Logarithmic Smoothing for Pessimistic Off-Policy Evaluation, Selection and Learning

Otmane Sakhi
Criteo AI Lab, Paris, France
o.sakhi@criteo.com

Imad Aouali
CREST, ENSAE
Criteo AI Lab, Paris, France
i.aouali@criteo.com

Pierre Alquier
ESSEC Business School, Singapore
alquier@essec.edu

Nicolas Chopin
CREST, ENSAE
nicolas.chopin@ensae.fr

Abstract

This work investigates the offline formulation of the contextual bandit problem, where the goal is to leverage past interactions collected under a behavior policy to evaluate, select, and learn new, potentially better-performing, policies. Motivated by critical applications, we move beyond point estimators. Instead, we adopt the principle of *pessimism* where we construct upper bounds that assess a policy's worst-case performance, enabling us to confidently select and learn improved policies. Precisely, we introduce novel, fully empirical concentration bounds for a broad class of importance weighting risk estimators. These bounds are general enough to cover most existing estimators and pave the way for the development of new ones. In particular, our pursuit of the tightest bound within this class motivates a novel estimator (LS), that *logarithmically smooths* large importance weights. The bound for LS is provably tighter than its competitors, and naturally results in improved policy selection and learning strategies. Extensive policy evaluation, selection, and learning experiments highlight the versatility and favorable performance of LS.

1 Introduction

In decision-making under uncertainty, offline contextual bandit [16] presents a practical framework for leveraging past interactions with an environment to optimize future decisions. This comes into play when we possess logged data summarizing an agent's past interactions [10]. These interactions, typically captured as context-action-reward tuples, hold valuable insights into the underlying dynamics of the environment. Each tuple represents a single round of interaction, where the agent observes a context (including relevant features), takes an action according to its current policy, often called *behavior policy*, and receives a reward that depends on both the observed context and the taken action. This framework is prevalent in interactive systems like online advertising, music streaming, and video recommendation. In online advertising, for instance, the user's profile is the context, the recommended product is the action, and the click-through rate (CTR) is the expected reward. By learning from past interactions, the recommender system tailors product suggestions to individual preferences, maximizing engagement and ultimately, business success.

To optimize future decisions without requiring real-time deployments, this framework presents us with three tasks: off-policy evaluation (OPE) [16], off-policy selection (OPS) [32], and off-policy learning (OPL) [55]. OPE estimates the risk: the *negative of expected reward* that a *target policy* would achieve, essentially predicting its performance if deployed. OPS selects the best-performing

policy from a finite set of options, and OPL finds the optimal policy within an infinite class of policies. In general, OPE is an intermediary step for OPS and OPL since its primary goal is policy comparison.

A significant amount of research in OPE has centered around Inverse Propensity Scoring (IPS) estimators [24, 16–18, 60, 19, 54, 38, 32, 45]. These estimators rely on importance weighting to address the discrepancy between the target and behavior policies. While unbiased under some conditions, IPS induces high variance. To mitigate this, regularization techniques have been proposed for IPS [10, 38, 54, 5, 21] trading some bias for reduced variance. However, these estimators can still deviate from the true risk, undermining their reliability for decision-making, especially in critical applications. In such scenarios, practitioners need estimates that cover the true risk with high confidence. To address this, several approaches focused on constructing either asymptotic [10, 48, 15] or finite sample [32, 21], high probability, empirical upper bounds on the risk. These bounds evaluate the performance of a policy in the worst-case scenario, adopting the principle of pessimism [27].

If this principle is used in OPE, it is central in OPS and OPL, where strategies are inspired by, or directly derived from, upper bounds on the risk [55, 35, 32, 49, 5, 59, 21]. Examples for OPS include Kuzborskij et al. [32] who employed an Efron-Stein bound for self-normalized IPS, or Gabbianelli et al. [21] that based their analysis on an upper bound constructed with the Implicit Exploration estimator. Focusing on OPL, Swaminathan and Joachims [55] exploited the empirical Bernstein bound [36] alongside the Clipping estimator to motivate sample variance penalization. This work was recently improved by either modifying the penalization [59] or analyzing the problem from the PAC-Bayesian lens [35]. The latter direction was further explored by Sakhi et al. [49], Aouali et al. [5, 7], Gabbianelli et al. [21] resulting in tight PAC-Bayesian bounds that can be directly optimized.

Existing *pessimistic* OPE, OPS, and OPL approaches involve analyzing the concentration properties of a *pre-defined risk estimator*, often chosen to simplify the analysis. We propose a different approach: we derive general concentration bounds applicable to a broad class of regularized IPS estimators and then identify the estimator within this class that achieves the tightest concentration bound. This leads to a tailored estimator, named Logarithmic Smoothing (LS). LS enjoys several desirable properties. It concentrates at a sub-Gaussian rate, and has a finite variance without being necessarily bounded. Its concentration upper bound allows us to evaluate the worst-case risk of any policy, enables us to derive a simple OPS strategy that directly minimizes our estimator akin to Gabbianelli et al. [21], and achieves state-of-the-art learning guarantees for OPL when analyzed within the PAC-Bayesian framework akin to [35, 49, 5, 7, 21].

This paper is structured as follows. Section 2 introduces the necessary background. In Section 3, we provide unified risk bounds for a broad class of regularized IPS estimators, for which LS enjoys the tightest upper bound. In Section 4, we analyze LS for OPS and OPL, and we further extend the analysis within the PAC-Bayesian framework. Extensive experiments in Section 5 highlight the favorable performance of LS, and Section 6 provides concluding remarks.

2 Setting and background

Offline contextual bandit. Let $\mathcal{X} \subset \mathbb{R}^d$ be the *context space*, which is a compact subset of $\mathbb{R}^d$, and let $\mathcal{A} = [K]$ be a finite *action set*. An agent's actions are guided by a *stochastic* and *stationary* policy $\pi \in \Pi$ within a policy space Π. Given a context $x \in \mathcal{X}$, $\pi(\cdot|x)$ is a probability distribution over the action set $\mathcal{A}$; $\pi(a|x)$ is the probability that the agent selects action a in context x. Then, an agent interacts with a contextual bandit over n rounds. In round $i \in [n]$, the agent observes a context $x_i \sim \nu$ where ν is a distribution with support $\mathcal{X}$. After this, the agent selects an action $a_i \sim \pi_0(\cdot|x_i)$, where π_0 is the *behavior policy* of the agent. Finally, the agent receives a stochastic cost $c_i \in [-1, 0]$ that depends on the observed context x_i and the taken action a_i. This cost c_i is sampled from a cost distribution $p(\cdot|x_i, a_i)$. This leads to n-sized logged data, $\mathcal{D}_n = (x_i, a_i, c_i)_{i \in [n]}$, where tuples (x_i, a_i, c_i) for $i \in [n]$ are i.i.d. The expected cost of taking action a in context x is $c(x, a) = \mathbb{E}_{c \sim p(\cdot|x,a)}[c]$, and the costs are negative because they are interpreted as the negative of rewards. The performance of a policy $\pi \in \Pi$ is evaluated through its *risk*, which aggregates the expected costs $c(x, a)$ over all possible contexts $x \in \mathcal{X}$ and taken actions $a \in \mathcal{A}$ by policy π, such as

$$R(\pi) = \mathbb{E}_{x \sim \nu, a \sim \pi(\cdot|x), c \sim p(\cdot|x,a)}[c] = \mathbb{E}_{x \sim \nu, a \sim \pi(\cdot|x)}[c(x, a)] . \tag{1}$$

The main goal is to use logged dataset $\mathcal{D}_n$ to enhance future decision-making without necessitating live deployments. This often entails three tasks: OPE, OPS, and OPL. First, OPE is concerned

with constructing an estimator $\hat{R}_n(\pi)$ of the risk $R(\pi)$ of a fixed *target policy* π and study its deviation, aspiring for $\hat{R}_n(\pi)$ to concentrate well around $R(\pi)$. Second, OPS focuses on selecting the best performing policy $\hat{\pi}_n^{\text{S}}$ from a *predefined* and *finite* collection of target policies $\{\pi_1, \ldots, \pi_m\}$, effectively seeking to determine $\operatorname{argmin}_{k \in [m]} R(\pi_k)$. Third, OPL aims to find a policy $\hat{\pi}_n^{\text{L}}$ within the *potentially infinite policy space* Π that achieves the lowest risk, essentially aiming to find $\operatorname{argmin}_{\pi \in \Pi} R(\pi)$. In general, both OPS and OPL rely on OPE's initial estimation of the risk.

Regularized IPS. Our work focuses on the inverse propensity scoring (IPS) estimator [24]. IPS approximates the risk of a policy π, $R(\pi)$, by adjusting the contribution of each sample in logged data according to its *importance weight (IW)*, which is the ratio of the probability of an action under the target policy π to its probability under the behavior policy π_0,

$$\hat{R}_n(\pi) = \frac{1}{n} \sum_{i=1}^{n} w_\pi(x_i, a_i) c_i \, , \tag{2}$$

where for any $(x, a) \in \mathcal{X} \times \mathcal{A}, w_\pi(x, a) = \pi(a|x)/\pi_0(a|x)$ are the IWs. IPS is unbiased under the coverage assumption (see for example Owen [39, Chapter 9]). However, it can suffer high variance, which tends to scale linearly with IWs [57]. This issue becomes pronounced when there is a significant discrepancy between the target policy π and the behavior policy π_0. To mitigate this, a common strategy consists in applying a regularization function $h : [0, 1]^2 \times [-1, 0] \to (-\infty, 0]$ to $\pi(a|x)$, $\pi_0(a|x)$ and c. This function is designed to reduce the estimator's variance at the cost of introducing some bias. Formally, the function h needs to satisfy the condition (**C1**), that is *defined* by

$$h \text{ satisfies (\textbf{C1})} \iff \forall (p, q, c) \in [0, 1]^2 \times [-1, 0], \quad pc/q \le h(p, q, c) \le 0. \tag{C1}$$

With such function h, the regularized IPS estimator reads

$$\hat{R}_n^h(\pi) = \frac{1}{n} \sum_{i=1}^{n} h\left(\pi(a_i|x_i), \pi_0(a_i|x_i), c_i\right) = \frac{1}{n} \sum_{i=1}^{n} h_i \, , \tag{3}$$

where $h_i = h\left(\pi(a_i|x_i), \pi_0(a_i|x_i), c_i\right)$. We recover standard IPS in (2) when $h(p, q, c) = pc/q$. Numerous regularization functions h were studied in the literature. For example,

$$h(p, q, c) = \min(p/q, M)c \, , M \in \mathbb{R}^+ \implies \text{Clipping [10]} \, , \tag{4}$$
$$h(p, q, c) = pc/q^\alpha \, , \alpha \in [0, 1] \implies \text{Exponential Smoothing [5]} \, ,$$
$$h(p, q, c) = pc/(q + \gamma) \, , \gamma \ge 0 \implies \text{Implicit Exploration [21]} \, .$$

Other IW regularizations include Harmonic [38] and Shrinkage [54]. With h satisfying (**C1**), we can derive our core result: a family of high-probability bounds that hold for regularized IPS.

3 Pessimistic off-policy evaluation

Standard OPE directly uses estimates of the risk, without capturing their associated uncertainty. This limits its effectiveness in critical applications. Pessimistic OPE addresses this issue by relying on finite sample, high-probability upper bounds to assess any policy's worst-case risk [15, 32]. This section contributes to this effort and focuses on providing novel, finite sample, tight upper bounds on the risk. This is achieved by deriving general bounds applicable to regularized IPS in (3).

3.1 Preliminaries and unified risk bounds

Let $\lambda > 0$, $\pi \in \Pi$, and h satisfying (**C1**), we define

$$\hat{\mathcal{M}}_n^{h,\ell}(\pi) = \frac{1}{n} \sum_{i=1}^{n} h_i^\ell, \qquad \text{and} \quad \psi_\lambda(x) = \frac{1}{\lambda}\left(1 - \exp(-\lambda x)\right), \, \forall x \in \mathbb{R}, \tag{5}$$

where $\hat{\mathcal{M}}_n^{h,\ell}(\pi)$ is the empirical ℓ-th moment of regularized IPS $\hat{R}_n^h(\pi)$, and $\psi_\lambda : \mathbb{R} \to \mathbb{R}$ is a *contraction* function satisfying $\psi_\lambda(x) \le x$ for any $x \in \mathbb{R}$. Then, we state our first result.

Proposition 1 (Empirical moments risk bound). *Let $\pi \in \Pi$, $L \ge 1$, $\delta \in (0, 1]$, $\lambda > 0$, and h satisfying (**C1**). Then it holds with probability at least $1 - \delta$ that*

$$R(\pi) \le U_L^{\lambda,h}(\pi), \quad \text{with} \quad U_L^{\lambda,h}(\pi) = \psi_\lambda\left(\hat{R}_n^h(\pi) + \sum_{\ell=2}^{2L} \frac{\lambda^{\ell-1}}{\ell} \hat{\mathcal{M}}_n^{h,\ell}(\pi) + \frac{\ln(1/\delta)}{\lambda n}\right), \tag{6}$$

where ψ_λ and $\hat{\mathcal{M}}_n^{h,\ell}(\pi)$ are both defined in (5), and recall that $\psi_\lambda(x) \le x$.

In Appendix F.1, we provide detailed proof, leveraging Chernoff bounds with a careful analysis of the moment-generating function. This results in the first empirical, high-order moment bound for offline contextual bandits, with several advantages. First, the bound applies to any regularization function h that satisfies the mild condition (C1), enabling the design of a tailored h that minimizes the bound. Second, it relies solely on empirical moments, without assuming the existence of theoretical moments. Third, the bound is fully empirical and tractable, facilitating efficient implementation of pessimism. Lastly, the parameter L controls the number of moments used, allowing a balance between bound tightness and computational cost. Specifically, for sufficiently small values of λ, higher values of L yield tighter bounds, though potentially at the cost of increased computational complexity as we would need to compute higher order moments. This is formally stated as follows.

Proposition 2 (Impact of L). *Let $\pi \in \Pi$, $\delta \in (0, 1]$, $\lambda > 0$, $L \geq 1$, and h satisfying* (C1). *Then,*

$$\lambda \leq \min_{i \in [n]} \left\{ \frac{2L + 2}{(2L + 1)|h_i|} \right\} \implies U_{L+1}^{\lambda, h}(\pi) \leq U_L^{\lambda, h}(\pi). \tag{7}$$

From (7), the bound $U_L^{\lambda, h}(\pi)$ in (6) becomes a decreasing function of L when $\lambda \leq \min_{i \in [n]}(1/|h_i|)$, suggesting that for sufficiently small λ, the tightest bound is achieved as $L \to \infty$. This condition on λ also depends on the values of h, highlighting the importance of the regularizer choice h. In fact, once we evaluate our bounds at their optimal regularizer function h, this condition on λ becomes unnecessary when comparing some of the optimal bounds. Specifically, we demonstrate in the following proposition that the bound with $L = 1$ can be always improved by increasing L.

Proposition 3 (Comparison of our bounds). *Let $\pi \in \Pi$, and $\lambda > 0$, we define*

$$U_L^{\lambda}(\pi) = \min_h U_L^{\lambda, h}(\pi), \quad and \quad h_{*, L} = \operatorname*{argmin}_h U_L^{\lambda, h}(\pi), \tag{8}$$

with the minimum taken over h satisfying (C1). *Then, for any $\lambda > 0$, it holds that for any $L > 1$, $U_L^{\lambda}(\pi) \leq U_1^{\lambda}(\pi)$. In particular, for any $\lambda > 0$,*

$$U_\infty^{\lambda}(\pi) \leq U_1^{\lambda}(\pi). \tag{9}$$

Proposition 3 shows that, irrespective of the value of λ, the bound with $L = 1$ can be always improved by bounds of increased moment order L, evaluated at their optimal regularizer $h_{*, L}$. This result encourages us to study bounds with high moment order L, especially if we can derive their optimal regularizers $h_{*, L}$. To this end, we examine two cases: $L = 1$, which results in an empirical second-moment bound, and $L \to \infty$, yielding a tight bound that does not require computing high-order moments. For each case, we identify the function h that minimizes the bound. If the minimizer for $L = 1$ is a variant of the clipping estimator [10], minimizing $L \to \infty$ motivates a novel logarithmic smoothing estimator. We begin by analyzing our empirical moment risk bound at $L = 1$.

3.2 Global clipping

Corollary 4 (Empirical second-moment risk bound with $L = 1$). *Let $\pi \in \Pi$, $\delta \in (0, 1]$, $\lambda > 0$, and h satisfying* (C1). *Then it holds with probability at least $1 - \delta$ that*

$$R(\pi) \leq \psi_\lambda \left(\hat{R}_n^h(\pi) + \frac{\lambda}{2} \hat{\mathcal{M}}_n^{h, 2}(\pi) + \frac{\ln(1/\delta)}{\lambda n} \right). \tag{10}$$

This is a direct consequence of (6) when $L = 1$. The bound holds for any h satisfying (C1). Thus we search for a function $h_{*, 1}$ that minimizes bound in (10). This function $h_{*, 1}$ writes

$$h_{*, 1}(p, q, c) = -\min(p|c|/q, 1/\lambda). \tag{11}$$

In particular, if we assume that costs are binary, $c \in \{-1, 0\}$, then $h_{*, 1}$ corresponds to clipping in (4) with parameter $M = 1/\lambda$. This is because $-\min(|c|p/q, 1/\lambda) = \min\left(p/q, \frac{1}{\lambda}\right)c$ when c is binary. This motivates the widely used clipping estimator [10]. However, this also suggests that the standard way of clipping (as in (4)) is only optimal[1] for binary costs. In general, the cost should also be clipped (as in (11)). Finally, with a suitable choice of $\lambda = \mathcal{O}(1/\sqrt{n})$, our bound in Corollary 4, using clipping (i.e., $h = h_{*, 1}$), outperforms the existing empirical Bernstein bound [55], which was specifically derived for clipping. This confirms the strength of our general bound, as minimizing it results in a bound with tighter concentration than specialized bounds. Appendix F.4 gives the the proof to find $h_{*, 1}$ and formal comparisons with empirical Bernstein are provided in Appendix F.5. In the next section, we study our general bound when we set $L \to \infty$.

