_{a'} \left(\frac{1}{N}\sum_{i=1}^N \sqrt{\pi_i(a')}\right)^2.$$

For the second term, we now apply Jensen to get

$$\sum_{a'} \left(\frac{1}{N}\sum_{i=1}^N \sqrt{\pi_i(a')}\right)^2 \leq \sum_{a'} \frac{1}{N}\sum_i \pi_i(a') = \frac{1}{N}\sum_i 1 = 1\,.$$

Next, we consider the first part and get

$$\frac{\pi_i(a)}{\left(\frac{1}{N}\sum_{i=1}^N \sqrt{\pi_i(a)}\right)^2} = \left(\frac{\sqrt{\pi_i(a)}}{\frac{1}{N}\sum_{i=1}^N \sqrt{\pi_i(a)}}\right)^2 \leq \left(\frac{\sqrt{\pi_i(a)}}{\frac{1}{N}\sqrt{\pi_i(a)}}\right)^2 = N^2,$$

which yields the desired upper bound. □

C.2 A continuous armed extension

In this section, we elaborate how our approach also works in the case of continuous armed bandits. First, note that our results do not scale with the number of arms, therefore we only need to check the assumptions on the derived theorems. The only theorem, where we explicitly needed discrete distributions is Lemma B.1. Fortunately, we can give a continuous extension of the theorem. Note that for χ^2 no extension is needed as Lemma B.3 already applies for the general case.

Lemma C.3. *Let $\Pi_N = \{\pi_1, \dots, \pi_N\}$ be a set of continuous distributions. Then, the solution to (3) KL as an objective is given by*

$$\pi_{\mathrm{KL}} := \frac{1}{N} \sum_{i=1}^{N} \pi_i.$$

We call this the continuous KL-barycenter.

Proof. First, we rewrite the average KL divergence and get

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}(\pi_N \| \pi_b) &= \frac{1}{N} \sum_{i=1}^{N} \mathrm{D}_{\mathrm{KL}}(\pi_i \| \pi_b) \\
&= \frac{1}{N} \sum_{i=1}^{N} \int_{-\infty}^{\infty} \pi_i(a) \log\left(\frac{\pi_i(a)}{\pi_b(a)}\right) \,\mathrm{da} \\
&= \frac{1}{N} \sum_{i=1}^{N} \int_{-\infty}^{\infty} \pi_i(a) \log\left(\pi_i(a)\right) \,\mathrm{da} - \frac{1}{N} \sum_{i=1}^{N} \int_{-\infty}^{\infty} \pi_i(a) \log\left(\pi_b(a)\right) \,\mathrm{da} \\
&= -\frac{1}{N} \sum_{i=1}^{N} \mathrm{H}(\pi_i) - \frac{1}{N} \sum_{i=1}^{N} \int_{-\infty}^{\infty} \pi_i(a) \log\left(\pi_b(a)\right) \,\mathrm{da},
\end{aligned}$$

where $\mathrm{H}(\pi_i)$ is the differential entropy of policy π_i. Note that, the first part is independent of the behavior policy π_b. Therefore, we will only consider the second part and get the following optimization problem

$$\max_{\pi_b} \frac{1}{N} \sum_{i=1}^{N} \int_{-\infty}^{\infty} \pi_i(a) \log\left(\pi_b(a)\right) \,\mathrm{da}.$$

We can write this as a Lagrange optimization problem with the constraint that $\int \pi_b(a) \,\mathrm{da} = 1$. The resulting Lagrange function is given by

$$\mathcal{L}(\pi_b(a), \lambda) \frac{1}{N} \sum_{i=1}^{N} \int_{-\infty}^{\infty} \pi_i(a) \log\left(\pi_b(a)\right) - \lambda \int \pi_b(a) \,\mathrm{da}.$$

Then, computing the derivative shows that

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \pi_b(a)} &= \frac{1}{N} \sum_{i=1}^{N} \frac{\pi_i(a)}{\pi_b(a)} - \lambda \stackrel{!}{=} 0 \\
\Leftrightarrow \frac{1}{\lambda N} \sum_{i=1}^{N} \pi_i(a) &= \pi_b(a)
\end{aligned}$$

Solving for λ gives, $\lambda = 1$ and therefore

$$\pi_b(a) = \frac{1}{N} \sum_{i=1}^{N} \pi_i(a).$$

□

This shows that we can translate our results to the continuous setting, we only need to introduce some further technical translations. In the case of density functions the importance weight is then defined by w_{IS}^{π} is then called the likelihood ratio and in the most general case, we need to introduce $\rho := \frac{d\pi_i}{d\pi_b}$, which is the probability density of π_i with respect to π_b, the so called *Radon- Nikodym derivative.*

C.3 Proof of Proposition 3.2

Proof. Recall, that by assuming R_*-subgaussian rewards, Hoeffding's inequality immediately implies

$$\mathbb{P}(\widehat{v}_n(\pi) - v(\pi) \geq \varepsilon) \leq \exp\left(-\frac{2\varepsilon^2 n}{(R_*\sigma_{\mathrm{IS}})^2}\right).$$

Hence, given $\delta \in (0,1)$, with probability at least $1-\delta$ we have

$$v(\pi) \geq \widehat{v}_n(\pi) - \frac{1}{\sqrt{2n}} R_*\sigma_{\mathrm{IS}}\sqrt{\log(\frac{N}{\delta})}$$

for all $\pi \in \Pi_N$. Similarly, it holds that

$$\widehat{v}_n(\pi) \geq v(\pi) - \frac{1}{\sqrt{2n}} R_*\sigma_{\mathrm{IS}}\sqrt{\log(\frac{N}{\delta})}$$

for all $\pi \in \Pi_N$. Furthermore, let $\delta \in (0,1)$ and $\widehat{\pi}_n := \arg\max_{i=1,\dots,N} \widehat{v}_n(\pi)$, then with probability at least $1-\delta$,

$$\begin{aligned} v(\widehat{\pi}_n) > \widehat{v}_n(\pi_n) - \frac{1}{\sqrt{2n}} R_*\sigma_{\mathrm{IS}}\sqrt{\log(\frac{N}{\delta})} &\geq \widehat{v}_n(\pi_*) - \frac{1}{\sqrt{2n}} R_*\sigma_{\mathrm{IS}}\sqrt{\log(\frac{N}{\delta})} \\ &> v(\pi_*) - \frac{2}{\sqrt{2n}} R_*\sigma_{\mathrm{IS}}\sqrt{\log(\frac{N}{\delta})}. \end{aligned}$$

It follows that

$$\mathcal{R}(\widehat{\pi}_n) < \frac{\sqrt{2}}{\sqrt{n}} R_*\sigma_{\mathrm{IS}}\sqrt{\log(\frac{N}{\delta})},$$

with probability at least $1-\delta$. □

C.4 Proof of Proposition 3.3

Proof. We define our multi-armed bandit model by $\mathcal{A} = (a_1, a_2)$, $\pi_1^{(N)} = (1 - \frac{1}{N}, \frac{1}{N})$ and $\pi_2^{(N)} = \dots = \pi_N^{(N)} = (\frac{1}{N}, 1 - 1/N)$. The behavior policy computes as

$$\pi_{\mathrm{KL}}(a_1) = \frac{2(N-1)}{N^2},\ \pi_{\mathrm{KL}}(a_2) = \frac{(N-1)^2+1}{N^2}.$$

Further, we consider deterministic rewards

$$R(a_1) := r_1,\ R(a_2) := r_2, \quad r_1 > r_2 \geq 0.$$

By the choice of r_1, r_2 as such that the policies satisfy that $v(\pi_1) > v(\pi_2) = \dots = v(\pi_N)$, which guarantees

$$\mathbb{P}(\mathcal{R}(\widehat{\pi}_n) > \varepsilon) \geq \mathbb{P}(\widehat{v}_n(\pi_2) > \widehat{v}_n(\pi_1)) = \mathbb{P}(\sum_{t=1}^{n} \frac{\pi_2(A_t) - \pi_1(A_t)}{\pi_{\mathrm{KL}}(A_t)} R(A_t) > 0) =: \mathbb{P}(\sum_{t=1}^{n} Z(A_t) > 0)$$

for all $0 \leq \varepsilon \leq v(\pi_1) - v(\pi_2)$. Next, we define

$$z_1 = \frac{\pi_2(a_1) - \pi_1(a_1)}{\pi_{\mathrm{KL}}(a_1)} r_1 = \frac{2N - N^2}{2(N-1)} r_1 < 0$$

and

$$z_2 = \frac{\pi_2(a_2) - \pi_1(a_2)}{\pi_{\mathrm{KL}}(a_2)} r_2 = \frac{N^2 - 2N}{(N-1)^2 + 1} r_2 > 0$$

for $N \geq 3$. Note that it holds that

$$\mathbb{P}(Z(A_t) = z_1) = \pi_{\mathrm{KL}}(a_1) = 2\frac{N-1}{N^2} := x_1 \quad \text{and} \quad \mathbb{P}(Z(A_t) = z_2) = \pi_{\mathrm{KL}}(a_2) = 1 - 2\frac{N-1}{N^2} := x_2,$$

such that

$$\tilde{Z}(A_t) := \frac{Z(A_t) - z_1}{z_2 - z_1} \sim \mathrm{Ber}(x_2).$$

Now, we define $r_2 = (1-\lambda)r_1$ for $\lambda > 0$. Note that, it immediately follows $r_1 > r_2$ and $v(\pi_1) > v(\pi_2)$. Furthermore, we get that

$$\begin{aligned} z_2 - z_1 &= \frac{1-\frac{2}{N}}{((N-1)^2+1)/N^2}(1-\lambda)r_1 + \frac{1-2/N}{2(N-1)/N^2} \\ &= r_1(1-\frac{2}{N})N^2\left(\frac{(1-\lambda)}{((N-1)^2+1)} + \frac{1}{2(N-1)}\right). \end{aligned}$$

It then follows

$$-\frac{z_1}{z_2-z_1} = \frac{1}{2(N-1)} \Big/ \left(\frac{1}{2(N-1)} + \frac{1-\lambda}{(N-1)^2+1}\right) = \frac{1}{1+2\frac{1-\lambda}{N-1+\frac{1}{N-1}}}.$$

$$\begin{aligned} \mathbb{P}\left(\sum_{t=1}^{n} \frac{\pi_2(A_t)-\pi_1(A_t)}{\pi_{\mathrm{KL}}(A_t)} R(A_t) > 0\right) &= \mathbb{P}\left(\sum_{t=1}^{n} \tilde{Z}(A_t) > -\frac{nz_1}{z_2-z_1}\right) \\ &\geq \frac{1}{\sqrt{2n}} \exp(-n\mathrm{D}(\frac{-z_1}{z_2-z_1}\|x_2)), \end{aligned} \tag{9}$$

where we applied Lemma 4.7.2 from Ash [1990], also documented for completeness just below the next proof.