---

[1] Here, optimality of a function h is defined with respect to our bound with $L = 1$ (Corollary 4).

3.3 Logarithmic smoothing

Corollary 5 (Empirical infinite-moment bound with $L \to \infty$). *Let $\pi \in \Pi$, $\delta \in (0,1]$, $\lambda > 0$, and h satisfying* (**C1**). *Then it holds with probability at least $1 - \delta$ that*

$$R(\pi) \leq \psi_\lambda \left(-\frac{1}{n} \sum_{i=1}^{n} \frac{1}{\lambda} \log\left(1 - \lambda h_i\right) + \frac{\ln(1/\delta)}{\lambda n} \right). \tag{12}$$

Appendix F.6 provides detailed proof. Setting $L \to \infty$ in (6) results in the bound in Corollary 5, which has different properties than Corollary 4. The resulting bound has a simple expression that does not require computing high order moments. This means that we can obtain the best of both worlds, a tight concentration bound with no additional computational complexity. As the bound is increasing in h, the function $h_{*,\infty}$ that minimizes this bound is $h_{*,\infty}(p, q, c) = pc/q$. This corresponds to the standard IPS in (2). This differs from the $L = 1$ bound in Corollary 4 that favored clipping. This shows the impact of the moment order L on the optimal function h. For any $\pi \in \Pi$, applying the bound in Corollary 5 with the optimal $h_{*,\infty}$ leads to $U_\infty^\lambda(\pi)$, of the following expression:

$$U_\infty^\lambda(\pi) = \psi_\lambda \left(\hat{R}_n^\lambda(\pi) + \frac{\ln(1/\delta)}{\lambda n} \right). \tag{13}$$

Even if we set $h_{*,\infty}(p, q, c) = pc/q$ (without IW regularization), $U_\infty^\lambda(\pi)$ can be seen as a risk upper bound of a novel regularized IPS estimator (satisfying (**C1**)), called Logarithmic Smoothing (LS):

$$\hat{R}_n^\lambda(\pi) = -\frac{1}{n} \sum_{i=1}^{n} \frac{1}{\lambda} \log\left(1 - \lambda w_\pi(x_i, a_i) c_i\right). \tag{14}$$

The LS estimator in (14) is defined for any non-negative $\lambda \geq 0$, with its bound in (13) holding for any positive $\lambda > 0$. Notably, $\lambda = 0$ retrieves the standard IPS estimator in (2), while $\lambda > 0$ introduces a bias-variance trade-off by logarithmically smoothing the IWs (Figure 1). This estimator acts as a soft, differentiable variant of clipping with parameter $1/\lambda$. A Taylor expansion of our estimator around $\lambda = 0$ yields

$$\hat{R}_n^\lambda(\pi) = \hat{R}_n(\pi) + \sum_{\ell=2}^{\infty} \frac{\lambda^{\ell-1}}{\ell} \left(\frac{1}{n} \sum_{i=1}^{n} \left(w_\pi(x_i, a_i) c_i\right)^\ell \right).$$

Figure 1: LS with different λs.

Thus, LS is a pessimistic estimator *by design*, implicitly implementing a form of *Sample All Moments Penalization*, which generalizes the *Sample Variance Penalization* [55]. To examine the statistical properties of our estimator, we introduce

$$\mathcal{S}_\lambda(\pi) = \mathbb{E}\left[\frac{(w_\pi(x, a) c)^2}{(1 - \lambda w_\pi(x, a) c)} \right], \tag{15}$$

which quantifies the discrepancy between π and π_0. Notably, $\mathcal{S}_\lambda(\pi)$ is always smaller than the second moment of the IW, effectively interpolating between a weighted first moment ($\lambda \gg 1$) and the second moment ($\lambda = 0$) of IPS. This quantity $\mathcal{S}_\lambda$ characterizes the concentration properties of the LS estimator akin to the coverage ratio for IX estimator [21]. With $\mathcal{S}_\lambda$ defined, we proceed by bounding the mean squared error (MSE) of our estimator, specifically bounding its bias and variance.

Proposition 6 (Bias-variance trade-off). *Let $\pi \in \Pi$ and $\lambda \geq 0$. Let $\mathcal{B}^\lambda(\pi)$ and $\mathcal{V}^\lambda(\pi)$ be respectively the bias and the variance of the LS estimator. Then we have that*

$$0 \leq \mathcal{B}^\lambda(\pi) \leq \lambda \mathcal{S}_\lambda(\pi), \quad \text{and} \quad \mathcal{V}^\lambda(\pi) \leq \frac{\mathcal{S}_\lambda(\pi)}{n}.$$

Moreover, it holds that for any $\lambda > 0$, the variance is finite as $\mathcal{V}^\lambda(\pi) \leq |R(\pi)|/\lambda n \leq 1/\lambda n$.

We observe that both the bias and variance are controlled by $\mathcal{S}_\lambda(\pi)$. Particularly, $\lambda = 0$ recovers the IPS estimator in (2), with zero bias and a variance bounded by $\mathbb{E}\left[w^2(x, a) c^2\right]/n$. When $\lambda > 0$, a bias-variance trade-off emerges. The bias is always non-negative and is capped at $\lambda \mathcal{S}_\lambda(\pi)$, which diminishes to zero when λ is small and goes to $|R(\pi)|$ as λ increases. Conversely, the variance decreases with a higher λ. Notably, $\lambda > 0$ ensures finite variance bounded by $1/\lambda n$, despite the estimator being unbounded. This is different from previous estimators that relied on bounded functions to ensure finite variance. We also prove in the following that a good choice of $\lambda = \mathcal{O}(1/\sqrt{n})$ ensures that our LS estimator enjoys a sub-Gaussian concentration [38].
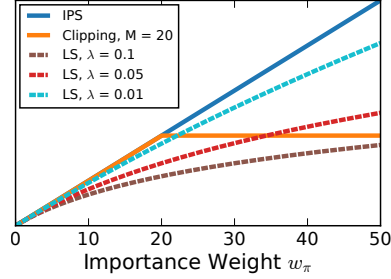
Proposition 7 (Sub-Gaussianity and comparison with Metelli et al. [38]). *Let* $\pi \in \Pi$, $\delta \in (0,1]$ *and* $\lambda > 0$. *Then the following inequalities holds with probability at least* $1 - \delta$:

$$R(\pi) - \hat{R}_n^\lambda(\pi) \leq \frac{\ln(2/\delta)}{\lambda n}, \qquad and \qquad \hat{R}_n^\lambda(\pi) - R(\pi) \leq \lambda \mathcal{S}_\lambda(\pi) + \frac{\ln(2/\delta)}{\lambda n}.$$

In particular, setting $\lambda = \lambda_* = \sqrt{\ln(2/\delta)/n\mathbb{E}\left[w_\pi(x,a)^2 c^2\right]}$ *yields that*

$$|R(\pi) - \hat{R}_n^{\lambda_*}(\pi)| \leq \sqrt{2\sigma^2 \ln(2/\delta)}, \qquad where \ \sigma^2 = 2\mathbb{E}\left[w_\pi(x,a)^2 c^2\right]/n. \tag{16}$$

Thus, a particular choice of λ^* ensures that $\hat{R}_n^{\lambda^*}(\pi)$ is sub-Gaussian, with a variance proxy σ^2 that improves on that obtained for the Harmonic estimator of Metelli et al. [38]. We refer the interested reader to Appendix E.2 for further discussions and proofs.

Next, we focus on the tightness of the LS upper bound in (13) as it will motivate our selection and learning strategies. Proposition 3 already showed that $U_\infty^\lambda(\pi)$, the bound of LS is tighter than $U_1^\lambda(\pi)$, the bound in Corollary 4 evaluated at the Global clipping function $h_{*,1}$. In this section, we compare the LS bound to the already tight IX bound presented by Gabbianelli et al. [21] and demonstrate in the following that the LS bound dominates it in all scenarios.

Proposition 8 (Comparison with IX of Gabbianelli et al. [21]). *Let* $\pi \in \Pi$, $\delta \in]0,1]$ *and* $\lambda > 0$, *the* IX *bound from [21] states that we have with probability at least* $1 - \delta$

$$R(\pi) \leq \hat{R}_n^{\lambda\text{-IX}}(\pi) + \frac{\ln(1/\delta)}{\lambda n}, \quad with \quad \hat{R}_n^{\lambda\text{-IX}}(\pi) = \frac{1}{n}\sum_{i=1}^n \frac{\pi(a_i|x_i)}{\pi_0(a_i|x_i) + \lambda/2} c_i. \tag{17}$$

Let $U_{\text{IX}}^\lambda(\pi)$ *be the upper bound of* (17)*, we have for any* $\lambda > 0$:

$$U_\infty^\lambda(\pi) \leq U_{\text{IX}}^\lambda(\pi). \tag{18}$$

This result states that no matter the scenario, for any evaluated policy π, and any chosen $\lambda > 0$, the LS bound will be always tighter than IX. The gap between the LS and IX bounds increases when n is small, or when the evaluated policy π is stochastic, as demonstrated and developed in Appendix F.8. These findings further validate the effectiveness of our approach, enabling us to identify the LS estimator, with an empirical bound that improves upon the tightest existing bounds. Consequently, we leverage the LS bound in the next section to derive our pessimistic OPS and OPL strategies.

4 Off-policy selection and learning

4.1 Off-policy selection

Let $\Pi_{\text{s}} = \{\pi_1, ..., \pi_m\}$ be a finite set of policies. In OPS, the goal is to find $\pi_*^{\text{s}} \in \Pi_{\text{s}}$ that satisfies

$$\pi_*^{\text{s}} = \operatorname*{argmin}_{\pi \in \Pi_{\text{s}}} R(\pi) = \operatorname*{argmin}_{k \in [m]} R(\pi_k). \tag{19}$$

As we do not have access to the true risk, we use a data-driven selection strategy that guarantees the identification of policies of performance close to that of π_*^{s}. Precisely, for $\lambda > 0$, we search for

$$\hat{\pi}_n^{\text{s}} = \operatorname*{argmin}_{\pi \in \Pi_{\text{s}}} \hat{R}_n^\lambda(\pi) = \operatorname*{argmin}_{k \in [m]} \hat{R}_n^\lambda(\pi_k). \tag{20}$$

To derive our strategy in (20), we minimize the bound of LS in (13), employing pessimism [27]. Fortunately, in our case, this boils down to minimizing $\hat{R}_n^\lambda(\pi)$, since the other terms in the bound are independent of the target policy π. This allows us to avoid computing complex statistics [55, 32] and does not require access to the behavior policy π_0. As we show next, it also ensures low suboptimality.

Proposition 9 (Suboptimality of our selection strategy in (20)). *Let* $\lambda > 0$ *and* $\delta \in (0,1]$. *Then, it holds with probability at least* $1 - \delta$ *that*

$$0 \leq R(\hat{\pi}_n^{\text{s}}) - R(\pi_*^{\text{s}}) \leq \lambda \mathcal{S}_\lambda(\pi_*^{\text{s}}) + \frac{2\ln(2|\Pi_{\text{s}}|/\delta)}{\lambda n}, \tag{21}$$

where $\mathcal{S}_\lambda(\pi)$, π_*^{s} *and* $\hat{\pi}_n^{\text{s}}$ *are defined in* (15)*,* (19) *and* (20)*.*

The derived suboptimality bound only requires coverage of the optimal actions (support of the optimal policy π_*^s), and improves on IX suboptimality [21], matching the minimax suboptimality lower bound of pessimistic methods [34, 27, 28]. Appendix G.1 provides proof of this suboptimality bound, and we discuss how this suboptimality improves upon existing strategies in Appendix E.3. By selecting $\lambda_n^s = \sqrt{2\ln(2|\Pi_S|/\delta)/n}$ for LS, we achieve a suboptimality scaling of $\mathcal{O}(1/\sqrt{n})$,

$$0 \leq R(\hat{\pi}_n^S) - R(\pi_*^S) \leq \left(1 + \mathcal{S}_{\lambda_n^s}(\pi_*^S)\right)\sqrt{2\ln(2|\Pi_S|/\delta)/n}, \tag{22}$$

which ensures finding the optimal policy with sufficient samples. Additionally, the multiplicative constant is smaller when π_0 is close to π_*^S, confirming the known observation that it is easier to identify the best policy if it is similar to the behavior policy π_0.

4.2 Off-policy learning

Similar to how we extended the evaluation bound in Corollary 5 (which applies to a single fixed target policy) to OPS (where it applies to a finite set of target policies), we can further derive bounds for an infinite policy class Π, enabling OPL. Several approaches have been proposed in previous work, primarily based on replacing the finite union bound over policies with more sophisticated uniform-convergence arguments. This was used by [55], which derived a variance-sensitive bound scaling with the covering number [61]. Since these approaches incorporate a complexity term that depends only on the policy class, the resulting pessimistic learning strategy (which minimizes the upper bound) would be similar to the selection strategy adopted earlier, leading, for a fixed λ, to

$$\hat{\pi}_n^L = \underset{\pi \in \Pi}{\operatorname{argmin}} \, \hat{R}_n^\lambda(\pi) + \frac{\mathcal{C}(\Pi)}{\lambda n} = \underset{\pi \in \Pi}{\operatorname{argmin}} \, \hat{R}_n^\lambda(\pi). \tag{23}$$

where $\mathcal{C}(\Pi)$ is a complexity measure [61]. This learning strategy is straightforward because it involves a smooth estimator that can be optimized using first-order methods and does not require second-order statistics. However, analyzing this approach is more challenging because the complexity measure $\mathcal{C}(\Pi)$ varies depending on the policy class considered, is often intractable [49] and can only be upper bounded with problem dependent constants [28].

Instead of the method described above, we derive PAC-Bayesian generalization bounds [37, 11] that apply to arbitrary policy classes. This framework has been shown to provide strong performance guarantees for OPL in practical scenarios [49, 5]. The PAC-Bayesian framework analyzes the performance of policies by viewing them as randomized predictors [35]. Specifically, let $\mathcal{F}(\Theta) = \{f_\theta : \mathcal{X} \to [K], \theta \in \Theta\}$ be a set of parameterized predictors that associate the context x with the action $f_\theta(x) \in [K]$. Let $\mathcal{P}(\Theta)$ be the set of all probability distributions on Θ. Each distribution $Q \in \mathcal{P}(\Theta)$ defines a policy π_Q by setting the probability of action a given context x as the probability that a random predictor $f_\theta \sim Q$ maps x to action a, that is,

$$\pi_Q(a|x) = \mathbb{E}_{\theta \sim Q}\left[\mathbb{1}\left[f_\theta(x) = a\right]\right], \qquad \forall(x,a) \in \mathcal{X} \times \mathcal{A}. \tag{24}$$

This characterization is not restrictive as any policy can be represented in this form [49]. Deriving PAC-Bayesian generalization bounds with this policy definition requires the regularized IPS to be linear in the target policy π [35, 5, 21]. Our estimator LS in (14) is non-linear in π. Therefore, for this PAC-Bayesian analysis, we introduce a linearized variant of LS, called LS-LIN, and defined as

$$\hat{R}_n^{\lambda\text{-LIN}}(\pi) = -\frac{1}{n}\sum_{i=1}^n \frac{\pi(a_i|x_i)}{\lambda}\log\left(1 - \frac{\lambda c_i}{\pi_0(a_i|x_i)}\right), \tag{25}$$

which smooths the impact of the behavior propensity π_0 instead of the IWs π/π_0. We provide in the following a core result of this section, the PAC-Bayesian bound that defines our learning strategy.

Proposition 10 (PAC-Bayes learning bound for $\hat{R}_n^{\lambda\text{-LIN}}$). *Given a prior $P \in \mathcal{P}(\Theta)$, $\delta \in (0, 1]$ and $\lambda > 0$, the following holds with probability at least $1 - \delta$:*

$$\forall Q \in \mathcal{P}(\Theta), \quad R(\pi_Q) \leq \psi_\lambda\left(\hat{R}_n^{\lambda\text{-LIN}}(\pi_Q) + \frac{\mathcal{KL}(Q||P) + \ln\frac{1}{\delta}}{\lambda n}\right), \tag{26}$$

where $\mathcal{KL}(Q||P)$ is the Kullback-Leibler divergence from P to Q.

PAC-Bayes bounds hold uniformly for all distributions $Q \in \mathcal{P}(\Theta)$ and replace the complexity measure $\mathcal{C}(\Pi)$ with the divergence $\mathcal{KL}(Q||P)$ from a reference *prior* distribution P. Extensive research focuses on identifying the best strategies for choosing this prior P [40]. While these bounds hold for any fixed prior P, in practice, it is typically set to the distribution inducing the behavior policy π_0, meaning P satisfies $\pi_0 = \pi_P$. This leads to an intuitive learning principle: by minimizing the upper bound, we seek policies with good empirical risk that do not deviate significantly from π_0.

Our bound can also be obtained using the truncation method from Alquier [1, Corollary 2.5]. This bound surpasses the already tight PAC-Bayesian bounds derived for Clipping [49], Exponential Smoothing [5], and Implicit Exploration [21], resulting in the tightest known generalization bound in OPL. Appendix G.2 gives formal proof of this bound and comparisons with existing PAC-Bayesian bounds can be found in Appendix E.4. For a fixed λ and a fixed prior P, we derive a learning strategy that minimizes the upper bound for a subset $\mathcal{L}(\Theta) \subseteq \mathcal{P}(\Theta)$ of distributions, seeking

$$Q_n = \underset{Q \in \mathcal{L}(\Theta)}{\operatorname{argmin}} \left\{ \hat{R}_n^{\lambda\text{-LIN}}(\pi_Q) + \frac{\mathcal{KL}(Q||P)}{\lambda n} \right\}, \quad \text{and setting } \hat{\pi}_n^{\mathtt{L}} = \pi_{Q_n}. \tag{27}$$

(27) is tractable and can be efficiently optimized for various policy classes [49, 5]. Below, we analyze its suboptimality compared to the best policy in the chosen class, $\pi_{Q^*} = \operatorname{argmin}_{Q \in \mathcal{L}(\Theta)} R(\pi_Q)$.