With the definition of D, we get

$$\mathrm{D}(\frac{-z_1}{z_2-z_1}\|x_2)) := -\frac{z_1}{z_2-z_1}\log\left(-\frac{z_1}{z_2-z_1}\frac{1}{x_2}\right) + (1-\frac{-z_1}{z_2-z_1})\log\left(\frac{1-\frac{-z_1}{z_2-z_1}}{1-x_2}\right).$$

With the choice of $\lambda = 1/N - \frac{1}{N(N-1)} > 0$, for $N \geq 3$, we get

$$-\frac{z_1}{z_2-z_1} = \frac{1}{1+\frac{2}{N}}.$$

It then follows that

$$\begin{aligned} \mathbb{P}\left(\sum_{t=1}^{n} \frac{\pi_2(A_t)-\pi_1(A_t)}{\pi_{\mathrm{KL}}(A_t)} R(A_t) > 0\right) &= \mathbb{P}\left(\sum_{t=1}^{n} \tilde{Z}(A_t) > -\frac{nz_1}{z_2-z_1}\right) \\ &\geq \frac{1}{\sqrt{2n}} \exp(-n\mathrm{D}(\frac{-z_1}{z_2-z_1}\|x_2)) \\ &\geq \frac{1}{\sqrt{2n}} \exp\left(-n\left(\frac{1}{1+\frac{2}{N}}\log\left(\frac{1}{1-\frac{2}{N^2}+\frac{4}{N^3}}\right)\right)\right) \\ &= \frac{1}{\sqrt{2n}} \exp\left(-n\left(\frac{1}{1+\frac{2}{N}}\log\left(1+\frac{\frac{2}{N^2}-\frac{4}{N^3}}{1-\frac{2}{N^2}+\frac{4}{N^3}}\right)\right)\right) \\ &\geq \frac{1}{\sqrt{2n}} \exp\left(-n\left(\frac{1}{1+\frac{2}{N}}\frac{\frac{2}{N^2}-\frac{4}{N^3}}{1-\frac{2}{N^2}+\frac{4}{N^3}}\right)\right) \\ &= \frac{1}{\sqrt{2n}} \exp\left(-n\left(\frac{2(N-2)N}{(N+2)^2(N^2-2N+2)}\right)\right) \\ &\geq \frac{1}{\sqrt{2n}} \exp\left(-\frac{n}{2\sigma_N^2}\right) \end{aligned} \tag{10}$$

□

The idea of the following proof is the same compared to the last one. However, a different hardness bandit construction is needed.

Proof. Recall, that by assuming R_*-subgaussian rewards, Hoeffding's inequality immediately implies

$$\mathbb{P}(\widehat{v}_n(\pi) - v(\pi) \geq \varepsilon) \leq \exp\left(-\frac{2\varepsilon^2 n}{(R_*\sigma_{\mathrm{IS}})^2}\right).$$

Hence, given $\delta \in (0,1)$, with probability at least $1-\delta$ we have

$$v(\pi) \geq \widehat{v}_n(\pi) - \frac{1}{\sqrt{2n}} R_* \sigma_{\mathrm{IS}} \sqrt{\log(\frac{N}{\delta})}$$

for all $\pi \in \Pi_N$. Similarly, it holds that

$$\widehat{v}_n(\pi) \geq v(\pi) - \frac{1}{\sqrt{2n}} R_* \sigma_{\mathrm{IS}} \sqrt{\log(\frac{N}{\delta})}$$

for all $\pi \in \Pi_N$. Furthermore, let $\delta \in (0,1)$ and $\widehat{\pi}_n := \arg\max_{i=1,\dots,N} \widehat{v}_n(\pi)$, then with probability at least $1-\delta$,

$$\begin{aligned} v(\widehat{\pi}_n) > \widehat{v}_n(\pi_n) - \frac{1}{\sqrt{2n}} R_* \sigma_{\mathrm{IS}} \sqrt{\log(\frac{N}{\delta})} &\geq \widehat{v}_n(\pi_*) - \frac{1}{\sqrt{2n}} R_* \sigma_{\mathrm{IS}} \sqrt{\log(\frac{N}{\delta})} \\ &> v(\pi_*) - \frac{2}{\sqrt{2n}} R_* \sigma_{\mathrm{IS}} \sqrt{\log(\frac{N}{\delta})}. \end{aligned}$$

It follows that

$$\mathcal{R}(\widehat{\pi}_n) < \frac{\sqrt{2}}{\sqrt{n}} R_* \sigma_{\mathrm{IS}} \sqrt{\log(\frac{N}{\delta})},$$

with probability at least $1-\delta$. □

Next, we give the case for π_{χ^2}.

Proof. We define our multi-armed bandit model by $\mathcal{A} = (a_1, a_2, \dots, a_N)$, $\pi_1^{(N)} = (1,0,\dots,0), \pi_2^{(N)} = (0,1,\dots,0), \dots, \pi_N^{(N)} = (0,0,\dots,1)$. The behavior policy computes as

$$\pi_{\chi^2}(a_1) = \frac{1}{N} = \dots = \pi_{\chi^2}(a_N) = \frac{1}{N}.$$

Further, we consider deterministic rewards

$$R(a_1) := r_1,\ R(a_2) := r_2, \quad r_1 > r_2 \geq 0.$$

By the choice of r_1, r_2 as such that the policies satisfy that $v(\pi_1) > v(\pi_2) = \dots = v(\pi_N)$, which guarantees

$$\mathbb{P}(\mathcal{R}(\widehat{\pi}_n) > \varepsilon) \geq \mathbb{P}(\widehat{v}_n(\pi_2) > \widehat{v}_n(\pi_1)) = \mathbb{P}(\sum_{t=1}^{n} \frac{\pi_2(A_t) - \pi_1(A_t)}{\pi_{\chi^2}(A_t)} R(A_t) > 0) =: \mathbb{P}(\sum_{t=1}^{n} Z(A_t) > 0)$$

for all $0 \leq \varepsilon \leq v(\pi_1) - v(\pi_2)$. Next, we define

$$z_1 = \frac{\pi_2(a_1) - \pi_1(a_1)}{\pi_{\mathrm{KL}}(a_1)} r_1 = \frac{0-1}{1/N} r_1 = -N r_1 < 0$$

and

$$z_2 = \frac{\pi_2(a_2) - \pi_1(a_2)}{\pi_{\chi^2}(a_2)} r_2 = N r_2 > 0$$

for $N \geq 3$. Note that it holds that

$$\mathbb{P}(Z(A_t) = z_1) = \pi_{\chi^2}(a_1) = \frac{1}{N} := x_1 \quad \text{and} \quad \mathbb{P}(Z(A_t) = z_2) = 1 - \pi_{\chi^2}(a_1) = 1 - \frac{1}{N} := x_2,$$

such that

$$\tilde{Z}(A_t) := \frac{Z(A_t) - z_1}{z_2 - z_1} \sim \mathrm{Ber}(x_2).$$

Now, we define $r_2 = (1-\lambda)r_1$ for $\lambda > 0$. Note that, it immediately follows $r_1 > r_2$ and $v(\pi_1) > v(\pi_2)$. Furthermore, we get that

$$z_2 - z_1 = Nr_1(2-\lambda).$$

It then follows

$$-\frac{z_1}{z_2 - z_1} = \frac{Nr_1}{Nr_1(2-\lambda)} = \frac{1}{2-\lambda}.$$

$$\begin{aligned}\mathbb{P}\left(\sum_{t=1}^{n} \frac{\pi_2(A_t) - \pi_1(A_t)}{\pi_{\mathrm{KL}}(A_t)} R(A_t) > 0\right) &= \mathbb{P}\left(\sum_{t=1}^{n} \tilde{Z}(A_t) > -\frac{nz_1}{z_2 - z_1}\right) \\ &\geq \frac{1}{\sqrt{2n}} \exp(-n\mathrm{D}(\frac{-z_1}{z_2 - z_1} \| x_2)),\end{aligned} \tag{11}$$

where we applied Lemma 4.7.2 from Ash [1990], also documented for completeness just below this proof.

With the definition of D, we get

$$\mathrm{D}(\frac{-z_1}{z_2 - z_1} \| x_2)) := -\frac{z_1}{z_2 - z_1} \log\left(-\frac{z_1}{z_2 - z_1}\frac{1}{x_2}\right) + (1 - \frac{-z_1}{z_2 - z_1}) \log\left(\frac{1 - \frac{-z_1}{z_2 - z_1}}{1 - x_2}\right).$$

With the choice of $\lambda = 1 - \frac{2}{N^2} > 0$, for $N \geq 3$, we get

$$-\frac{z_1}{z_2 - z_1} = \frac{1}{(1 + \frac{2}{N^2})}.$$

It then follows that

$$\begin{aligned}\mathbb{P}\left(\sum_{t=1}^{n} \frac{\pi_2(A_t) - \pi_1(A_t)}{\pi_{\mathrm{KL}}(A_t)} R(A_t) > 0\right) &= \mathbb{P}\left(\sum_{t=1}^{n} \tilde{Z}(A_t) > -\frac{nz_1}{z_2 - z_1}\right) \\ &\geq \frac{1}{\sqrt{2n}} \exp(-n\mathrm{D}(\frac{-z_1}{z_2 - z_1} \| x_2)) \\ &\geq \frac{1}{\sqrt{2n}} \exp\left(-n\left(\frac{1}{1 + \frac{2}{N^2}} \log\left(\frac{1}{1 + \frac{2}{N^2} - \frac{1}{N} + \frac{2}{N^3}}\right)\right)\right) \\ &= \frac{1}{\sqrt{2n}} \exp\left(-n\left(\frac{1}{1 + \frac{2}{N^2}} \log\left(1 + \frac{N^2 - 2N - 2}{N^3 - N^2 + 2N + 2}\right)\right)\right) \\ &\geq \frac{1}{\sqrt{2n}} \exp\left(-n\left(\frac{1}{1 + \frac{2}{N^2}} \frac{N^2 - 2N - 2}{N^3 - N^2 + 2N + 2}\right)\right) \\ &= \frac{1}{\sqrt{2n}} \exp\left(-n\left(\frac{N^4 - 2N^3 - 2N^2}{N^5 - N^4 + 4N^3 + 4N + 4}\right)\right) \\ &\geq \frac{1}{\sqrt{2n}} \exp\left(-\frac{n}{2\sigma_N}\right)\end{aligned} \tag{12}$$

□

Lemma C.4 (Lemma 4.7.2 in Ash [1990])**.** *For $k > np$ it holds true that*

$$\mathbb{P}(\mathrm{Bin(n,p)} \geq k) \geq \frac{1}{\sqrt{8k(1 - k/n)}} \exp(-n\mathrm{D}(k/n \| p)),$$

where $\mathrm{D}(\cdot\|\cdot)$ is the Binary entropy function.