Proposition 11 (Suboptimality of the learning strategy in (27)). *Let $\lambda > 0$, $P \in \mathcal{L}(\Theta)$ and $\delta \in (0, 1]$. Then, it holds with probability at least $1 - \delta$ that*

$$0 \le R(\hat{\pi}_n^{\mathtt{L}}) - R(\pi_{Q^*}) \le \lambda \mathcal{S}_\lambda^{\text{LIN}}(\pi_{Q^*}) + \frac{2\left(\mathcal{KL}(Q^*||P) + \ln(2/\delta)\right)}{\lambda n}, \tag{28}$$

where $\mathcal{S}_\lambda^{\text{LIN}}(\pi) = \mathbb{E}\left[\pi(a|x)c^2/(\pi_0^2(a|x) - \lambda\pi_0(a|x)c)\right]$ and $\hat{\pi}_n^{\mathtt{L}}$ is defined in (27).

Our suboptimality bound only requires coverage of the support of the optimal policy π_{Q_*}. This bound matches the minimax suboptimality lower bound of pessimistic learning with deterministic policies [28]. Appendix G.3 provides a proof of Proposition 11, while Appendix E.5 discusses the suboptimality bound further and proves that it improves on the IX learning strategy of [21, Section 5]. Setting $\lambda_n^l = 2/\sqrt{n}$ guarantees us a suboptimality that scales with $\mathcal{O}(1/\sqrt{n})$ as

$$0 \le R(\hat{\pi}_n^{\mathtt{L}}) - R(\pi_{Q^*}) \le (2\mathcal{S}_{\lambda_n^l}^{\text{LIN}}(\pi_{Q^*}) + \mathcal{KL}(Q^*||P) + \ln(2/\delta))/\sqrt{n}.$$

By setting the reference P to the distribution inducing π_0, we find that the learning suboptimality is reduced when the behavior policy π_0 is close to the optimal policy π_{Q^*}. This is similar to the suboptimality for our selection strategy. The suboptimality upper bound reflects a common intuition in the OPL literature: pessimistic learning algorithms converge faster when π_0 is close to π_{Q^*}.

5 Experiments

Our experimental setup follows the standard multiclass-to-bandit conversion used in prior studies [18, 55]. Each multi-class dataset has features and labels and we convert it to contextual bandit problems where contexts correspond to features and actions to labels. Precisely, the reward r for taking action (label) a with context (features) x is modeled as Bernoulli with probability $p_x = \epsilon + \mathbb{1}\left[a = \rho(x)\right](1 - 2\epsilon)$, where $\rho(x)$ be the true label of features x, and ϵ is a noise parameter. In particular, the true label $\rho(x)$ represents the action with the highest average reward for context x. This setup ensures an average reward of $1 - \epsilon$ for the optimal action $\rho(x)$ and ϵ for all others, constructing a logged bandit feedback dataset in the form $\{x_i, a_i, c_i\}_{i \in [n]}$, where $c_i = -r_i$ is the associated cost.

5.1 Off-policy evaluation and selection experiments

For both evaluation and selection, we adopt the same experimental design as [32] to facilitate the comparison. We consider exponential target policies $\pi(a|x) \propto \exp(\frac{1}{\tau} f(a, x))$, with τ a temperature controlling the policy's entropy and $f(a, x)$ the score of the item a for the context x. We use this to define ideal policies as $\pi^{\mathtt{ideal}}(a|x) \propto \exp(\frac{1}{\tau}\mathbb{I}\{\rho(x) = a\})$, and also create faulty, mismatching policies for which the peak is shifted to another, wrong action for a set of faulty actions $F \subset [K]$. To recreate real world scenarios, we also consider policies directly learned from logged bandit feedback, of the form $\pi_{\theta^{\mathtt{IPS}}}(a|x) \propto \exp(\frac{1}{\tau} x^t \theta_a^{\mathtt{IPS}})$ and $\pi_{\theta^{\mathtt{SN}}}(a|x) \propto \exp(\frac{1}{\tau} x^t \theta_a^{\mathtt{SN}})$, with their parameters learned

Table 1: Bound's tightness ($|U(\pi)/R(\pi) - 1|$) with varying number of samples of the `kropt` dataset.

Number of samples	SN-ES	cIPS-EB	IX	cIPS-L=1 (Ours)	LS (Ours)
2^8	1.000	0.917	0.373	<u>0.364</u>	**0.362**
2^9	1.000	0.732	<u>0.257</u>	0.289	**0.236**
2^{10}	0.794	0.554	<u>0.226</u>	0.240	**0.213**
2^{11}	0.649	0.441	<u>0.171</u>	0.197	**0.159**
2^{12}	0.472	0.327	<u>0.126</u>	0.147	**0.117**
2^{13}	0.374	0.204	<u>0.062</u>	0.077	**0.054**
2^{14}	0.257	0.138	<u>0.041</u>	0.049	**0.035**

by respectively minimizing the IPS [24] and SN [56] empirical risks. More details on the definition of the different policies are given in Appendix H. Finally, 11 real multiclass classification datasets are chosen from the UCI ML Repository [8] (See Table 3 in Appendix H.1.1) with various number of samples, dimensions and action space sizes to conduct our experiments[2].

(OPE) Tightness of the bounds. Evaluating the worst case performance of a policy is done through evaluating risk upper bounds [10, 32]. This means that a better evaluation will solely depend on the tightness of the bounds used. To this end, given a policy π, we are interested in bounds $U(\pi)$ with a small relative radius $|U(\pi)/R(\pi) - 1|$. We compare our newly derived bounds (cIPS-L=1 for U_1^λ and LS for U_∞^λ both with $\lambda = 1/\sqrt{n}$) to empirical evaluation bounds of the literature: SN-ES: the Efron Stein bound for Self Normalized IPS [32], cIPS-EB: Empirical Bernstein for Clipping [55] and the recent IX: Implicit Exploration bound [21]. The first experiment uses the `kropt` dataset with $\epsilon = 0.2$, collects bandit feedback with faulty behavior policy (with $\tau = 0.25$) to evaluate an ideal policy ($\tau = 0.1$), and explores how the relative radiuses of the considered bounds shrink while varying the number of datapoints. Table 1 compiles the results of the experiments and suggest that the LS bound is tighter than its competitors no matter the size of the feedback collected. The second experiments uses all 11 datasets, with different behavior policies ($\tau_0 \in \{0.2, 0.25, 0.3\}$) and different noise levels ($\epsilon \in \{0., 0.1, 0.2\}$) to evaluate ideal policies with different temperatures ($\tau \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$), defining ~ 500 different scenarios to validate our findings. We plot in Figure 2 the cumulative distribution of the relative radius of the considered bounds. We observe that while cIPS-L=1 and IX can be comparable, the LS bound is tighter than all its competitors. We also provide detailed results in Appendix H.1.2 that further confirm the superiority of the LS bound.

(OPS) Find the best, avoid the worst policy. Policy selection aims at identifying the best policy among a set of finite candidates. In practice, we are interested in finding policies that improve on π_0 and avoid policies that perform worse than π_0. To replicate real world scenarios, we design an experiment where π_0 is a faulty policy ($\tau = 0.2$), that collects noisy ($\epsilon = 0.2$) interaction data, some of which is used to learn $\pi_{\theta^{IPS}}, \pi_{\theta^{SN}}$, and that we add to our discrete set of policies $\Pi_{k=4} = \{\pi_0, \pi^{\text{ideal}}, \pi_{\theta^{IPS}}, \pi_{\theta^{SN}}\}$. The goal is to measure the ability of our selection strategies to choose from $\Pi_{k=4}$, better performing policies than π_0. We thus define three possible outcomes: a strategy can select *worse* performing policies, *better* performing or the *best* policy. Our goal in these experiments is to empirically validate the pitfalls of point estimators while confirming the benefits of using the pessimism principle. To this end, we compare *pessimistic* selection strategies to policy selection using the classical point estimators IPS [24] and SN [56]. The comparison is conducted on the 11 UCI datasets with 10 different seeds resulting in 110 scenarios. We plot in Figure 2 the percentage of time each method selected the best policy, a better or a worse policy than π_0. While risk estimators can identify the best policy, they are unreliable as they can choose worse performing policies than π_0, a catastrophic outcome in critical applications. Pessimistic selection is more conservative, as it avoids poor performing policies completely and empirically confirms that tighter upper bounds result in better selection strategies: LS upper bound is less conservative and finds best policies the most (comparable to SN) while never selecting poor performing policies. Fine grained results (for each dataset) can be found in Appendix H.1.3.

5.2 Off-policy learning experiments

We follow the successful off policy learning paradigm based on directly minimizing PAC-Bayesian risk generalization bounds [49, 5] as it comes with guarantees of improvement and avoids hyper-

---

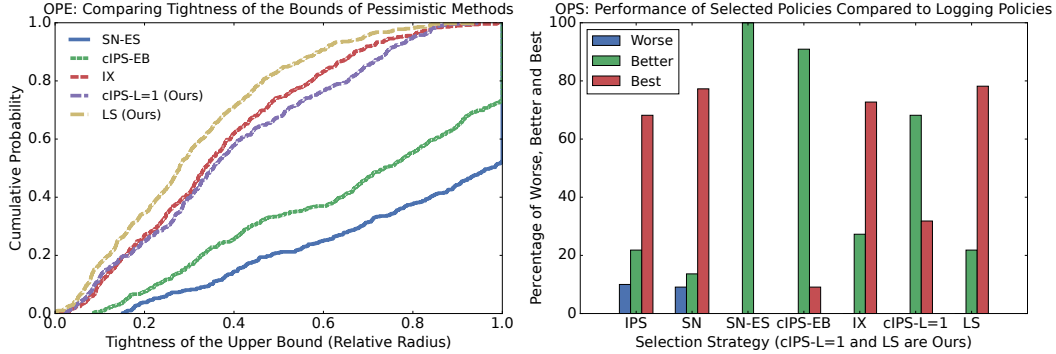[2]The code can be found at https://github.com/otmhi/offpolicy_ls.

Figure 2: Results for OPE and OPS experiments.

	cIPS	cvcIPS	ES	IX	LS-LIN (Ours)
$rI(U(\hat{\pi}_n^L))$	14.48%	21.28%	7.78%	<u>24.74%</u>	**26.31%**
$rI(R(\hat{\pi}_n^L))$	28.13%	<u>33.64%</u>	29.44%	**36.70%**	**36.76%**

Table 2: OPL: Relative Improvement of guaranteed risk and true risk averaged over 200 scenarios.

parameter tuning. For comparable results, we use the same 4 datasets (described in Appendix H.2, Table 7) as in [49, 5] and adopt the **LGP**: Linear Gaussian Policies [49] as our class of parametrized policies. For each dataset, we use behavior policies trained on a small fraction of the data in a supervised fashion, combined with different inverse temperature parameters $\alpha \in \{0.1, 0.3, 0.5, 0.7, 1.\}$ to cover cases of diffused and peaked behavior policies. These policies generate for 10 different seeds, 10 logged bandit feedback datasets resulting in 200 different scenarios to test our learning approaches. In the PAC-Bayesian OPL paradigm, we minimize the empirical upper bounds $U(\pi)$ directly and obtain the learned policy as the bound's minimizer $\hat{\pi}_n^L$ (as in (27)). With $\hat{\pi}_n^L$ obtained, we are interested in two quantities: The guaranteed risk by the bound, which is the value of the bound $U(\hat{\pi}_n^L)$ at its minimizer. This quantity reflects the worst case performance of the learned policy, a lower value implies stronger performance guarantees. We are also interested in the true risk of the minimizer of the bound $R(\hat{\pi}_n^L)$ as it translates the performance of the obtained policy acting on unseen data. As this learning paradigm is based on optimizing tractable, generalization bounds, we only compare our approach to methods that provide them. Precisely, we compare our LS-LIN learning strategy in (27) to strategies based on minimizing off-policy PAC Bayesian bounds from the literature: clipped IPS (cIPS) and Control Variate clipped IPS (cvcIPS) [49], Exponential Smoothing (ES) [5] and Implicit Exploration (IX) [21]. The results are summarized in Table 2 where we compute:

$$rI(x) = (R(\pi_0) - x)/(R(\pi_0) - R(\pi^*)) = (R(\pi_0) - x)/(R(\pi_0) + 1),$$

the improvement over $R(\pi_0)$ achieved by minimizing the different bounds in terms of $x \in \{U, R\}$ (guaranteed risk and true risk respectively), relative to an ideal improvement. This metric helps us normalize the results, and we report its average over 200 different scenarios, with results in bold being significantly better. Fine grained results can be found in Appendix H.2.4. We observe that the LS-LIN PAC-Bayesian bound improves substantially on its competitors in terms of the guaranteed risk, and also obtains the best performing policies (on par with the IX PAC-Bayesian bound).

6 Conclusion

Motivated by the *pessimism* principle, we have derived novel, empirical risk upper bounds tailored for the regularized IPS family of estimators. Minimizing these bounds within this family unveiled Logarithmic Smoothing, a simple estimator with good concentration properties. With its tight upper bound, LS confidently evaluates a policy, and shows provably better guarantees for both selecting and learning policies than all competitors. Our upper bounds remain broadly applicable, only requiring *negative costs*. While this condition does not impact importance weighting estimators, it does not hold for doubly robust estimators. Extending our approach to derive empirical bounds for this type of estimators presents a nontrivial, yet interesting task to explore in future work. Another potential extension would be to relax the i.i.d. assumption of the contextual bandit problem to address, the general offline Reinforcement Learning setting. This direction will introduce a more challenging estimation task and requires developing new concentration bounds.

References

[1] Pierre Alquier. *Transductive and Inductive Adaptative Inference for Regression and Density Estimation*. Theses, ENSAE ParisTech, December 2006. URL https://pastel.hal.science/tel-00119593.

[2] Pierre Alquier. User-friendly introduction to PAC-Bayes bounds. *Foundations and Trends® in Machine Learning*, 17(2), 2024.

[3] Imad Aouali, Amine Benhalloum, Martin Bompaire, Achraf Ait Sidi Hammou, Sergey Ivanov, Benjamin Heymann, David Rohde, Otmane Sakhi, Flavian Vasile, and Maxime Vono. Reward Optimizing Recommendation using Deep Learning and Fast Maximum Inner Product Search. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 4772–4773, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393850. doi: 10.1145/3534678.3542622. URL https://doi.org/10.1145/3534678.3542622.

[4] Imad Aouali, Achraf Ait Sidi Hammou, Sergey Ivanov, Otmane Sakhi, David Rohde, and Flavian Vasile. Probabilistic Rank and Reward: A Scalable Model for Slate Recommendation, 2022.

[5] Imad Aouali, Victor-Emmanuel Brunel, David Rohde, and Anna Korba. Exponential Smoothing for Off-Policy Learning. In *Proceedings of the 40th International Conference on Machine Learning*, pages 984–1017. PMLR, 2023.

[6] Imad Aouali, Victor-Emmanuel Brunel, David Rohde, and Anna Korba. Bayesian off-policy evaluation and learning for large action spaces. *arXiv preprint arXiv:2402.14664*, 2024.

[7] Imad Aouali, Victor-Emmanuel Brunel, David Rohde, and Anna Korba. Unified PAC-Bayesian Study of Pessimism for Offline Policy Learning with Regularized Importance Sampling. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024. URL https://openreview.net/forum?id=d7W4HOsTXU.

[8] A. Asuncion and D. J. Newman. UCI machine learning repository, 2007. URL http://www.ics.uci.edu/$\sim$mlearn/{MLR}epository.html.

[9] Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

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

[11] Olivier Catoni. PAC-Bayesian supervised classification: The thermodynamics of statistical learning. *IMS Lecture Notes Monograph Series*, page 1–163, 2007. ISSN 0749-2170. doi: 10.1214/074921707000000391. URL http://dx.doi.org/10.1214/074921707000000391.

[12] Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H. Chi. Top-K Off-Policy Correction for a REINFORCE Recommender System. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, page 456–464, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359405. doi: 10.1145/3289600.3290999. URL https://doi.org/10.1145/3289600.3290999.

[13] Victor Chernozhukov, Mert Demirer, Greg Lewis, and Vasilis Syrgkanis. Semi-parametric efficient policy learning with continuous actions. *Advances in Neural Information Processing Systems*, 32, 2019.

[14] Matej Cief, Jacek Golebiowski, Philipp Schmidt, Ziawasch Abedjan, and Artur Bekasov. Learning action embeddings for off-policy evaluation. In *European Conference on Information Retrieval*, pages 108–122. Springer, 2024.

[15] Bo Dai, Ofir Nachum, Yinlam Chow, Lihong Li, Csaba Szepesvari, and Dale Schuurmans. Coindice: Off-policy confidence interval estimation. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9398–9411. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6aaba9a124857622930ca4e50f5afed2-Paper.pdf.

[16] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 1097–1104, 2011.

[17] Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. Sample-efficient nonstationary policy evaluation for contextual bandits. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, UAI'12, page 247–254, Arlington, Virginia, USA, 2012. AUAI Press.

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

[19] Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, pages 1447–1456. PMLR, 2018.

[20] Hamish Flynn, David Reeb, Melih Kandemir, and Jan Peters. PAC-Bayes Bounds for Bandit Problems: A Survey and Experimental Comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):15308–15327, 2023. doi: 10.1109/TPAMI.2023.3305381.

[21] Germano Gabbianelli, Gergely Neu, and Matteo Papini. Importance-weighted offline learning done right. In *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.

[22] Alexandre Gilotte, Clément Calauzènes, Thomas Nedelec, Alexandre Abraham, and Simon Dollé. Offline A/B testing for recommender systems. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 198–206, 2018.