C.5 Proof of Proposition 3.6

Proof. First, note that the χ^2 divergence and KL divergence can be lower bounded by the squared Hellinger distance in the sense that

$$D_H^2(\pi_i, \pi_b) = \frac{1}{2}\sum_{a\in\mathcal{A}}\left(\sqrt{\pi_i(a)} - \sqrt{\pi_b(a)}\right)^2 \leq \frac{1}{2}D_{\mathrm{KL}}(\pi_i \mid \pi_b) \leq \chi^2(\pi_i \mid \pi_b).$$

Hence, by assumption it holds that $D_H^2(\pi_i, \pi_b) \leq \eta$ implying that, for all $a \in \mathcal{A}$,

$$\left(\sqrt{\pi_i(a)} - \sqrt{\pi_b(a)}\right)^2 \leq 2\eta\,. \tag{13}$$

Let $i \in \{1, \dots, N\}$ and $a \in \mathcal{A}$ be arbitrary. First note that $\frac{\pi_i(a)}{\pi_b(a)} \leq N$ is always satisfied by construction of π_b. For $\pi_i(a) \leq \pi_b(a)$ we obviously have $\frac{\pi_i(a)}{\pi_b(a)} \leq 1$. Next, consider the case $\pi_i(a) > \pi_b(a)$: Using (13), we have

$$0 \leq \sqrt{\pi_i(a)} \leq \sqrt{2\eta} + \sqrt{\pi_b(a)}\,.$$

This implies that

$$\frac{\sqrt{\pi_i(a)}}{\sqrt{\pi_b(a)}} \leq \frac{\sqrt{2\eta} + \sqrt{\pi_b(a)}}{\sqrt{\pi_b(a)}} \leq 1 + \frac{\sqrt{2\eta}}{\sqrt{\min_a \pi_b(a)}}\,.$$

The assertion follows by taking the square on both sides. □

C.6 Upper bound of the weights for the safe policy

Proposition C.5. *Suppose that $D_\iota(\pi_i \mid \pi_b) \leq \eta$ for $\iota \in \{\mathrm{KL}, \chi^2\}$ for all $i \in \{1, \dots, N\}$ and $\pi_b(a) > 0$ for all $a \in \mathcal{A}$. For any $\lambda \in (0, 1)$ it holds that*

$$\sigma_{\mathrm{safe}} := \max_{a\in\mathcal{A}} \frac{\pi_i(a)}{\pi_{\mathrm{safe}}^\lambda(a)} \leq \min\left(\frac{N}{1-\lambda}, \frac{1}{1-\lambda} + \frac{2\eta K}{\lambda} + \frac{2\sqrt{2\eta K}}{\sqrt{1-\lambda}\sqrt{\lambda}}\right).$$

Moreover, suppose that $\eta < 1$ and let $\lambda(\eta) := \sqrt{\eta}$, then

$$\sigma_{\mathrm{safe}} \leq \min\left(\frac{N}{1-\sqrt{\eta}}, \frac{1}{1-\sqrt{\eta}} + 2\sqrt{\eta}K + \frac{2\sqrt{2K\sqrt{\eta}}}{\sqrt{1-\sqrt{\eta}}}\right).$$

Proof. Recall, that the χ^2 divergence and the KL divergence is lower bounded by the squared Hellinger distance. Let $i \in \{1, \dots, N\}$ and $a \in \mathcal{A}$ be arbitrary. First, in the case $\pi_i(a) \leq \pi_{\mathrm{KL}}(a)$, we have

$$\frac{\pi_i(a)}{\pi_{\mathrm{safe}}^\lambda(a)} \leq \frac{\pi_b(a)}{(1-\lambda)\pi_b(a) + \lambda/K} \leq \frac{1}{1-\lambda}\,.$$

In the case $\pi_i(a) > \pi_b(a)$, it holds that

$$0 \leq \sqrt{\pi_i(a)} \leq \sqrt{2\eta} + \sqrt{\pi_b(a)}\,,$$

and therefore, we have

$$\begin{aligned}
\frac{\pi_i(a)}{\pi_{\mathrm{safe}}^\lambda(a)} &\leq \frac{2\eta + 2\sqrt{2\eta}\sqrt{\pi_b(a)} + \pi_b(a)}{(1-\lambda)\pi_b(a) + \lambda/K} \\
&\leq \frac{2\eta K}{\lambda} + \frac{2\sqrt{2\eta}\sqrt{\pi_b(a)}}{\sqrt{(1-\lambda)\pi_b(a) + \lambda/K}\sqrt{\lambda/K}} + \frac{\pi_b(a)}{(1-\lambda)\pi_b(a) + \lambda/K} \\
&\leq \frac{2\eta K}{\lambda} + 2\sqrt{2\eta}\frac{\sqrt{(1-\lambda)\pi_b(a)}}{\sqrt{(1-\lambda)\pi_b(a) + \lambda/K}}\frac{\sqrt{K}}{\sqrt{1-\lambda}\sqrt{\lambda}} + \frac{1}{1-\lambda} \\
&\leq \frac{2\eta K}{\lambda} + 2\sqrt{2\eta}\frac{\sqrt{K}}{\sqrt{1-\lambda}\sqrt{\lambda}} + \frac{1}{1-\lambda}
\end{aligned}$$

which verifies the first claim. The second claim is a direct consequence. □

D Proofs and additional details of Section 4

In this section, we provide the omitted proofs and additional discussions of Section 4.

D.1 Proof of Proposition 4.2

Proof. First, observe that

$$\begin{aligned}\mathbb{P}(\widehat{\pi}_n \notin K_1) &= \mathbb{P}(\bigcup_{j=2,\dots,M}\{\max_{\pi\in K_j}\widehat{v}_n^{(j)}(\pi) - \max_{\pi\in K_1}\widehat{v}_n^{(1)}(\pi) \geq \Delta\}) \\ &\leq \sum_{j=2}^{M}\mathbb{P}(\max_{\pi\in K_j}\widehat{v}_n^{(j)}(\pi) - \max_{\pi\in K_1}\widehat{v}_n^{(1)}(\pi) \geq \min(\Delta,\varepsilon)) \\ &\leq \sum_{j=2}^{M}\mathbb{P}(\max_{\pi\in K_j}\widehat{v}_n^{(j)}(\pi) - \max_{\pi\in K_1}\widehat{v}_n^{(1)}(\pi) \geq \varepsilon)\end{aligned}$$

and we proceed by considering each probability in the last line separately. Let $j \in \{2,\dots,M\}$ and decompose as follows

$$\begin{aligned}\max_{\pi\in K_j}\widehat{v}_n^{(j)}(\pi) - \max_{\pi\in K_1}\widehat{v}_n^{(1)}(\pi) &= \max_{\pi\in K_j}\widehat{v}_n^{(j)}(\pi) - v(\pi_*^{(j)}) + v(\pi_*^{(j)}) - v(\pi_*^{(1)}) \\ &\quad + v(\pi_*^{(1)}) - \max_{\pi\in K_1}\widehat{v}_n^{(1)}(\pi) \\ &\leq \max_{\pi\in K_j}\widehat{v}_n^{(j)}(\pi) - v(\pi_*^{(j)}) + v(\pi_*^{(1)}) - \max_{\pi\in K_1}\widehat{v}_n^{(1)}(\pi)\end{aligned}$$

almost surely due to the assumption $v(\pi_*^{(j)}) - v(\pi_*^{(1)}) < 0$. Hence, we have

$$\begin{aligned}\mathbb{P}(\max_{\pi\in K_j}\widehat{v}_n^{(j)}(\pi) - \max_{\pi\in K_1}\widehat{v}_n^{(1)}(\pi) \geq \varepsilon) &\leq \mathbb{P}(|\max_{\pi\in K_j}\widehat{v}_n^{(j)}(\pi) - v(\pi_*^{(j)})| \geq \varepsilon/2) \\ &\quad + \mathbb{P}(|v(\pi_*^{(1)}) - \max_{\pi\in K_1}\widehat{v}_n^{(1)}(\pi)| \geq \varepsilon/2)\,.\end{aligned}$$

For all $j \in \{1,\dots,M\}$ it holds that

$$\begin{aligned}\mathbb{P}(|\max_{\pi\in K_j}\widehat{v}_n^{(j)}(\pi) - v(\pi_*^{(j)})| \geq \tfrac{\varepsilon}{2}) &\leq \mathbb{P}(\max_{\pi\in K_j}\widehat{v}_n^{(j)}(\pi) - v(\pi_*^{(j)}) \geq \tfrac{\varepsilon}{2}) + \mathbb{P}(v(\pi_*^{(j)}) - \max_{\pi\in K_j}\widehat{v}_n^{(j)}(\pi) \geq \tfrac{\varepsilon}{2}) \\ &\leq \mathbb{P}(\bigcup_{\pi\in K_j}\{\widehat{v}_n^{(j)}(\pi) - v(\pi_*^{(j)}) \geq \tfrac{\varepsilon}{2}\}) + \mathbb{P}(v(\pi_*^{(j)}) - \widehat{v}_n^{(j)}(\pi_*^{(j)}) \geq \tfrac{\varepsilon}{2}) \\ &\leq (N_j+1)\exp(-\frac{\varepsilon^2 n_j}{2R_*^2\sigma_{\mathrm{c}}^2}) \\ &= (N_j+1)\exp(-\frac{\varepsilon^2 n}{2MR_*^2\sigma_{\mathrm{c}}^2})\,,\end{aligned}$$

where we have used Hoeffding's inequality for R_*-subgaussian rewards. In total we obtain

$$\begin{aligned}\mathbb{P}(\widehat{\pi}_n \notin K_1) &\leq (M-1)(N_1+1)\exp(-\frac{\varepsilon^2 n}{2MR_*^2\sigma_{\mathrm{c}}^2}) + \exp(-\frac{\varepsilon^2 n}{2MR_*^2\sigma_{\mathrm{c}}^2})\sum_{j=2}^{M}(N_j+1) \\ &= \exp(-\frac{\varepsilon^2 n}{2MR_*^2\sigma_{\mathrm{c}}^2})(N+M+(M-2)(N_1+1)\end{aligned}$$

which is bounded by δ by the choice $n \geq n(\varepsilon,\delta)$. □

Remark D.1. *The derived upper bound could be further tightened by leveraging the fact that all policies within a cluster are inherently "similar". Specifically, we could refine the bound by incorporating the observation that the difference $|\hat{v}(\pi) - \hat{v}(\tilde{\pi})|$ remains small for any pair of policies π and $\tilde{\pi}$ within the same cluster. However, this refinement would only lead to an improvement in the logarithmic factor of the bound. Given its limited impact on the overall result, we omit this adjustment for simplicity.*

D.2 Proof of Theorem 4.3

Proof. Let $\pi_* = \arg\max_{\pi \in \Pi_N} v(\pi)$ and $\widehat{\pi}_n$ be defined in (6). Using the law of total probability we obtain

$$\begin{aligned}
\mathbb{P}(v(\pi_*) - v(\widehat{\pi}_n) \geq \varepsilon) &= \mathbb{P}(\widehat{\pi}_n \notin K_1)\mathbb{P}(v(\pi_*) - v(\widehat{\pi}_n) \geq \varepsilon \mid \widehat{\pi}_n \notin K_1) \\
&\quad + \mathbb{P}(\widehat{\pi}_n \in K_1)\mathbb{P}(v(\pi_*) - v(\widehat{\pi}_n) \geq \varepsilon \mid \widehat{\pi}_n \in K_1) \\
&\leq \mathbb{P}(\widehat{\pi}_n \notin K_1) + \mathbb{P}(v(\pi_*) - v(\widehat{\pi}_n) \geq \varepsilon \mid \widehat{\pi}_n \in K_1) \\
&\leq \mathbb{P}(\widehat{\pi}_n \notin K_1) + \mathbb{P}(|v(\pi_*) - \widehat{v}_n^{(1)}(\pi_*)| \geq \varepsilon/2 \mid \widehat{\pi}_n \in K_1) \\
&\quad + \mathbb{P}(|\widehat{v}_n^{(1)}(\pi_*) - v(\widehat{\pi}_n)| \geq \varepsilon/2 \mid \widehat{\pi}_n \in K_1).
\end{aligned}$$