[23] Torben Hagerup and Christine Rüb. A Guided Tour of Chernoff Bounds. *Inf. Process. Lett.*, 33 (6):305–308, 1990. URL http://dblp.uni-trier.de/db/journals/ipl/ipl33.html#HagerupR90.

[24] Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

[25] Edward L Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008.

[26] Olivier Jeunen and Bart Goethals. Pessimistic reward models for off-policy learning in recommendation. In *Fifteenth ACM Conference on Recommender Systems*, pages 63–74, 2021.

[27] Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021.

[28] Ying Jin, Zhimei Ren, Zhuoran Yang, and Zhaoran Wang. Policy learning "without" overlap: Pessimism and generalized empirical Bernstein's inequality, 2023.

[29] Nathan Kallus and Angela Zhou. Policy evaluation and optimization with continuous treatments. In *International conference on artificial intelligence and statistics*, pages 1243–1251. PMLR, 2018.

[30] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[31] Akshay Krishnamurthy, John Langford, Aleksandrs Slivkins, and Chicheng Zhang. Contextual bandits with continuous actions: Smoothing, zooming, and adapting. *Journal of Machine Learning Research*, 21(137):1–45, 2020.

[32] Ilja Kuzborskij, Claire Vernade, Andras Gyorgy, and Csaba Szepesvári. Confident off-policy evaluation and selection through self-normalized importance weighting. In *International Conference on Artificial Intelligence and Statistics*, pages 640–648. PMLR, 2021.

[33] Tor Lattimore and Csaba Szepesvari. *Bandit Algorithms*. Cambridge University Press, 2019.

[34] Lihong Li, Remi Munos, and Csaba Szepesvari. Toward Minimax Off-policy Value Estimation. In Guy Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 608–616, San Diego, California, USA, 09–12 May 2015. PMLR. URL https://proceedings.mlr.press/v38/li15b.html.

[35] Ben London and Ted Sandler. Bayesian counterfactual risk minimization. In *International Conference on Machine Learning*, pages 4125–4133. PMLR, 2019.

[36] Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

[37] David A. McAllester. Some PAC-Bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, page 230–234, New York, NY, USA, 1998. Association for Computing Machinery. ISBN 1581130570. doi: 10.1145/279943.279989. URL https://doi.org/10.1145/279943.279989.

[38] Alberto Maria Metelli, Alessio Russo, and Marcello Restelli. Subgaussian and differentiable importance sampling for off-policy evaluation and learning. *Advances in Neural Information Processing Systems*, 34:8119–8132, 2021.

[39] Art B. Owen. *Monte Carlo theory, methods and examples*. https://artowen.su.domains/mc/, 2013.

[40] Emilio Parrado-Hernández, Amiran Ambroladze, John Shawe-Taylor, and Shiliang Sun. PAC-Bayes Bounds with Data Dependent Priors. *Journal of Machine Learning Research*, 13(112): 3507–3531, 2012. URL http://jmlr.org/papers/v13/parrado12a.html.

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

[42] James M Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

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

[44] Noveen Sachdeva, Lequn Wang, Dawen Liang, Nathan Kallus, and Julian McAuley. Off-policy evaluation for large action spaces via policy convolution. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 3576–3585, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400701719. doi: 10.1145/3589334.3645501. URL https://doi.org/10.1145/3589334.3645501.

[45] Yuta Saito and Thorsten Joachims. Off-policy evaluation for large action spaces via embeddings. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 19089–19122. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/saito22a.html.

[46] Yuta Saito, Qingyang Ren, and Thorsten Joachims. Off-policy evaluation for large action spaces via conjunct effect modeling. In *international conference on Machine learning*, pages 29734–29759. PMLR, 2023.

[47] Otmane Sakhi, Stephen Bonner, David Rohde, and Flavian Vasile. BLOB: A Probabilistic model for recommendation that combines organic and bandit signals. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 783–793, 2020.

[48] Otmane Sakhi, Louis Faury, and Flavian Vasile. Improving Offline Contextual Bandits with Distributional Robustness, 2020.

[49] Otmane Sakhi, Pierre Alquier, and Nicolas Chopin. PAC-Bayesian Offline Contextual Bandits with Guarantees. In *International Conference on Machine Learning*, pages 29777–29799. PMLR, 2023.

[50] Otmane Sakhi, David Rohde, and Nicolas Chopin. Fast Slate Policy Optimization: Going Beyond Plackett-Luce. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=f7a8XCRtUu.

[51] Otmane Sakhi, David Rohde, and Alexandre Gilotte. Fast Offline Policy Optimization for Large Scale Recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37 (8):9686–9694, Jun. 2023. doi: 10.1609/aaai.v37i8.26158. URL https://ojs.aaai.org/index.php/AAAI/article/view/26158.

[52] Yevgeny Seldin, Nicolò Cesa-Bianchi, Peter Auer, François Laviolette, and John Shawe-Taylor. PAC-Bayes-Bernstein Inequality for Martingales and its Application to Multiarmed Bandits. In Dorota Glowacka, Louis Dorard, and John Shawe-Taylor, editors, *Proceedings of the Workshop on On-line Trading of Exploration and Exploitation 2*, volume 26 of *Proceedings of Machine Learning Research*, pages 98–111, Bellevue, Washington, USA, 02 Jul 2012. PMLR. URL https://proceedings.mlr.press/v26/seldin12a.html.

[53] Anshumali Shrivastava and Ping Li. Asymmetric LSH (ALSH) for Sublinear Time Maximum Inner Product Search (MIPS). In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/310ce61c90f3a46e340ee8257bc70e93-Paper.pdf.

[54] Yi Su, Maria Dimakopoulou, Akshay Krishnamurthy, and Miroslav Dudík. Doubly robust off-policy evaluation with shrinkage. In *International Conference on Machine Learning*, pages 9167–9176. PMLR, 2020.

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

[56] Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. *advances in neural information processing systems*, 28, 2015.

[57] Adith Swaminathan, Akshay Krishnamurthy, Alekh Agarwal, Miro Dudik, John Langford, Damien Jose, and Imed Zitouni. Off-policy evaluation for slate recommendation. *Advances in Neural Information Processing Systems*, 30, 2017.

[58] Muhammad Faaiz Taufiq, Arnaud Doucet, Rob Cornish, and Jean-Francois Ton. Marginal density ratio for off-policy evaluation in contextual bandits. *Advances in Neural Information Processing Systems*, 36, 2024.

[59] Lequn Wang, Akshay Krishnamurthy, and Aleksandrs Slivkins. Oracle-efficient pessimism: Offline policy optimization in contextual bandits. *arXiv preprint arXiv:2306.07923*, 2023.

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

[61] Ding-Xuan Zhou. The covering number in learning theory. *J. Complex.*, 18(3):739–767, sep 2002. ISSN 0885-064X. doi: 10.1006/jcom.2002.0635. URL https://doi.org/10.1006/jcom.2002.0635.

Table of Contents for Supplementary Material

A   Limitations

This work develops theoretically grounded and practical pessimistic approaches for the offline contextual bandit setting. Even if the proposed algorithms are general, and provably better than competitors, they still suffer from the intrinsic limitations of importance weighting estimators. Specifically, our method, as presented, will perform poorly in *extremely* large action spaces. However, these limitations can be mitigated by incorporating additional structure as in Saito and Joachims [45], Saito et al. [46]. Another limitation arises from the offline contextual bandit setting itself, which assumes i.i.d. observations. While this assumption is valid in simple scenarios, it becomes unsuitable once we want to capture the long term effect of interventions. Extending our results to the more general, reinforcement learning setting would be an interesting research direction as it comes with a challenging estimation task and will require developing new concentration bounds.

B   Broader impact

Our work contributes to the development of theoretically grounded and practical pessimistic approaches for the offline contextual bandit setting. The derived algorithms can improve the robustness of decision-making processes by prioritizing safety and minimizing uncertainty associated risks. By leveraging pessimistic strategies, we ensure that decisions are made with a conservative bias, thereby potentially improving outcomes in high-stakes environments where the cost of errors is substantial. Although our framework and algorithms have broad, potentially good applications, their specific social impacts will solely depend on the chosen application domain.

C   Extended related work

Offline contextual bandits. Contextual bandit is a widely adopted framework for online learning in uncertain environments [33]. However, some real-world applications present challenges for existing online algorithms, and thus offline methods that leverage historical data to optimize decision-making have gained traction [10]. Fortunately, large datasets summarizing past interactions are often available, allowing agents to improve their policies offline [55]. Our work explores this offline approach, known as offline (or off-policy) contextual bandits [16]. In this setting, off-policy evaluation (OPE) estimates policy performance using historical data, mimicking real-time evaluations. Depending on the application, the goal might be to find the best policy within a predefined finite set (off-policy selection (OPS)) or the optimal policy overall (off-policy learning (OPL)).

Off-policy evaluation. In recent years, OPE has experienced a noticeable surge of interest, with numerous significant contributions [16–18, 60, 19, 54, 38, 32, 45, 47, 26]. The literature on OPE can be broadly classified into three primary approaches. The first, referred to as the direct method (DM) [26, 6], involves the development of a model designed to approximate expected costs for any context-action pair. This model is subsequently employed to estimate the performance of the policies. This approach is often designed for specific applications such as large-scale recommender systems [47, 26, 4]. The second approach, known as inverse propensity scoring (IPS) [24, 17], aims to estimate the costs associated with the evaluated policies by correcting for the inherent preference bias of the behavior policy within the dataset. While IPS maintains its unbiased nature when operating under the assumption that the evaluation policy is absolutely continuous with respect to the behavior policy, it can be susceptible to high variance and substantial bias when this assumption is violated [43]. In response to the variance issue, various techniques have been introduced, including clipping [25, 10], shrinkage [54], power-mean correction [38], implicit exploration [21], self-normalization [56], among others [22]. The third approach, known as doubly robust (DR) [42, 9, 16, 18, 19], combines elements from both the direct method (DM) and inverse propensity scoring (IPS). This work focuses on regularized IPS.

Off-policy selection and learning. as in OPE, three key approaches dominate: DM, IPS and DR in OPS and OPL. In OPS, all these methods share the same core objective: identifying the policy with the highest estimated reward from a finite set of candidates. However, they differ in their reward estimation techniques, as discussed in the OPE section above. In contrast, in OPL, DM either deterministically selects the action with the highest estimated reward or constructs a distribution based on these estimates. IPS and DR, on the other hand, employ gradient descent for policy learning [55], updating a parameterized policy denoted by π_θ as $\theta_{t+1} \leftarrow \theta_t - \nabla_\theta R(\pi_\theta)$ for each iteration t.

Since the true risk R is unknown, $\nabla_\theta R(\pi_\theta)$ is unknown and needs to be estimated using techniques like IPS or DR.

Pessimism in offline contextual bandits. Most OPE studies directly use their point estimators of the risk in OPE, OPS and OPL. However, point estimators can deviate from the true value of the risk, rendering them unreliable for decision-making. Therefore, and to increase safety, alternative approaches focus on constructing bounds on the risk. These bounds, either asymptotic [10, 48, 15] or finite sample [32, 21], aim to evaluate a policy's worst-case performance, adhering to the principle of *pessimism in face of uncertainty* [27]. The principle of pessimism transcends OPE, influencing both OPS and OPL. In these domains, strategies are predominantly inspired by, or directly derived from, upper bounds on the true risk [55, 35, 32, 49, 5, 59]. Consider OPS: [32] leveraged an Efron-Stein bound for the self-normalized IPS estimator, while [21] anchored their analysis on a bound constructed with the Implicit Exploration estimator. Shifting focus to OPL, [55] combined the empirical Bernstein bound [36] with the clipping estimator, motivating sample variance penalization for policy learning. Recent advancements include modifications to the penalization term [59] to be scalable and efficient.

PAC-Bayes extension. The PAC-Bayesian paradigm [37, 11] (see Alquier [2] for a recent introduction) provides a rich set of tools to prove generalization bounds for different statistical learning problems. The classical (online) contextual bandit problem received a lot of attention from the PAC-Bayesian community with the seminal work of Seldin et al. [52]. It is just recently that these tools were adapted to the offline contextual bandit setting, with [35] that introduced a clean and scalable PAC-Bayesian perspective to OPL. This perspective was further explored by [20, 49, 5, 7, 21], leading to the development of tight, tractable PAC-Bayesian bounds suitable for direct optimization.

Large action space extension. While regularization techniques can improve IPS properties, they often fall short when dealing with extremely large action spaces. Additional assumptions regarding the structure of the contextual bandit problem become necessary. For example, Saito and Joachims [45] introduced the Marginalized IPS (MIPS) framework and estimator. MIPS leverages auxiliary information about the actions in the form of action embeddings. Roughly speaking, MIPS assumes access to embeddings e_i within logged data and defines the risk estimator as

$$\hat{R}_n^{\text{MIPS}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \frac{\pi\left(e_i \mid x_i\right)}{\pi_0\left(e_i \mid x_i\right)} c_i = \frac{1}{n} \sum_{i=1}^{n} w\left(x_i, e_i\right) c_i \,,$$

where the logged data $\mathcal{D}_n = \{(x_i, a_i, e_i, r_i)\}_{i=1}^{n}$ now includes action embeddings for each data point. The marginal importance weight

$$w(x, e) = \frac{\pi(e \mid x)}{\pi_0(e \mid x)} = \frac{\sum_a p(e \mid x, a)\pi(a \mid x)}{\sum_a p(e \mid x, a)\pi_0(a \mid x)}$$

is a key component of this approach. Compared to IPS and DR, MIPS achieves significantly lower variance in large action spaces [45] while maintaining unbiasedness if the action embeddings directly influence costs c. This necessitates informative embeddings that capture the causal effects of actions on costs. However, high-dimensional embeddings can still lead to high variance for MIPS, similar to IPS. Additionally, high bias can arise if the direct effect assumption is violated and embeddings fail to capture these causal effects. This bias is particularly present when performing action feature selection for dimensionality reduction. Recent work proposes learning such embeddings directly from logged data [41, 44, 14], or loosen this assumption [58, 46]. Our proposed importance weight regularization can be potentially combined with these estimators under their respective assumptions on the underlying structure of the contextual bandit problem, extending our approach to large action spaces, and we posit that this will be beneficial when, for example, the action embedding dimension is high. Another line of research in large action spaces is more interested with the learning problem, precisely solving the optimization issues arising from policies defined on large action spaces. Indeed, naive optimization tends to be slow and scales linearly with the number of actions K [12]. Recent work [51, 50] solve this by leveraging fast maximum inner product search [53, 3] in the training loop, reducing the optimization complexity to *logarithmic* in the action space size. These methods however require a linear objective on the target policy. Luckily, our PAC-Bayesian learning objective is linear in the policy and its optimization is amenable to such acceleration.

Continuous action space extension. While research has predominantly focused on discrete action spaces, a limited number of studies have tackled the continuous case [29, 13, 59]. For example, [29]

explored non-parametric evaluation and learning of continuous action policies using kernel smoothing, while [13] investigated the semi-parametric setting. Recently, [59] leveraged the smoothing approach from [31] to extend their discrete OPL method to continuous actions. Our work can either use the densities directly, or be similarly extended to continuous actions through a well-defined discretization of the space. Imagine a scenario with infinitely many actions, where policies are defined by density functions. For any context x, $\pi(a \mid x)$ represents the density function that maps actions a to probabilities. The discretization process transforms the original contextual bandit problem characterized by the density-based policy class Π into an OPL problem defined by a discrete, mass-based policy class Π_K (for a finite number of actions K). Each policy within Π_K approximates a policy in Π through a smoothing process.

D   Useful lemmas

In the following, and for any quantity Z, all expectations are computed w.r.t to the distribution of the data when playing actions under the behaviour policy π_0, as in:

$$\mathbb{E}\left[Z\right] = \mathbb{E}_{x \sim \nu, a \sim \pi_0(\cdot \mid x), c \sim p(\cdot \mid x, a)}\left[Z\right].$$

A lot of the results derived in the paper are based on the use of the well known Chernoff Inequality, that we state below for a sum of i.i.d. random variables:

Lemma 12 (Chernoff Inequality for a sum of i.i.d. random variables.). *Let $a \in \mathbb{R}$, $n \in \mathbb{N}^*$ and $\{X_i, i \in [n]\}$ a collection of n i.i.d. random variables. The following concentration bounds on the right tail of $\sum_{i \in [n]} X_i$ hold for any $\lambda \geq 0$:*

$$P\left(\sum_{i \in [n]} X_i > a\right) \leq \left(\mathbb{E}\left[\exp\left(\lambda X_1\right)\right]\right)^n \exp(-\lambda a)$$

This result is classical in the literature [23] and we omit its proof. We will also need the following lemma, that states the monotonous nature of a key function in our analysis, and that we take the time to prove.

Lemma 13. *Let $L \geq 1$ and f_L be the following function:*

$$f_L(x) = \frac{\log(1+x) - \sum_{\ell=1}^{L} \frac{(-1)^{\ell-1}}{\ell} x^\ell}{(-1)^L x^{L+1}}.$$

We have that f_L is a decreasing function in $\mathbb{R}^+$ for all $L \in \mathbb{N}^*$.