Firstly, by the proof of Proposition 4.2 we have

$$\mathbb{P}(\widehat{\pi}_n \notin K_1) \leq (N + M + (M-2)(N_1+1)) \exp(-\frac{\varepsilon^2 n}{2MR_*^2\sigma_{\mathrm{c}}^2}).$$

Secondly, due to Assumption 4.1, $\pi_* \in K_1$, we have

$$\mathbb{P}(|v(\pi_*) - \widehat{v}_n^{(1)}(\pi_*)| \geq \varepsilon/2 \mid \pi_n \in K_1) = \mathbb{P}(|v(\pi_*) - \widehat{v}_n^{(1)}(\pi_*)| \geq \varepsilon/2) \leq 2\exp(-\frac{\varepsilon^2 n}{2MR_*^2\sigma_{\mathrm{c}}^2}),$$

where we have again used Hoeffding's inequality and the choice $n = n(\varepsilon, \delta)$. Finally, we consider the last term

$$\begin{aligned}
\mathbb{P}(|\widehat{v}_n^{(1)}(\pi_*) - v(\widehat{\pi}_n)| \geq \varepsilon/2 \mid \widehat{\pi}_n \in K_1) &\leq \mathbb{P}(\widehat{v}_n^{(1)}(\pi_*) - v(\widehat{\pi}_n) \geq \varepsilon/2 \mid \widehat{\pi}_n \in K_1) \\
&\quad + \mathbb{P}(v(\widehat{\pi}_n) - \widehat{v}_n^{(1)}(\pi_*) \geq \varepsilon/2 \mid \widehat{\pi}_n \in K_1).
\end{aligned}$$

Using the fact $\widehat{v}_n^{(1)}(\pi_*) \leq \widehat{v}_n^{(1)}(\widehat{\pi}_n)$ conditioned on $\widehat{\pi}_n \in K_1$, it holds that

$$\begin{aligned}
\mathbb{P}(\widehat{v}_n^{(1)}(\pi_*) - v(\widehat{\pi}_n) \geq \varepsilon/2 \mid \widehat{\pi}_n \in K_1) &\leq \mathbb{P}(\widehat{v}_n^{(1)}(\widehat{\pi}_n) - v(\widehat{\pi}_n) \geq \varepsilon/2 \mid \widehat{\pi}_n \in K_1) \\
&\leq \mathbb{P}(\bigcup_{\pi \in K_1} \{\widehat{v}_n^{(1)}(\pi) - v(\pi) \geq \varepsilon/2\} \mid \widehat{\pi}_n \in K_1) \\
&= \mathbb{P}(\bigcup_{\pi \in K_1} \{\widehat{v}_n^{(1)}(\pi) - v(\pi) \geq \varepsilon/2\}) \\
&\leq \sum_{\pi \in K_1} \mathbb{P}(\widehat{v}_n^{(1)}(\pi) - v(\pi) \geq \varepsilon/2) \\
&\leq N_1 \exp(-\frac{\varepsilon^2 n}{2MR_*^2\sigma_{\mathrm{c}}^2}),
\end{aligned}$$

where we have used Hoeffding's inequality. Similarly, since $v(\widehat{\pi}_n) \leq v(\pi_*)$ almost surely, by Hoeffding's inequality we obtain

$$\begin{aligned}
\mathbb{P}(v(\widehat{\pi}_n) - \widehat{v}_n^{(1)}(\pi_*) \geq \varepsilon/2 \mid \widehat{\pi}_n \in K_1) &\leq \mathbb{P}(v(\pi_*) - \widehat{v}_n^{(1)}(\pi_*) \geq \varepsilon/2 \mid \widehat{\pi}_n \in K_1) \\
&= \mathbb{P}(v(\pi_*) - \widehat{v}_n^{(1)}(\pi_*) \geq \varepsilon/2) \\
&\leq \exp(-\frac{\varepsilon^2 n}{2MR_*^2\sigma_{\mathrm{c}}^2}).
\end{aligned}$$

Overall, we proved that

$$\begin{aligned}
\mathbb{P}(v(\pi_*) - v(\widehat{\pi}_n) \geq \varepsilon) &\leq (N_1 + 3 + N + M + (M-2)(N_1+1)) \exp(-\frac{\varepsilon^2 n}{2MR_*^2\sigma_{\mathrm{c}}^2}) \\
&= (2 + N + M + (M-1)(N_1+1)) \exp(-\frac{\varepsilon^2 n}{2MR_*^2\sigma_{\mathrm{c}}^2}).
\end{aligned}$$

Choosing $n = n(\varepsilon, \delta) = \frac{2MR_*^2\sigma_{\mathrm{c}}^2 \log(\frac{2+N+M+(M-1)(N_1+1)}{\delta})}{\varepsilon^2}$ the righthand side simplifies to δ and the claim follows. □

D.3 Problem-independent expected regret bound

Corollary D.2. *Suppose that Assumption 4.1 is in place, and $n_1 = \cdots = n_M$. For any $M \in \{1, \ldots, N\}$ the regret is bounded by*

$$\mathbb{E}\left[\mathcal{R}(\hat{\pi}_n))\right] \leq \frac{\Delta_{\max}}{\sqrt{n}} \left(1 + \frac{MN_1}{N} + \frac{2M}{N}\right) + \sqrt{2} M^{3/2} R_* \sigma_c \sqrt{\frac{\log(N\sqrt{n})}{n}},$$

where $n = M \cdot n_1$.

Proof. Let $\Delta_j := v(\pi_*) - v(\pi_*^{(j)})$, $j = \{1, \ldots, M\}$. Without loss of generality assume that $\Delta_1 = 0$ and $\Delta_j > 0$ for all $j = \{2, \ldots, M\}$ with $\Delta_j \leq \max_j \Delta_j =: \Delta_{\max}$. For arbitrary $\eta > 0$, we have

$$\begin{aligned}
\mathbb{E}\left[\mathcal{R}(\hat{\pi}_n))\right] &= \mathbb{E}[v(\pi_*) - v(\widehat{\pi}_n)] \\
&= \sum_{j:\Delta_j > \eta} \mathbb{E}[(v(\pi_*) - v(\widehat{\pi}_n))\mathbb{1}_{\widehat{\pi}_n \in K_j}] + \sum_{j:\Delta_j < \eta} \mathbb{E}[(v(\pi_*) - v(\widehat{\pi}_n))\mathbb{1}_{\widehat{\pi}_n \in K_j}] \\
&\leq \sum_{j:\Delta_j > \eta} \Delta_{\max} \mathbb{P}(\hat{\pi}_n \in K_j) + \sum_{j:\Delta_j < \eta} \eta \mathbb{P}(\hat{\pi}_n \in K_j) \\
&\leq \sum_{j:\Delta_j > \eta} \Delta_{\max} \mathbb{P}(\max_{\pi \in K_j} \hat{v}_n(\pi) > \max_{\pi \in K_1} \hat{v}_n(\pi)) + \eta M,
\end{aligned}$$

where we have used $\sum_{j=1}^{M} \mathbb{1}_{\widehat{\pi}_n \in K_j} = 1$ almost surely. By the proof of Proposition 4.2, for $j \in \{2, \ldots, M\}$ with $\Delta_j > \eta$ we have

$$\mathbb{P}(\max_{\pi \in K_j} \hat{v}_n(\pi) > \max_{\pi \in K_1} \hat{v}_n(\pi)) \leq (N_j + N_1 + 2) \exp\left(-\frac{\eta^2 n}{2MR_*^2\sigma_c^2}\right).$$

Thus, for the choice of

$$\eta = \sqrt{\frac{\log(N\sqrt{n}) 2MR_*^2\sigma_c^2}{n}}$$

we achieve an overall expected regret bound

$$\begin{aligned}
\mathbb{E}\left[\mathcal{R}(\hat{\pi}_n))\right] &\leq \sum_{j:\Delta_j > \eta} \Delta_{\max}(N_j + N_1 + 2) \exp\left(-\frac{\eta^2 n}{2MR_*^2\sigma_c^2}\right) + \eta M \\
&\leq \Delta_{\max}(N + MN_1 + 2M) \exp\left(-\frac{\eta^2 n}{2MR_*^2\sigma_c^2}\right) + \eta M \\
&\leq \frac{\Delta_{\max}}{\sqrt{n}} \left(1 + \frac{MN_1}{N} + \frac{2M}{N}\right) + \sqrt{2} M^{3/2} R_* \sigma_c \sqrt{\frac{\log(N\sqrt{n})}{n}}.
\end{aligned}$$

□

E Details on experimental setup

In this section, we provide details on the experimental setup and the full algorithm for the clustering procedure.

The clustering algorithm is given in Algorithm 2.

The experiments are conducted by following the procedure below:

1. We construct a multi-armed bandit environment by defining a reward distribution and the number of arms. The reward distribution for each arm follows a Gaussian distribution. Specifically, we set the number of arms to 100, with the highest mean reward of 3 assigned to arm 1, decreasing linearly by 0.05 per arm until arm 100. The variance is sampled uniformly and independent from $(1, 3)$.

Algorithm 2 Hellinger-based clustering for behavior policy design

Require: Number of clusters M, set of target policies Π_N
Ensure: Cluster assignments and computed behavior policies $\{\pi^1_{\mathrm{KL}}, \ldots, \pi^M_{\mathrm{KL}}\}$

Initialize an empty set $\Pi_{\mathrm{sqrttargets}}$
for $\pi^i_{\text{target}} \in \Pi_N$:
    Compute element-wise square root: $\pi^i_{\text{sqrttarget}} \leftarrow \sqrt{\pi^i_{\text{target}}}$
    Add $\pi^i_{\text{sqrttarget}}$ to $\Pi_{\mathrm{sqrttargets}}$
end
Apply KMeans clustering to $\Pi_{\mathrm{sqrttargets}}$:
clusters $\leftarrow$ KMeans$(\Pi_{\mathrm{sqrttargets}}, M, \text{Metric} = D^2_{\mathrm{H}})$
for $j = 1$ to M:
    Compute behavior policies π^j_b according to chosen f-divergence, ensuring it remains a valid probability distribution
end

2. We set the number of target policies to $N = 1000$. For each policy and each arm, we sample a weight uniformly from $(1, 2)$. To introduce structured dependencies among target policies, we form groups of policies that prioritize specific arms by adding an additional random weight sampled uniformly from $(1, 10)$ on those arms. In particular, we create 6 groups of the following sizes $[25, 50, 25, 825, 50, 25]$, with preferred arms $[[2], [3, 5], [22, 24, 34], [23, 99], [99], [53]]$. To ensure that every policy assigns positive probability to all arms, we transform the sampled weights into softmax policies using a temperature parameter of 1.
3. Given the set of target policies, we apply Algorithm 2 to obtain the KL-barycenters and the χ^2-barycenters corresponding to different clusters.
4. Finally, we compute the importance sampling estimates for all target policies using a fixed sample size. In the case of clustering, the samples are distributed uniformly across the clusters.