Proof. Let $L \geq 1$ and f_L be the following function:

$$f_L(x) = \frac{\log(1+x) - \sum_{\ell=1}^{L} \frac{(-1)^{\ell-1}}{\ell} x^\ell}{(-1)^L x^{L+1}}.$$

Let $x \in \mathbb{R}^+$, we have the following identity holding $\forall t > 0$ and $\forall n \geq 0$:

$$\frac{1 + (-1)^n t^{n+1}}{1+t} = \sum_{k=0}^{n} (-1)^k t^k \iff \frac{1}{1+t} = \sum_{k=0}^{n} (-1)^k t^k + \frac{(-1)^{n+1} t^{n+1}}{1+t}. \tag{29}$$

Recall the integral form of the log function:

$$\log(1+x) = \int_0^x \frac{1}{1+t} dt.$$

We integrate both sides of the Equality (29) and show that the numerator of $f_L(x)$ is equal to:

$$\log(1+x) - \sum_{k=1}^{K} \frac{(-1)^{k-1}}{k} x^k = (-1)^K \int_0^x \frac{t^K}{1+t} dt.$$

18

This result enables us to rewrite the function f_L as:

$$f_L(x) = \frac{1}{x^{L+1}} \int_0^x \frac{t^L}{1+t} dt.$$

Using the change of variable $t = ux$, we obtain:

$$f_L(x) = \int_0^1 \frac{u^L}{1+xu} dt$$

which is clearly decreasing for in $\mathbb{R}^+$. This ends the proof. $\qquad\square$

Finally, we also state the important change of measure lemma:

Lemma 14 (Change of measure). *Let g be a function of the parameter θ and data $\mathcal{D}_n$, for any distribution Q that is P continuous, for any $\delta \in (0, 1]$, we have with probability $1 - \delta$:*

$$\mathbb{E}_{\theta \sim Q}[g(\theta, \mathcal{D}_n)] \leq \mathcal{KL}(Q||P) + \ln \frac{\Psi_g}{\delta} \tag{30}$$

with $\Psi_g = \mathbb{E}_{\mathcal{D}_n} \mathbb{E}_{\theta \sim P}[e^{g(\theta, \mathcal{D}_n)}]$.

Lemma 14 is the backbone of a multitude of PAC-Bayesian bounds. It is proven in many references, see for example [2] or Lemma 1.1.3 in [11]. With this result, the recipe of constructing a generalization bound reduces to choosing an adequate function g for which we can control Ψ_g.

E   Additional results and discussions

E.1   Plots of the empirical moments bounds (Proposition 1)

For any $\pi \in \Pi$, let $U_L^{\lambda,h}(\pi)$ be the upper bound of Proposition 1:

$$U_L^{\lambda,h}(\pi) = \psi_\lambda \left(\hat{R}_n^h(\pi) + \frac{\ln(1/\delta)}{\lambda n} + \sum_{\ell=2}^{2L} \frac{\lambda^{\ell-1}}{\ell} \hat{\mathcal{M}}_n^{h,\ell}(\pi) \right).$$

One can observe that the bound $U_L^{\lambda,h}$ depends on three parameters, the regularized IPS function h, the free parameter λ and the moment order L. We choose a dataset (balance-scale) with $n = 612$, and evaluate a policy π with $R(\pi) = -0.93$ to evaluate our bound for different parameters. We fix $\lambda = \sqrt{1/n}$ and plot the value of $U_L^{\lambda,h}$ for different values of the moment order $L \in \{1, 2, 3, 4, 6, 8, \infty\}$ and for 4 different regularization functions, namely IPS, clipped IPS ($M = \sqrt{n}$), Implicit Exploration (IX) ($\lambda = \sqrt{1/n}$) and Exponential Smoothing (ES) ($\alpha = 1 - \sqrt{1/n}$). The results are shown in Figure 3. One can observe from the plot that The decreasing nature of $U_L^{\lambda,h}$ depends on λ and the regularization function h. Indeed, Proposition 2 states that $\lambda < \min_{i \in [n]} 1/|h_i|$ implies that the bound is decreasing w.r.t L. Which means that once this condition is not verified, we do not know if the bound will keep decreasing with L. If the bound seems decreasing for CIPS and IX, One can observe that for both IPS and ES, the bound increased from $L = 4$ to $L = 8$, but achieved its minimum at $L = \infty$, with IPS being optimal for this value. This highlights the connection between L, the value of λ and the regularizer h.

E.2   The study of Logarithmic Smoothing estimator and proofs

Recall the form of the Logarithmic Smoothing estimator, defined for any $\lambda \geq 0$:

$$\hat{R}_n^\lambda(\pi) = -\frac{1}{n} \sum_{i=1}^n \frac{1}{\lambda} \log \left(1 - \lambda w_\pi(x_i, a_i) c_i\right) . \tag{31}$$

Our estimator $\hat{R}_n^\lambda(\pi)$, is defined for a non-negative $\lambda \geq 0$. In particular, $\lambda = 0$ recovers the unbiased IPS estimator in (2) and $\lambda > 0$ introduces a bias variance trade-off. This estimator can be interpreted
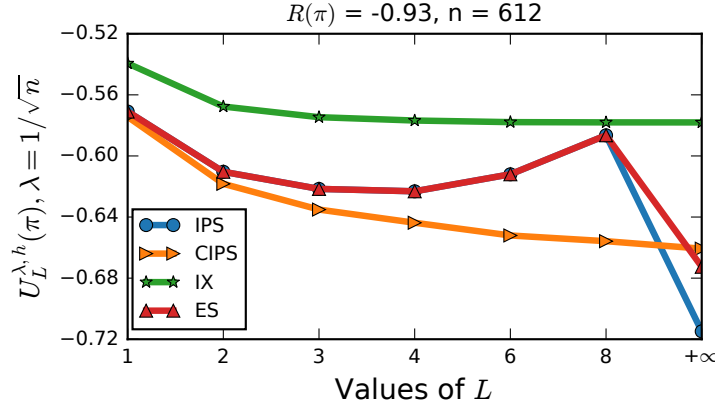
Figure 3: Proposition 1 for different values of L and with different regularized IPS h.

as Logarithmic Soft Clipping, and have a similar behavior than Clipping of Bottou et al. [10]. Indeed, $1/\lambda$ plays a similar role to the clipping parameter M, as for any $i \in [n]$, we have:

$$w_\pi(x_i, a_i)c_i \ll \frac{1}{\lambda} \implies -\frac{1}{\lambda} \log\left(1 - \lambda w_\pi(x_i, a_i)c_i\right) \approx w_\pi(x_i, a_i)c_i.$$

$$w_\pi(x_i, a_i)c_i < M \implies \min\left(w_\pi(x_i, a_i), M\right)c_i = w_\pi(x_i, a_i)c_i.$$

LS can be seen as a smooth, differentiable version of clipping. We plot the graph of the two functions in Figure 4. One can observe that once $\lambda > 0$, LS exhibits a bias-variance trade-off, with a declining bias with $\lambda \to 0$. This is different than Clipping as no bias is suffered once M is bigger than the support of w_π, this comes however with the price of suffering the full variance of IPS. In the following, we study the bias-variance trade-off that emerges with the new Logarithmic Smoothing estimator.
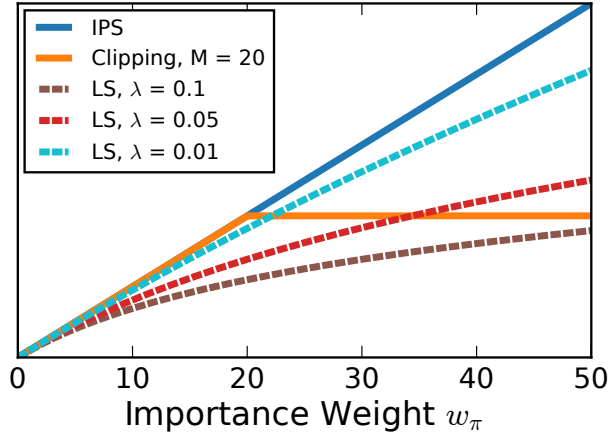
Figure 4: Comparison of Logarithmic Smoothing and Clipping.

We begin by defining the bias and variance of $\hat{R}_n^\lambda(\pi)$:

$$\mathcal{B}^\lambda(\pi) = \mathbb{E}\left[\hat{R}_n^\lambda(\pi)\right] - R(\pi), \qquad \mathcal{V}^\lambda(\pi) = \mathbb{E}\left[\left(\hat{R}_n^\lambda(\pi) - \mathbb{E}\left[\hat{R}_n^\lambda(\pi)\right]\right)^2\right]. \qquad (32)$$

Moreover, for any $\lambda \geq 0$, we define the following quantity

$$\mathcal{S}_\lambda(\pi) = \mathbb{E}\left[\frac{w_\pi(x, a)^2 c^2}{1 - \lambda w_\pi(x, a)c}\right], \qquad (33)$$

that will be essential in studying the properties of this estimator akin to the coverage ratio used for the IX-estimator [21]. In the following, we study the properties of our estimator $\hat{R}_n^\lambda(\pi)$ in (14). We start with bounding its mean squared error (MSE), which involves bounding its bias and variance.

20

> **Proposition** (Bias-variance trade-off). *Let $\pi \in \Pi$ and $\lambda \geq 0$. Then we have that*
> $$0 \leq \mathcal{B}^\lambda(\pi) \leq \lambda \mathcal{S}_\lambda(\pi), \quad \text{and} \quad \mathcal{V}^\lambda(\pi) \leq \mathcal{S}_\lambda(\pi)/n.$$
> *Moreover, it holds that for any $\lambda > 0$:*
> $$\mathcal{V}^\lambda(\pi) \leq \frac{|R(\pi)|}{n\lambda} \leq \frac{1}{n\lambda}.$$

Proof. Let us start with bounding the bias. We have for any $\lambda \geq 0$:

$$
\begin{aligned}
\mathcal{B}^\lambda(\pi) &= \mathbb{E}\left[\hat{R}_n^\lambda(\pi)\right] - R(\pi) \\
&= \mathbb{E}\left[-\frac{1}{\lambda}\log(1 - \lambda w_\pi(x,a)c) - w_\pi(x,a)c\right] \quad \text{(IPS is unbiased)}.
\end{aligned}
$$

Using $\log(1 + x) \leq x$ for any $x \geq 0$ proves that the bias is positive. For its upper bound, we use the following inequality $\log(1 + x) \geq \frac{x}{1+x}$ holding for $x \geq 0$:

$$
\begin{aligned}
\mathcal{B}^\lambda(\pi) &= \mathbb{E}\left[-\frac{1}{\lambda}\log(1 - \lambda w_\pi(x,a)c) - w_\pi(x,a)c\right] \\
&\leq \mathbb{E}\left[\frac{w_\pi(x,a)c}{1 - \lambda w_\pi(x,a)c} - w_\pi(x,a)c\right] = \lambda\mathbb{E}\left[\frac{(w_\pi(x,a)c)^2}{1 - \lambda w_\pi(x,a)c}\right] = \lambda\mathcal{S}_\lambda(\pi).
\end{aligned}
$$

Now focusing on the variance, we have:

$$
\begin{aligned}
\mathcal{V}^\lambda(\pi) &= \mathbb{E}\left[\left(\hat{R}_n^\lambda(\pi) - \mathbb{E}\left[\hat{R}_n^\lambda(\pi)\right]\right)^2\right] \\
&\leq \frac{1}{n\lambda^2}\mathbb{E}\left[\log(1 - \lambda w_\pi(x,a)c)^2\right].
\end{aligned}
$$

We use the following inequality $\log(1 + x) \leq x/\sqrt{x+1}$ holding for $x \geq 0$ to obtain our result:

$$\mathcal{V}^\lambda(\pi) \leq \frac{1}{n}\mathcal{S}_\lambda(\pi).$$

Notice that once $\lambda > 0$, we have:

$$\mathcal{S}_\lambda(\pi) = \mathbb{E}\left[\frac{w_\pi(x,a)^2 c^2}{1 - \lambda w_\pi(x,a)c}\right] \leq \frac{1}{\lambda}\mathbb{E}\left[w_\pi(x,a)|c|\right] = \frac{|R(\pi)|}{\lambda},$$

resulting in a finite variance whenever $\lambda > 0$:

$$\mathcal{V}^\lambda(\pi) \leq \frac{|R(\pi)|}{n\lambda} \leq \frac{1}{n\lambda}.$$

$\square$

$\lambda = 0$ recovers the IPS estimator in (2), with zero bias and variance bounded by $\mathbb{E}\left[w^2(x,a)c^2\right]/n$. When $\lambda > 0$, a bias-variance trade-off emerges. The bias is always non-negative as we still recover an estimator that verifies **(C1)**. The bias is capped at $\lambda\mathcal{S}_\lambda(\pi)$, which diminishes to zero when λ is small and goes to $|R(\pi)|$ as λ increases. Conversely, the variance decreases with a higher λ. Notably, $\lambda > 0$ ensures finite variance bounded by $1/\lambda n$, despite the estimator being unbounded. This is different from previous regularizations that relied on bounded functions to ensure finite variance.

While prior evaluations of estimators often relied on bias and variance analysis, Metelli et al. [38] argued for studying the non-asymptotic concentration rate of the estimators, advocating for sub-Gaussianity as a desired property. Even if our estimator is not bounded, we prove in the following that it is sub-Gaussian.

Proposition (Sub-Gaussianity). *Let $\pi \in \Pi$, $\delta \in (0,1]$ and $\lambda > 0$. Then the following inequalities holds with probability at least $1 - \delta$:*

$$R(\pi) - \hat{R}_n^\lambda(\pi) \leq \frac{\ln(2/\delta)}{\lambda n}, \qquad \text{and} \qquad \hat{R}_n^\lambda(\pi) - R(\pi) \leq \lambda \mathcal{S}_\lambda(\pi) + \frac{\ln(2/\delta)}{\lambda n}.$$

In particular, setting $\lambda = \lambda_* = \sqrt{\ln(2/\delta)/n\mathbb{E}\left[w_\pi(x,a)^2 c^2\right]}$ yields that

$$|R(\pi) - \hat{R}_n^{\lambda_*}(\pi)| \leq \sqrt{2\sigma^2 \ln(2/\delta)}, \qquad \text{where } \sigma^2 = 2\mathbb{E}\left[w_\pi(x,a)^2 c^2\right]/n. \qquad (34)$$

Proof. Let $\pi \in \Pi$, $\lambda > 0$ and $\delta > 0$. To prove sub-Gaussianity, we need both upper bounds and lower bounds on $R(\pi)$ using $\hat{R}_n^\lambda(\pi)$. For the upper bound, we can use the bound of Corollary 5, and recall that $\psi_\lambda(x) \leq x$ for all x. We then obtain with a probability $1 - \delta$:

$$R(\pi) \leq \psi_\lambda\left(\hat{R}_n^\lambda(\pi) + \frac{\ln(1/\delta)}{\lambda n}\right) \implies R(\pi) - \hat{R}_n^\lambda(\pi) \leq \frac{\ln(1/\delta)}{\lambda n}.$$

For the lower bound on the risk, we go back to our Chernoff Lemma 12, and use the collection of i.i.d. random variable, that for any $i \in [n]$, are defined as:

$$\bar{X}_i = -\frac{1}{\lambda} \log\left(1 - \lambda w_\pi(x_i, a_i) c_i\right).$$

This gives for $a \in \mathbb{R}$:

$$P\left(\sum_{i \in [n]} \bar{X}_i > a\right) \leq \left(\mathbb{E}\left[\exp\left(\lambda \bar{X}_1\right)\right]\right)^n \exp(-\lambda a)$$

$$P\left(\sum_{i \in [n]} \bar{X}_i > a\right) \leq \left(\mathbb{E}\left[\frac{1}{1 - \lambda w_\pi(x,a)c}\right]\right)^n \exp(-\lambda a)$$

Solving for $\delta = \left(\mathbb{E}\left[\frac{1}{1 - \lambda w_\pi(x,a)c}\right]\right)^n \exp(-\lambda a)$, we get:

$$P\left(\frac{1}{n}\sum_{i \in [n]} \bar{X}_i > \frac{1}{\lambda}\log\left(\mathbb{E}\left[\frac{1}{1 - \lambda w_\pi(x,a)c}\right]\right) + \frac{\ln(1/\delta)}{\lambda n}\right) \leq \delta$$

The complementary event holds with at least probability $1 - \delta$:

$$\hat{R}_n^\lambda(\pi) \leq \frac{1}{\lambda}\log\left(\mathbb{E}\left[\frac{1}{1 - \lambda w_\pi(x,a)c}\right]\right) + \frac{\ln(1/\delta)}{\lambda n},$$

which implies using the inequality $\log(x) \leq x - 1$ for all $x > 0$:

$$\hat{R}_n^\lambda(\pi) - R(\pi) \leq \frac{1}{\lambda}\log\left(\mathbb{E}\left[\frac{1}{1 - \lambda w_\pi(x,a)c}\right]\right) - R(\pi) + \frac{\ln(1/\delta)}{\lambda n}$$

$$\leq \frac{1}{\lambda}\left(\mathbb{E}\left[\frac{1}{1 - \lambda w_\pi(x,a)c}\right] - 1\right) - R(\pi) + \frac{\ln(1/\delta)}{\lambda n}$$

$$\leq \mathbb{E}\left[\frac{w_\pi(x,a)c}{1 - \lambda w_\pi(x,a)c} - w_\pi(x,a)c\right] + \frac{\ln(1/\delta)}{\lambda n}$$

$$\leq \lambda \mathbb{E}\left[\frac{w_\pi(x,a)^2 c^2}{1 - \lambda w_\pi(x,a)c}\right] + \frac{\ln(1/\delta)}{\lambda n} = \lambda \mathcal{S}_\lambda(\pi) + \frac{\ln(1/\delta)}{\lambda n},$$

which proves the lower bound on the risk. As both results hold with high probability, we use a union argument to have them both holding for probability at least $1 - \delta$:

$$R(\pi) - \hat{R}_n^\lambda(\pi) \leq \frac{\ln(2/\delta)}{\lambda n}, \qquad \text{and} \qquad \hat{R}_n^\lambda(\pi) - R(\pi) \leq \lambda \mathcal{S}_\lambda(\pi) + \frac{\ln(2/\delta)}{\lambda n},$$

which implies that:

$$|R(\pi) - \hat{R}_n^\lambda(\pi)| \leq \lambda \mathcal{S}_\lambda(\pi) + \frac{\ln(2/\delta)}{\lambda n} \leq \lambda \mathbb{E}\left[w_\pi(x,a)^2 c^2\right] + \frac{\ln(2/\delta)}{\lambda n}.$$

This means that setting $\lambda = \lambda_* = \sqrt{\ln(2/\delta)/n\mathbb{E}\left[w_\pi(x,a)^2 c^2\right]}$ yields a sub-Gaussian concentration:

$$|R(\pi) - \hat{R}_n^{\lambda_*}(\pi)| \leq 2\sqrt{\frac{\mathbb{E}\left[w_\pi(x,a)^2 c^2\right]\ln(2/\delta)}{n}}.$$

This ends the proof. $\qquad\square$

From (34), $\hat{R}_n^{\lambda_*}(\pi)$ is sub-Gaussian with variance proxy $\sigma^2 = 2\mathbb{E}\left[\omega(x,a)^2 c^2\right]/n$, which is lower that the variance proxy of the Harmonic estimator of Metelli et al. [38]. Indeed, the Harmonic estimator has a slightly worse variance proxy of $\sigma_H^2 = \frac{(2+\sqrt{3})^2}{3}\mathbb{E}\left[\omega(x,a)^2 c^2\right]/n$, giving $\sigma^2 < \sigma_H^2$.

E.3   OPS: Formal comparison with IX suboptimality

Let us begin by stating results from the IX work [21]. Recall that the IX estimator is defined for any $\lambda > 0$, by:

$$\hat{R}_n^{\lambda\text{-IX}}(\pi) = \frac{1}{n}\sum_{i=1}^n \frac{\pi(a_i|x_i)}{\pi_0(a_i|x_i) + \lambda/2}c_i.$$

Let $\Pi_s = \{\pi_1, ..., \pi_m\}$ be a finite set of predefined policies. In OPS, the goal is to find $\pi_*^s \in \Pi_s$ that satisfies

$$\pi_*^s = \operatorname{argmin}_{\pi \in \Pi_s} R(\pi) = \operatorname{argmin}_{k \in [m]} R(\pi_k).$$

for $\lambda > 0$, the selection strategy suggested in Gabbianelli et al. [21] was to search for:

$$\hat{\pi}_n^{s,\text{ IX}} = \operatorname*{argmin}_{\pi \in \Pi_s} \hat{R}_n^{\lambda\text{-IX}}(\pi) = \operatorname*{argmin}_{k \in [m]} \hat{R}_n^{\lambda\text{-IX}}(\pi). \tag{35}$$

Proposition 15 (Suboptimality of the IX selection strategy)**.** *Let $\lambda > 0$ and $\delta \in (0,1]$. Then, it holds with probability at least $1 - \delta$ that*

$$0 \leq R(\hat{\pi}_n^{s,\text{ IX}}) - R(\pi_*^s) \leq \lambda\mathcal{C}_{\lambda/2}(\pi_*^s) + \frac{2\ln(2|\Pi_s|/\delta)}{\lambda n}, \tag{36}$$

where

$$\mathcal{C}_\lambda(\pi) = \mathbb{E}\left[\frac{\pi(a|x)}{\pi_0^2(a|x) + \lambda\pi_0(a|x)}|c|\right].$$

Both suboptimalities (LS and IX) have the same form, they only depend on two different quantities ($\mathcal{S}_\lambda$ and $\mathcal{C}_\lambda$ respectively). For a $\pi \in \Pi$ and $\lambda > 0$, If we can identify when $\mathcal{S}_\lambda(\pi) \leq \mathcal{C}_{\lambda/2}(\pi)$, then we can prove that the sub-optimality of LS selection strategy is better than the one of IX. Luckily, this is always the case, and it is stated formally below.

Proposition 16. *Let $\pi \in \Pi$ and $\lambda > 0$. We have:*

$$\mathcal{S}_\lambda(\pi) \leq \mathcal{C}_{\lambda/2}(\pi). \tag{37}$$

Proof. Let $\pi \in \Pi$ and $\lambda > 0$, we have:

$$\mathcal{C}_{\lambda/2}(\pi) - \mathcal{S}_\lambda(\pi) = \mathbb{E}\left[\frac{\pi(a|x)}{\pi_0^2(a|x) + \frac{\lambda}{2}\pi_0(a|x)}|c| - \frac{w_\pi(x,a)^2 c^2}{1 - \lambda w_\pi(x,a)c}\right]$$

$$= \mathbb{E}\left[\frac{\pi(a|x)}{\pi_0^2(a|x) + \frac{\lambda}{2}\pi_0(a|x)}|c| - \frac{\pi(a|x)^2 c^2}{\pi_0^2(a|x) - \lambda\pi_0(a|x)\pi(a|x)c}\right]$$

$$= \mathbb{E}\left[\pi(a|x)|c|\left(\frac{1}{\pi_0^2(a|x) + \frac{\lambda}{2}\pi_0(a|x)} - \frac{\pi(a|x)|c|}{\pi_0^2(a|x) + \lambda\pi_0(a|x)\pi(a|x)|c|}\right)\right]$$

$$= \mathbb{E}\left[\pi(a|x)|c|\left(\frac{\pi_0^2(a|x)\left(1 - \pi(a|x)|c|\right) + \frac{\lambda}{2}\pi_0(a|x)\pi(a|x)|c|}{(\pi_0^2(a|x) + \frac{\lambda}{2}\pi_0(a|x))(\pi_0^2(a|x) + \lambda\pi_0(a|x)\pi(a|x)|c|)}\right)\right]$$

$$\geq 0.$$

$\square$

This means that the suboptimality of LS selection strategy is better bounded than the one of IX. Our experiments confirm that the LS selection strategy is better than IX in practical scenarios.

Minimax optimality of our selection strategy. As discussed in Gabbianelli et al. [21], pessimistic algorithms tend to have the property that their regret scales with the minimax sample complexity of estimating the value of the optimal policy [27]. For the case of multi-armed bandit (one context x), this estimation minimax sample complexity is proved by Li et al. [34] and is of the rate $\mathcal{O}(\mathbb{E}[w_{\pi^*}(x,a)^2 c^2])$, with π^* being the optimal policy. Our bound matches the lower bound proved by Li et al. [34], as:

$$\mathcal{S}_\lambda(\pi^*) = \mathbb{E}\left[\frac{w_{\pi^*}(x,a)^2 c^2}{1 - \lambda w_{\pi^*}(x,a)c}\right] \leq \mathbb{E}\left[w_{\pi^*}(x,a)^2 c^2\right],$$

which is not the case for the suboptimality of IX, that only matches it in the deterministic setting with binary costs, as:

$$\mathcal{C}_\lambda(\pi^*) = \mathbb{E}\left[\frac{\pi^*(a|x)}{\pi_0^2(a|x) + \lambda\pi_0(a|x)}|c|\right] \leq \mathbb{E}\left[\frac{\pi^*(a|x)}{\pi_0^2(a|x)}|c|\right] = \mathbb{E}\left[\left(\frac{\pi^*(a|x)}{\pi_0(a|x)}\right)^2 c^2\right],$$

with the last inequality only holding when π^* is deterministic and the costs are binary. For deterministic policies and the general contextual bandit, we invite the reader to see a formal proof of the minimax lower bound of pessimism in Jin et al. [28, Theorem 4.4], matched for both IX and LS.

E.4   OPL: Formal comparison of PAC-Bayesian bounds

As it is easier to work with linear estimators within the PAC-Bayesian framework, we define the following estimator of the risk $\hat{R}_n^{p-\text{LIN}}(\pi)$, with the help of a function $p : \mathbb{R} \to \mathbb{R}$ as:

$$\hat{R}_n^{p-\text{LIN}}(\pi) = \frac{1}{n}\sum_{i=1}^n \frac{\pi(a_i|x_i)}{p(\pi_0(a_i|x_i))}c_i$$

with the only condition on p to be $\{C_1^{\text{LIN}} : \forall x, p(x) \geq x\}$. This condition helps us control the impact of actions with low probabilities under π_0. This risk estimator encompasses well known risk estimators depending on the choice of p.

Now that we defined the family of estimators covered by our analysis, we attack the problem of deriving generalization bounds. We derive our empirical high order bound expressed in the following:

Proposition 17 (Empirical High Order PAC-Bayes bound)**.** *Let $L \geq 1$. Given a prior P on $\mathcal{F}_\Theta$, $\delta \in (0, 1]$ and $\lambda > 0$, the following bound holds with probability at least $1 - \delta$ uniformly for all distribution Q over $\mathcal{F}_\Theta$:*

$$R(\pi_Q) \leq \psi_\lambda \left(\hat{R}_n^{p-\text{LIN}}(\pi_Q) + \frac{\mathcal{KL}(Q||P) + \ln \frac{1}{\delta}}{\lambda n} + \sum_{\ell=2}^{2L} \frac{\lambda^{\ell-1}}{\ell} \hat{\mathcal{M}}_n^{p-\text{LIN},\ell}(\pi_Q) \right) \quad (38)$$

with:

$$\hat{\mathcal{M}}_n^{p-\text{LIN},\ell}(\pi_Q) = \frac{1}{n} \sum_{i=1}^n \frac{\pi_Q(a_i|x_i)}{p(\pi_0(a_i|x_i))^\ell} c_i^\ell$$

$$\psi_\lambda = x :\to \frac{1 - \exp(-\lambda x)}{\lambda}.$$

Proof. Let $L \geq 1$, we have from Lemma 13, and for any positive random variable $X \geq 0$ and $\lambda > 0$:

$$f_{2L-1}(0) = \frac{1}{2L} \geq f_{2L-1}(\lambda X) = -\frac{\log(1 + \lambda X) - \sum_{\ell=1}^{2L-1} \frac{(-1)^{\ell-1}}{\ell}(\lambda X)^\ell}{(\lambda X)^{2L}}$$

which is equivalent to:

$$\sum_{\ell=1}^{2L} \frac{(-1)^{\ell-1}}{\ell}(\lambda X)^\ell \leq \log(1 + \lambda X) \iff \exp\left(\sum_{\ell=1}^{2L} \frac{(-1)^{\ell-1}}{\ell}(\lambda X)^\ell \right) \leq 1 + \lambda X$$

$$\implies \mathbb{E}\left[\exp\left(\sum_{\ell=1}^{2L} \frac{(-1)^{\ell-1}}{\ell}(\lambda X)^\ell \right) \right] \leq 1 + \mathbb{E}\left[\lambda X \right]$$

$$\implies \mathbb{E}\left[\exp\left(\sum_{\ell=1}^{2L} \frac{(-1)^{\ell-1}}{\ell}(\lambda X)^\ell \right) \right] \leq \exp\left(\log(1 + \mathbb{E}\left[\lambda X \right]) \right),$$

which implies that:

$$\mathbb{E}\left[\exp\left(\lambda(X - \frac{1}{\lambda} \log(1 + \mathbb{E}\left[\lambda X \right])) + \sum_{\ell=2}^{2L} \frac{(-1)^{\ell-1}}{\ell}(\lambda X)^\ell \right) \right] \leq 1.$$

For any $X \leq 0$, we can inject $-X \geq 0$ to obtain:

$$\forall X \leq 0, \quad \mathbb{E}\left[\exp\left(\lambda \left(-\frac{1}{\lambda} \log(1 + \mathbb{E}\left[\lambda X \right]) - X \right) - \sum_{k=2}^{2K} \frac{1}{k}(\lambda X)^k \right) \right] \leq 1. \quad (39)$$

Let:

$$d_\theta(a|x) = \mathbb{1}\left[f_\theta(x) = a \right], \forall(x, a) \in \mathcal{X} \times \mathcal{A},$$

it means that:

$$\pi_Q(a|x) = \mathbb{E}_{\theta \sim Q}\left[d_\theta(a|x) \right], \forall(x, a) \in \mathcal{X} \times \mathcal{A}.$$

Let $\lambda > 0$. The adequate function g we are going to use in combination with Lemma 14 is:

$$g(\theta, \mathcal{D}_n) = \sum_{i=1}^n \lambda \left(-\frac{1}{\lambda} \log(1 + \lambda R^{p-\text{LIN}}(d_\theta)) - \frac{d_\theta(a_i|x_i)}{p(\pi_0(a_i|x_i))} c_i \right) - \sum_{\ell=2}^{2L} \frac{1}{\ell} \left(\lambda \frac{d_\theta(a_i|x_i)}{p(\pi_0(a_i|x_i))} c_i \right)^\ell$$

$$= \sum_{i=1}^n \lambda \left(-\frac{1}{\lambda} \log(1 + \lambda R^{p-\text{LIN}}(d_\theta)) - \frac{d_\theta(a_i|x_i)}{p(\pi_0(a_i|x_i))} c_i \right) - \sum_{\ell=2}^{2L} \frac{d_\theta(a_i|x_i)}{\ell} \left(\frac{\lambda}{p(\pi_0(a_i|x_i))} c_i \right)^\ell.$$

By exploiting the i.i.d. nature of the data and exchanging the order of expectations (P is independent of $\mathcal{D}_n$), we can naturally prove using (39) that:

$$\Psi_g = \mathbb{E}_P \left[\prod_{i=1}^n \mathbb{E} \left[\exp \left(\lambda \left(-\frac{1}{\lambda} \log(1 + \lambda R^{p-\text{LIN}}(d_\theta)) - X_i(\theta) \right) - \sum_{k=2}^{2K} \frac{1}{k} (\lambda X_i(\theta))^k \right) \right] \right] \leq 1,$$

as we have :

$$X_i(\theta) = \frac{d_\theta(a_i|x_i)}{p(\pi_0(a_i|x_i))} c_i \leq 0 \quad \forall i.$$

Injecting Ψ_g in Lemma 14, rearranging terms and using that $\hat{R}_n^{p-\text{LIN}}(\pi)$ has positive bias concludes the proof. $\qquad\square$

Similarly to the OPE section, we use this general bound to obtain a PAC-Bayesian Empirical Second Moment bound and the PAC-Bayesian LS-LIN bound. That we state directly below:

Empirical second moment bound. With $L = 1$, we obtain the following:

Corollary 18 (Second Moment Upper bound). *Given a prior P on $\mathcal{F}_\Theta$, $\delta \in (0,1]$ and $\lambda > 0$. The following bound holds with probability at least $1 - \delta$ uniformly for all distribution Q over $\mathcal{F}_\Theta$:*

$$R(\pi_Q) \leq \psi_\lambda \left(\hat{R}_n^p(\pi_Q) + \frac{\mathcal{KL}(Q||P) + \ln \frac{1}{\delta}}{\lambda n} + \frac{\lambda}{2} \hat{\mathcal{M}}_n^{p-\text{LIN},2}(\pi_Q) \right). \tag{40}$$

Log Smoothing PAC-Bayesian Bound. With $L \to \infty$, we obtain the following:

Proposition 19 ($\hat{R}_n^{\lambda-\text{LIN}}$ PAC-Bayes bound). *Given a prior P on $\mathcal{F}_\Theta$, $\delta \in (0,1]$ and $\lambda > 0$, the following bound holds with probability at least $1 - \delta$ uniformly for all distribution Q over $\mathcal{F}_\Theta$:*

$$R(\pi_Q) \leq \psi_\lambda \left(\hat{R}_n^{\lambda-\text{LIN}}(\pi_Q) + \frac{\mathcal{KL}(Q||P) + \ln \frac{1}{\delta}}{\lambda n} \right). \tag{41}$$

with:

$$\hat{R}_n^{\lambda\text{-LIN}}(\pi) = -\frac{1}{n} \sum_{i=1}^n \frac{\pi(a_i|x_i)}{\lambda} \log \left(1 - \frac{\lambda c_i}{\pi_0(a_i|x_i)} \right).$$

Following the same proof schema as of the OPE section, we can demonstrate that the Log Smoothing PAC-Bayesian bound dominates the Empirical Second moment PAC-Bayesian bound $L = 1$. However, we use the bound of $L = 1$ as an intermediary to state the dominance of the Log Smoothing PAC-Bayesian bound.

Indeed, we can easily compare the result obtained with $L = 1$ to previously derived PAC-Bayesian bounds for off-policy learning. We start by writing down the conditional Bernstein bound of Sakhi et al. [49] holding for the (linear) cIPS ($p : x \to \max(x, \tau)$). For a policy π_Q and a $\lambda > 0$, we have:

$$R(\pi_Q) \leq \hat{R}_n^\tau(\pi_Q) + \sqrt{\frac{\mathcal{KL}(Q||P) + \ln \frac{4\sqrt{n}}{\delta}}{2n}} + \frac{\mathcal{KL}(Q||P) + \ln \frac{2}{\delta}}{\lambda n} + \lambda g\left(\lambda/\tau\right) \mathcal{V}_n^\tau(\pi_Q).$$

$$R(\pi_Q) \leq \hat{R}_n^\tau(\pi_Q) + \frac{\mathcal{KL}(Q||P) + \ln \frac{1}{\delta}}{\lambda n} + \frac{\lambda}{2} \hat{\mathcal{S}}_n^\tau(\pi_Q). \tag{L = 1}$$

We can observe that the previously derived conditional Bernstein bound has several terms that make it less tight:

- It has an additional, strictly positive square root KL divergence term.

- The multiplicative factor $g(\lambda/\tau)$ is always bigger than $1/2$, and diverges when $\tau \to 0$.
- With enough data ($n \gg 1$), we also have:

$$\hat{\mathcal{S}}_n^\tau(\pi_Q) \approx \mathbb{E}\left[\frac{\pi_Q(a|x)}{\max\{\pi_0(a|x), \tau\}^2}c(a,x)^2\right] \leq \mathbb{E}\left[\frac{\pi_Q(a|x)}{\max\{\pi_0(a|x), \tau\}^2}\right] \approx \mathcal{V}_n^\tau(\pi_Q).$$

These observations confirm that the new bound derived with $L = 1$ is tighter than what was previously proposed for $\mathtt{cIPS}$, especially when $n \gg 1$. As our bound can work for other estimators, we also compare it to a recently proposed PAC-Bayes bound in Aouali et al. [5] for the exponentially-smoothed estimator ($p : x \to x^\alpha$) with $\alpha \in [0, 1]$:

$$R(\pi_Q) \leq \hat{R}_n^\alpha(\pi_Q) + \sqrt{\frac{\mathcal{KL}(Q||P) + \ln\frac{4\sqrt{n}}{\delta}}{2n}} + \frac{\mathcal{KL}(Q||P) + \ln\frac{2}{\delta}}{\lambda n} + \frac{\lambda}{2}\left(\mathcal{V}_n^\alpha(\pi_Q) + \hat{\mathcal{S}}_n^\alpha(\pi_Q)\right).$$

$$R(\pi_Q) \leq \hat{R}_n^\alpha(\pi_Q) + \frac{\mathcal{KL}(Q||P) + \ln\frac{1}{\delta}}{\lambda n} + \frac{\lambda}{2}\hat{\mathcal{S}}_n^\alpha(\pi_Q). \tag{L = 1}$$

We can clearly see that the previously proposed bound for the exponentially smoothed estimator has two additional positive quantities that makes it less tight than our bound. In addition, computing our bound does not rely on expectations under π_0 (contrary to the previous bounds that have $\mathcal{V}_n$) which alleviates the need to access the logging policy and reduce the computations.

This demonstrates the superiority of $L = 1$ compared to existing variance sensitive PAC-Bayesian bounds. It means that $L \to \infty$ is even better. We can also prove that the Log smoothing PAC-Bayesian Bound is better than the one of IX in Gabbianelli et al. [21]. Indeed, using $\log(1 + x) \geq \frac{x}{1+x/2}$ for all $x \geq 0$, we have for any $P, Q \in \mathcal{P}(\Theta)$ and $\lambda > 0$:

$$\psi_\lambda\left(\hat{R}_n^{\lambda-\mathrm{LIN}}(\pi_Q) + \frac{\mathcal{KL}(Q||P) + \ln\frac{1}{\delta}}{\lambda n}\right) \leq \hat{R}_n^{\lambda-\mathrm{LIN}}(\pi_Q) + \frac{\mathcal{KL}(Q||P) + \ln\frac{1}{\delta}}{\lambda n}$$

$$\leq -\frac{1}{n}\sum_{i=1}^n \frac{\pi_Q(a_i|x_i)}{\lambda}\log\left(1 - \frac{\lambda c_i}{\pi_0(a_i|x_i)}\right) + \frac{\mathcal{KL}(Q||P) + \ln\frac{1}{\delta}}{\lambda n}$$

$$\leq \frac{1}{n}\sum_{i=1}^n \frac{\pi_Q(a_i|x_i)}{\pi_0(a_i|x_i) - \lambda c_i/2} + \frac{\mathcal{KL}(Q||P) + \ln\frac{1}{\delta}}{\lambda n}$$

$$\leq \hat{R}_n^{\lambda-\mathrm{IX}}(\pi_Q) + \frac{\mathcal{KL}(Q||P) + \ln\frac{1}{\delta}}{\lambda n}, \tag{IX-bound}$$

with the last implication leveraging that the cost is always bigger than $-1/$ This proves that our bound is better than the IX bound. This means that our PAC-Bayesian bound is better than all existing PAC-Bayesian off-policy learning bounds.

E.5 OPL: Formal comparison with IX PAC-Bayesian learning suboptimality

Let us begin by stating results from the IX work [21]. Recall that the IX estimator is defined for any $\lambda > 0$, by:

$$\hat{R}_n^{\mathrm{IX}-\lambda}(\pi) = \frac{1}{n}\sum_{i=1}^n \frac{\pi(a_i|x_i)}{\pi_0(a_i|x_i) + \lambda/2}c_i,$$

and that we used the linearized version of the $\mathtt{LS}$ estimator, $\mathtt{LS-LIN}$ defined as:

$$\hat{R}_n^{\lambda\text{-}\mathrm{LIN}}(\pi) = -\frac{1}{n}\sum_{i=1}^n \frac{\pi(a_i|x_i)}{\lambda}\log\left(1 - \frac{\lambda c_i}{\pi_0(a_i|x_i)}\right).$$

Let Θ be a parameter space and $\mathcal{P}(\Theta)$ be the set of all probability distribution on Θ. Our goal is to find the best policy in a chosen class $\mathcal{L}(\Theta) \subset \mathcal{P}(\Theta)$:

$$\pi_{Q^*} = \underset{Q \in \mathcal{L}(\Theta)}{\mathrm{argmin}}\, R(\pi_Q).$$

For $\lambda > 0$ and a prior $P \in \mathcal{P}(\Theta)$, the PAC-Bayesian learning strategy suggested in Gabbianelli et al. [21] is to find in $\mathcal{L}(\Theta) \subset \mathcal{P}(\Theta)$:

$$\hat{\pi}_{Q_n}^{\mathrm{IX}} = \operatorname*{argmin}_{Q \in \mathcal{L}(\Theta)} \left\{ \hat{R}_n^{\mathrm{IX}-\lambda}(\pi_Q) + \frac{\mathcal{KL}(Q||P)}{\lambda n} \right\}.$$

This learning strategy suffers from a suboptimality bounded in the result below:

Proposition 20 (Suboptimality of the IX PAC-Bayesian learning strategy from [21])**.** *Let* $\lambda > 0$ *and* $\delta \in (0,1]$*. Then, it holds with probability at least* $1 - \delta$ *that*

$$0 \leq R(\hat{\pi}_{Q_n}^{\mathrm{IX}}) - R(\pi_{Q^*}) \leq \lambda \mathcal{C}_{\lambda/2}(\pi_{Q^*}) + \frac{2\left(\mathcal{KL}(Q^*||P) + \ln(2/\delta)\right)}{\lambda n},$$

where

$$\mathcal{C}_\lambda(\pi) = \mathbb{E}\left[\frac{\pi(a|x)}{\pi_0^2(a|x) + \lambda\pi_0(a|x)} |c| \right].$$

Similarly for PAC-Bayesian learning, both suboptimalities (LS and IX) have the same form, they only depend on two different quantities ($\mathcal{S}_\lambda^{\mathrm{LIN}}$ and $\mathcal{C}_\lambda$ respectively). For a $\pi \in \Pi$ and $\lambda > 0$, If we can identify when $\mathcal{S}_\lambda^{\mathrm{LIN}}(\pi) \leq \mathcal{C}_{\lambda/2}(\pi)$, then we can prove that the sub-optimality of LS PAC-Bayesian learning strategy is better than the one of IX in certain cases. Luckily, this is always the case, and it is stated formally below.

Proposition 21. *Let* $\pi \in \Pi$ *and* $\lambda > 0$*. We have:*

$$\mathcal{S}_\lambda^{\mathrm{LIN}}(\pi) \leq \mathcal{C}_{\lambda/2}(\pi). \tag{42}$$

Proof. Let $\pi \in \Pi$ and $\lambda > 0$, and recall that:

$$\mathcal{S}_\lambda^{\mathrm{LIN}}(\pi) = \mathbb{E}\left[\frac{\pi(a|x)c^2}{\pi_0^2(a|x) - \lambda\pi_0(a|x)c} \right].$$

We have:

$$\mathcal{C}_{\lambda/2}(\pi) - \mathcal{S}_\lambda^{\mathrm{LIN}}(\pi) = \mathbb{E}\left[\frac{\pi(a|x)}{\pi_0^2(a|x) + \frac{\lambda}{2}\pi_0(a|x)} |c| - \frac{\pi(a|x)c^2}{\pi_0^2(a|x) - \lambda\pi_0(a|x)c} \right]$$

$$= \mathbb{E}\left[\pi(a|x)|c| \left(\frac{1}{\pi_0^2(a|x) + \frac{\lambda}{2}\pi_0(a|x)} - \frac{|c|}{\pi_0^2(a|x) + \lambda\pi_0(a|x)|c|} \right) \right]$$

$$= \mathbb{E}\left[\pi(a|x)|c| \left(\frac{\pi_0^2(a|x)(1 - |c|) + \frac{\lambda}{2}\pi_0(a|x)|c|}{(\pi_0^2(a|x) + \frac{\lambda}{2}\pi_0(a|x))(\pi_0^2(a|x) + \lambda\pi_0(a|x)|c|)} \right) \right]$$

$$\geq 0.$$

$\square$

Similarly, this means that the suboptimality of LS-LIN PAC-Bayesian learning strategy is also, better bounded than the one of IX.

Minimax optimality of our learning strategy. From Jin et al. [28, Theorem 4.4] we can state that the minimax suboptimality lower bound, in the case of deterministic optimal policies is of the rate $\mathcal{O}(1/\sqrt{nC^*})$ with $\inf_{x \in \mathcal{X}} \pi_0(\pi^*(x)|x) > C^*$. Our bound as well as IX bound match this minimax lower bound, as:

$$\mathcal{S}_\lambda^{\mathrm{LIN}}(\pi^*) = \mathbb{E}_{x,c}\left[\frac{c^2}{\pi_0(\pi^*(x)|x) - \lambda c} \right] \leq \frac{1}{C^*}$$

$$\mathcal{C}_\lambda(\pi^*) = \mathbb{E}_{x,c}\left[\frac{|c|}{\pi_0(\pi^*(x)|x) + \lambda} \right] \leq \frac{1}{C^*}.$$

One can see that for both, selecting a

$$\lambda^* = \sqrt{\frac{2\left(\mathcal{KL}(Q^*||P) + \ln(2/\delta)\right)C^*}{n}},$$

gets you the desired bound, matching this minimax rate.

F   Proofs of OPE

F.1   Proof of high order empirical moments bound (Proposition 1)

> **Proposition** (Empirical moments risk bound). *Let $\pi \in \Pi$, $L \geq 1$, $\delta \in (0,1]$, $\lambda > 0$ and h satisfying (C1). Then it holds with probability at least $1 - \delta$ that*
>
> $$R(\pi) \leq \psi_\lambda\Big(\hat{R}_n^h(\pi) + \sum_{\ell=2}^{2L} \frac{\lambda^{\ell-1}}{\ell} \hat{\mathcal{M}}_n^{h,\ell}(\pi) + \frac{\ln(1/\delta)}{\lambda n}\Big),$$
>
> *where ψ_λ and $\hat{\mathcal{M}}_n^{h,\ell}(\pi)$ are defined in (5), respectively, and recall that $\psi_\lambda(x) \leq x$.*

Proof. Let $L \in \mathbb{N}^*$, $\lambda > 0$ and $X \geq 0$ a **positive random variable**. We have $2L - 1 \geq 1$, and with the decreasing nature of $f_{(2L-1)}$ (Lemma 13), we also have:

$$f_{(2L-1)}(0) \geq f_{2L-1}(\lambda X) \iff \frac{1}{2L} \geq -\frac{\log(1 + \lambda X) - \sum_{l=1}^{2L-1} \frac{(-1)^{\ell-1}}{k}(\lambda X)^\ell}{(\lambda X)^{2L}}$$

$$\iff \sum_{\ell=1}^{2L} \frac{(-1)^{\ell-1}}{k}(\lambda X)^\ell \leq \log(1 + \lambda X)$$

$$\iff \exp\left(\sum_{\ell=1}^{2L} \frac{(-1)^{\ell-1}}{\ell}(\lambda X)^\ell\right) \leq 1 + \lambda X$$

$$\implies \mathbb{E}\left[\exp\left(\sum_{\ell=1}^{2L} \frac{(-1)^{\ell-1}}{\ell}(\lambda X)^\ell\right)\right] \leq 1 + \lambda \mathbb{E}[X]$$

$$\implies \mathbb{E}\left[\exp\left(\sum_{\ell=1}^{2L} \frac{(-1)^{\ell-1}}{\ell}(\lambda X)^\ell\right)\right] \leq \exp\left((\log(1 + \lambda \mathbb{E}[X]))\right)$$

$$\implies \mathbb{E}\left[\exp\left(\lambda(X - \frac{1}{\lambda}\log\left(1 + \lambda \mathbb{E}[X]\right)) + \sum_{\ell=2}^{2L} \frac{(-1)^{\ell-1}}{\ell}(\lambda X)^\ell\right)\right] \leq 1.$$

For any $X \leq 0$, we can inject $-X \geq 0$ to obtain:

$$\forall X \leq 0, \quad \mathbb{E}\left[\exp\left(\lambda\left(-\frac{1}{\lambda}\log\left(1 - \lambda\mathbb{E}[X]\right) - X\right) - \sum_{\ell=2}^{2L} \frac{1}{\ell}(\lambda X)^\ell\right)\right] \leq 1. \qquad (43)$$

The result in Equation (43) will be combined with Chernoff Inequality (Lemma 12) to finally prove our bound. Let $\lambda > 0$, for our problem, we define the random variable X_i to use in the Chernoff Inequality as:

$$X_i = -\frac{1}{\lambda}\log\left(1 - \lambda\mathbb{E}[h]\right) - h_i - \sum_{\ell=2}^{2L} \frac{1}{\ell}(\lambda h_i)^\ell.$$

For any $a \in \mathbb{R}$, this gives us the following:

$$P\left(\sum_{i\in[n]} X_i > a\right) \leq (\mathbb{E}\left[\exp\left(\lambda X_1\right)\right])^n \exp(-\lambda a)$$

$$P\left(-\frac{n}{\lambda}\log\left(1 - \lambda\mathbb{E}\left[h\right]\right) - \sum_{i\in[n]}\left(h_i + \sum_{\ell=2}^{2L}\frac{1}{\ell}(\lambda h_i)^\ell\right) > a\right) \leq (\mathbb{E}\left[\exp\left(\lambda X_1\right)\right])^n \exp(-\lambda a)$$

$$P\left(-\frac{n}{\lambda}\log\left(1 - \lambda\mathbb{E}\left[h\right]\right) - \sum_{i\in[n]}\left(h_i + \sum_{\ell=2}^{2L}\frac{1}{\ell}(\lambda h_i)^\ell\right) > a\right) \leq \exp(-\lambda a) \quad \text{(Use of Equation (43))}$$

Solving for $\delta = \exp(-\lambda a)$, we get:

$$P\left(-\frac{n}{\lambda}\log\left(1 - \lambda\mathbb{E}\left[h\right]\right) - \sum_{i\in[n]}\left(h_i + \sum_{\ell=2}^{2L}\frac{1}{\ell}(\lambda h_i)^\ell\right) > \frac{\ln(1/\delta)}{\lambda}\right) \leq \delta$$

$$P\left(-\frac{1}{\lambda}\log\left(1 - \lambda\mathbb{E}\left[h\right]\right) - \frac{1}{n}\sum_{i\in[n]}\left(h_i + \sum_{\ell=2}^{2L}\frac{\lambda^\ell}{\ell}h_i^\ell\right) > \frac{\ln(1/\delta)}{\lambda n}\right) \leq \delta$$

$$P\left(-\frac{1}{\lambda}\log\left(1 - \lambda\mathbb{E}\left[h\right]\right) - \hat{R}_n^h(\pi) - \sum_{\ell=2}^{2L}\frac{\lambda^{\ell-1}}{\ell}\hat{\mathcal{M}}_n^{h,\ell}(\pi) > \frac{\ln(1/\delta)}{\lambda n}\right) \leq \delta$$

$$P\left(-\frac{1}{\lambda}\log\left(1 - \lambda\mathbb{E}\left[h\right]\right) > \hat{R}_n^h(\pi) + \sum_{\ell=2}^{2L}\frac{\lambda^{\ell-1}}{\ell}\hat{\mathcal{M}}_n^{h,\ell}(\pi)\frac{\ln(1/\delta)}{\lambda n}\right) \leq \delta.$$

This means that the following, complementary event will hold with probability at least $1 - \delta$:

$$-\frac{1}{\lambda}\log\left(1 - \lambda\mathbb{E}\left[h\right]\right) \leq \hat{R}_n^h(\pi) + \sum_{\ell=2}^{2L}\frac{\lambda^{\ell-1}}{\ell}\hat{\mathcal{M}}_n^{h,\ell}(\pi)\frac{\ln(1/\delta)}{\lambda n}.$$

ψ_λ being a non-decreasing function, applying it to the two sides of this inequality gives us:

$$\mathbb{E}\left[h\right] \leq \psi_\lambda\left(\hat{R}_n^h(\pi) + \sum_{\ell=2}^{2L}\frac{\lambda^{\ell-1}}{\ell}\hat{\mathcal{M}}_n^{h,\ell}(\pi) + \frac{\ln(1/\delta)}{\lambda n}\right).$$

Finally, h satisfies (**C1**), this means that the bound is also an upper bound on the true risk, giving:

$$R(\pi) \leq \psi_\lambda\left(\hat{R}_n^h(\pi) + \sum_{\ell=2}^{2L}\frac{\lambda^{\ell-1}}{\ell}\hat{\mathcal{M}}_n^{h,\ell}(\pi) + \frac{\ln(1/\delta)}{\lambda n}\right),$$

which concludes the proof. $\qquad\square$

F.2 Proof of the impact of L on the bound's tightness (Proposition 2)

> **Proposition** (Impact of L on the bound's tightness). *Let $\pi \in \Pi$, $\delta \in (0,1]$, $\lambda > 0$, $L \geq 1$ and h satisfying* (C1). *Let*
>
> $$U_L^{\lambda,h}(\pi) = \psi_\lambda \left(\hat{R}_n^h(\pi) + \frac{\ln(1/\delta)}{\lambda n} + \sum_{\ell=2}^{2L} \frac{\lambda^{\ell-1}}{\ell} \hat{\mathcal{M}}_n^{h,\ell}(\pi) \right)$$
>
> *be the upper bound in Equation* (6)*. Then,*
>
> $$\lambda \leq \min_{i \in [n]} \left\{ \frac{2L+2}{(2L+1)|h_i|} \right\} \implies U_{L+1}^{\lambda,h}(\pi) \leq U_L^{\lambda,h}(\pi). \tag{44}$$
>
> *which implies that:*
>
> $$\lambda \leq \min_{i \in [n]} \left\{ \frac{1}{|h_i|} \right\} \implies U_L^{\lambda,h}(\pi) \text{ is a decreasing function w.r.t } L.$$

Proof. We want to prove the implication (44) from which the condition on the decreasing nature of our bound will follow. Indeed, Let us suppose that (44) is true, we have:

$$\lambda \leq \min_{i \in [n]} \left\{ \frac{1}{|h_i|} \right\} \implies \forall L \geq 1, \quad \lambda \leq \min_{i \in [n]} \left\{ \frac{2L+2}{(2L+1)|h_i|} \right\}$$

$$\implies \forall L \geq 1, \quad U_{L+1}^{\lambda,h}(\pi) \leq U_L^{\lambda,h}(\pi) \quad \text{(Using (44))}$$

$$\implies U_L^{\lambda,h}(\pi) \text{ is a decreasing function w.r.t } L.$$

Now let us prove the implication in (44). We have for any $L \geq 1$:

$$U_{L+1}^{\lambda,h}(\pi) \leq U_L^{\lambda,h}(\pi) \iff \sum_{\ell=2L+1}^{2L+2} \frac{\lambda^{\ell-1}}{\ell} \hat{\mathcal{M}}_n^{h,\ell}(\pi) \leq 0$$

$$\iff \frac{\lambda^{2L}}{n} \sum_{i=1}^n h_i^{2L+1} \left(\frac{1}{2L+1} + \frac{\lambda h_i}{2L+2} \right) \leq 0$$

As $h_i \leq 0$, we can ensure this inequality by choosing a λ that verifies:

$$\forall i \in [n], \quad \lambda \leq \left\{ \frac{2L+2}{(2L+1)|h_i|} \right\} \iff \lambda \leq \min_{i \in [n]} \left\{ \frac{2L+2}{(2L+1)|h_i|} \right\}$$

which concludes the proof. $\qquad \square$

F.3 Comparisons of the bounds U_L^λ (Proposition 3)

We compare the bounds evaluated in their optimal regularisation function h. We start by stating the proposition and proving it.

> **Proposition.** *Let $\pi \in \Pi$, and $\lambda > 0$, we define:*
>
> $$U_L^\lambda(\pi) = \min_h U_L^{\lambda,h}(\pi).$$
>
> *Then, for any $\lambda > 0$, it holds that for any $L > 1$:*
>
> $$U_L^\lambda(\pi) \leq U_1^\lambda(\pi).$$
>
> *In particular, $\forall \lambda > 0$:*
>
> $$U_\infty^\lambda(\pi) \leq U_1^\lambda(\pi), \tag{45}$$

Proof. Let $\pi \in \Pi$, $\lambda > 0$ and

$$U_L^\lambda(\pi) = \min_h U_L^{\lambda,h}(\pi).$$

We can prove (see Appendix F.4) that:

$$U_1^\lambda(\pi) = U_1^{\lambda,h_{*,1}}(\pi) = \psi_\lambda \left(\hat{R}_n^{h_{*,1}}(\pi) + \frac{\lambda}{2} \hat{\mathcal{M}}_n^{h_{*,1},2}(\pi) + \frac{\ln(1/\delta)}{\lambda n} \right)$$

with:

$$h_{*,1}(p,q,c) = -\min(|c|p/q, 1/\lambda),$$

and that (see Appendix F.7):

$$U_\infty^\lambda(\pi) = \psi_\lambda \left(\hat{R}_n^\lambda(\pi) + \frac{\ln(1/\delta)}{\lambda n} \right).$$

From Proposition 2, we have that for any h:

$$\lambda \le \min_{i \in [n]} \left\{ \frac{1}{|h_i|} \right\} \implies U_L^{\lambda,h}(\pi) \text{ is a decreasing function w.r.t } L.$$

It appears that the optimal function $h_{*,1}$ respects this condition, as by definition:

$$\min_{i \in [n]} \left\{ \frac{1}{|(h_{*,1})_i|} \right\} \ge \lambda,$$

meaning that:

$$U_L^{\lambda,h_{*,1}}(\pi) \text{ is a decreasing function w.r.t } L.$$

This result suggests that the Empirical Second Moment bound, evaluated in its optimal function $h_{*,1}$, is always bigger than bounds with additional moments (evaluated in the same $h_{*,1}$). This leads us to the result wanted, as for any $L > 1$:

$$U_L^\lambda(\pi) = \min_h U_L^{\lambda,h}(\pi) \le U_L^{\lambda,h_{*,1}}(\pi) \le U_1^{\lambda,h_{*,1}}(\pi) = U_1^\lambda(\pi).$$

In particular, we get:

$$U_\infty^\lambda(\pi) \le U_1^\lambda(\pi),$$

which ends the proof. $\qquad\square$

This means that U_∞^λ is tighter than U_1^λ, and thus can also be tighter than empirical Bernstein.

F.4 Proof of the optimality of global clipping for Corollary 4

Proposition (Optimal h for $L = 1$)**.** *Let $\lambda > 0$. The function h that minimizes the bound for $L = 1$, giving the tightest result is:*

$$\forall i, \quad h_i = h(\pi(a_i|x_i), \pi_0(a_i|x_i), c_i)) = -\min \left\{ \frac{\pi(a_i|x_i)}{\pi_0(a_i|x_i)}|c_i|, \frac{1}{\lambda} \right\}$$

This means that when the costs are binary, we obtain the classical Clipping estimator of parameter $1/\lambda$:

$$h_i = \min \left\{ \frac{\pi(a_i|x_i)}{\pi_0(a_i|x_i)}, \frac{1}{\lambda} \right\} c_i.$$

Proof. We want to look for the value of h that minimizes the bound. Formally, by fixing all variables of the bound, this problem reduces to:

$$\underset{h \in \mathbf{(C1)}}{\mathrm{argmin}} \ \hat{R}_n^h(\pi) + \frac{\lambda}{2} \hat{\mathcal{M}}_n^{h,2}(\pi) = \underset{h \in \mathbf{(C1)}}{\mathrm{argmin}} \ \frac{1}{n} \sum_{i=1}^{n} \left(h_i + \frac{\lambda}{2} h_i^2 \right).$$

The objective decomposes across data points, so we can solve it for every h_i independently. Let us fix a $j \in [n]$, the following problem:

$$\operatorname*{argmin}_{h_j \in \mathbb{R}} \hat{R}_n^h(\pi) + \frac{\lambda}{2} \hat{\mathcal{M}}_n^{h,2}(\pi) = \operatorname*{argmin}_{h_j \in \mathbb{R}} \left\{ h_j + \frac{\lambda}{2} h_j^2 \right\}$$

$$\text{subject to} \quad h_j \geq \frac{\pi(a_j|x_j)}{\pi_0(a_j|x_j)} c_j$$

is strongly convex in h_j. We write the KKT conditions for h_j to be optimal; there exists α^* that verifies:

$$1 + \lambda h_j - \alpha^* = 0 \tag{46}$$

$$\alpha^* \geq 0 \tag{47}$$

$$\alpha^* \left(\frac{\pi(a_j|x_j)}{\pi_0(a_j|x_j)} c_j - h_j \right) = 0 \tag{48}$$

$$h_j \geq \frac{\pi(a_j|x_j)}{\pi_0(a_j|x_j)} c_j \tag{49}$$

We study the two following two cases:

Case 1: $h_j \leq -\frac{1}{\lambda}$:
we have $\alpha^* = 1 + \lambda h_j \leq 0 \implies \alpha^* = 0$, meaning that:

$$h_j = -\frac{1}{\lambda}$$

Case 2: $h_j > -\frac{1}{\lambda}$:
we have $\alpha^* = 1 + \lambda h_j > 0$, which combined to condition (36) gives:

$$h_j = \frac{\pi(a_j|x_j)}{\pi_0(a_j|x_j)} c_j.$$

The two results combined mean that we always have:

$$h_j \geq -\frac{1}{\lambda}, \text{ and whenever } h_j > -\frac{1}{\lambda} \implies h_j = \frac{\pi(a_j|x_j)}{\pi_0(a_j|x_j)} c_j.$$

We deduce that h_j has the following form:

$$h_j = h(\pi(a_j|x_j), \pi_0(a_j|x_j), c_j) = -\min\left\{ \frac{\pi(a_j|x_j)}{\pi_0(a_j|x_j)} |c_j|, \frac{1}{\lambda} \right\} \tag{50}$$

$$\alpha^* = 1 - \lambda \min\left\{ \frac{\pi(a_j|x_j)}{\pi_0(a_j|x_j)} |c_j|, \frac{1}{\lambda} \right\} \tag{51}$$

These values verify the KKT conditions. As the problem is strongly convex, h_j has a unique possible value and must be equal to equation (38). The form of h_j is a global clipping that includes the cost in the function as well. In the case where the cost function c is binary:

$$\forall i \quad c_i \in \{-1, 0\},$$

we recover the classical Clipping with parameter $1/\lambda$ as an optimal solution for h:

$$h_j = \min\left\{ \frac{\pi(a_j|x_j)}{\pi_0(a_j|x_j)}, \frac{1}{\lambda} \right\} c_j.$$

$\square$

F.5 Comparison with empirical Bernstein

We begin by comparing the Second Moment Bound with Swaminathan and Joachims [55]'s bound as they both manipulate similar quantities. The bound of [55] uses the Empirical Bernstein bound of [36] applied to the Clipping Estimator. We recall its expression below for a parameter $M > 0$:

$$\hat{R}_n^M(\pi) = \frac{1}{n} \sum_{i=1}^n \min\left\{ \frac{\pi(a_i|x_i)}{\pi_0(a_i|x_i)}, M \right\} c_i.$$

We also give below the Empirical Bernstein Bound applied to this estimator:

Proposition (Empirical Bernstein for Clipping of [55]). *Let $\pi \in \Pi$, $\delta \in (0,1]$ and $M > 0$. Then it holds with probability at least $1 - \delta$ that*

$$R(\pi) \leq \hat{R}_n^M(\pi) + \sqrt{\frac{2\hat{V}_n^M(\pi)\ln(2/\delta)}{n}} + \frac{7M\ln(2/\delta)}{3(n-1)}, \tag{52}$$

with $\hat{V}_n^M(\pi)$ the empirical variance of the clipping estimator.

We are usually interested in the case where π and π_0 are different, leading to substantial importance weights. In this practical scenario, the variance and the second moment are of the same magnitude of M. Indeed, one can see it from the following equality:

$$\underbrace{\hat{V}_n^M(\pi)}_{\mathcal{O}(M)} = \underbrace{\hat{\mathcal{M}}_n^{M,2}(\pi)}_{\mathcal{O}(M)} - \underbrace{\left(\hat{R}_n^M(\pi)\right)^2}_{\mathcal{O}(\bar{c}^2)}$$

$$\approx \underbrace{\hat{\mathcal{M}}_n^{M,2}(\pi)}_{\mathcal{O}(M)} \quad (M \gg \bar{c}^2 = o(1).)$$

This means that in practical scenarios, the empirical variance and the empirical second moment are approximately the same. Recall that the Second Moment Bound works for any regularizer h, As Clipping satisfies **(C1)**, we give the Second Moment Upper of Corollary 4 with Clipping below:

$$\psi_\lambda\left(\hat{R}_n^M(\pi) + \frac{\lambda}{2}\hat{\mathcal{M}}_n^{M,2}(\pi) + \frac{\ln(1/\delta)}{\lambda n}\right) \leq \hat{R}_n^M(\pi) + \frac{\lambda}{2}\hat{\mathcal{M}}_n^{M,2}(\pi) + \frac{\ln(1/\delta)}{\lambda n} \quad (\psi_\lambda(x) \leq x, \forall x)$$

$$\leq \hat{R}_n^M(\pi) + \frac{\lambda}{2}\hat{\mathcal{M}}_n^{M,2}(\pi) + \frac{\ln(1/\delta)}{\lambda n}.$$

Choosing a $\lambda \approx \sqrt{2\ln(1/\delta)/(n\hat{\mathcal{M}}_n^{M,2}(\pi))}$ gives us an upper bound that is close to:

$$\hat{R}_n^M(\pi) + \frac{\lambda}{2}\hat{\mathcal{M}}_n^{M,2}(\pi) + \frac{\ln(1/\delta)}{\lambda n} \approx \hat{R}_n^M(\pi) + \sqrt{\frac{2\hat{\mathcal{M}}_n^{M,2}(\pi)\ln(1/\delta)}{n}}$$

$$\approx \hat{R}_n^M(\pi) + \sqrt{\frac{2\hat{V}_n^M(\pi)\ln(1/\delta)}{n}}$$

$$\leq \hat{R}_n^M(\pi) + \sqrt{\frac{2\hat{V}_n^M(\pi)\ln(2/\delta)}{n}} + \frac{7M\ln(2/\delta)}{3(n-1)}.$$

This means that in practical scenarios, and with a good choice of $\lambda \sim \mathcal{O}(1/\sqrt{n})$, the Second Moment bound would be better than the Empirical Bernstein bound, and this difference will be even greater when $M \gg 1$. This is aligned with our experiments, where we see that the new Second Moment bound is much tighter in practice. This also confirms that the Logarithmic smoothing bound is even tighter, because it is smaller than the Second Moment bound as stated in Proposition 3.

F.6 Proof of the $L \to \infty$ bound (Corollary 5)

> **Proposition** (Empirical Logarithmic Smoothing bound with $L \to \infty$). *Let $\pi \in \Pi$, $\delta \in (0, 1]$ and $\lambda > 0$. Then it holds with probability at least $1 - \delta$ that*
>
> $$R(\pi) \leq \psi_\lambda\Big(-\frac{1}{n} \sum_{i=1}^{n} \frac{1}{\lambda} \log\left(1 - \lambda h_i\right) + \frac{\ln(1/\delta)}{\lambda n} \Big).$$

Taking the limit of L naively recovers this form of the bound, but imposes a condition on λ for the bound to converge. We instead, take another path of proof that does not impose any condition on λ, developed below. The main idea is to take the limit of L to recover the variable to use along Chernoff.

Proof. Recall that for the proof of the Empirical moments bounds, we used the following random variable defined with $\lambda > 0$:

$$X_i = -\frac{1}{\lambda} \log\left(1 - \lambda \mathbb{E}\left[h\right]\right) - h_i - \sum_{\ell=2}^{2L} \frac{1}{\ell} (\lambda h_i)^\ell,$$

combined with Chernoff Inequality (Lemma 12) to prove our bound. If we take the limit $L \to \infty$ for our random variable, we obtain the following random variable:

$$
\begin{aligned}
\tilde{X}_i &= -\frac{1}{\lambda} \log\left(1 - \lambda \mathbb{E}\left[h\right]\right) + \frac{1}{\lambda} \log\left(1 - \lambda h_i\right) \\
&= \frac{1}{\lambda} \log\left(\frac{1 - \lambda h_i}{1 - \lambda \mathbb{E}\left[h\right]} \right).
\end{aligned}
$$

We use the random variable $\tilde{X}_i$ with the Chernoff Inequality. For any $a \in \mathbb{R}$, we have:

$$P\left(\sum_{i \in [n]} \tilde{X}_i > a \right) \leq \left(\mathbb{E}\left[\exp\left(\lambda \tilde{X}_1 \right) \right] \right)^n \exp(-\lambda a)$$

$$P\left(-\frac{n}{\lambda} \log\left(1 - \lambda \mathbb{E}\left[h\right]\right) + \sum_{i \in [n]} \left(\frac{1}{\lambda} \log\left(1 - \lambda h_i\right) \right) > a \right) \leq \left(\mathbb{E}\left[\exp\left(\lambda \tilde{X}_1 \right) \right] \right)^n \exp(-\lambda a)$$

On the other hand, we have:

$$\mathbb{E}\left[\exp\left(\lambda \tilde{X}_1 \right) \right] = \frac{\mathbb{E}\left[1 - \lambda h_i\right]}{1 - \lambda \mathbb{E}\left[h\right]} = 1.$$

Using this equality and solving for $\delta = \exp(-\lambda a)$, we get:

$$P\left(-\frac{n}{\lambda} \log\left(1 - \lambda \mathbb{E}\left[h\right]\right) + \sum_{i \in [n]} \left(\frac{1}{\lambda} \log\left(1 - \lambda h_i\right) \right) > \frac{\ln(1/\delta)}{\lambda} \right) \leq \delta$$

$$P\left(-\frac{1}{\lambda} \log\left(1 - \lambda \mathbb{E}\left[h\right]\right) + \frac{1}{n} \sum_{i \in [n]} \frac{1}{\lambda} \log\left(1 - \lambda h_i\right) > \frac{\ln(1/\delta)}{\lambda n} \right) \leq \delta$$

This means that the following, complementary event will hold with probability at least $1 - \delta$:

$$-\frac{1}{\lambda} \log\left(1 - \lambda \mathbb{E}\left[h\right]\right) \leq -\frac{1}{n} \sum_{i=1}^{n} \frac{1}{\lambda} \log\left(1 - \lambda h_i\right) + \frac{\ln(1/\delta)}{\lambda n}.$$

ψ_λ being a non-decreasing function, applying it to the two sides of this inequality gives us:

$$\mathbb{E}\left[h\right] \leq \psi_\lambda\Big(-\frac{1}{n} \sum_{i=1}^{n} \frac{1}{\lambda} \log\left(1 - \lambda h_i\right) + \frac{\ln(1/\delta)}{\lambda n} \Big).$$

As h satisfies **(C1)**, we obtain the required inequality:

$$R(\pi) \leq \psi_\lambda\Big(-\frac{1}{n} \sum_{i=1}^{n} \frac{1}{\lambda} \log\left(1 - \lambda h_i\right) + \frac{\ln(1/\delta)}{\lambda n} \Big).$$

and conclude the proof. $\qquad\square$

F.7 Proof of the optimality of IPS for Corollary 5

> **Proposition** (Optimal h for $L \to \infty$)**.** *Let $\lambda > 0$. The function h that minimizes the bound for $L \to \infty$, giving the tightest result is:*
>
> $$\forall i, \quad h_i = h(\pi(a_i|x_i), \pi_0(a_i|x_i), c_i)) = \frac{\pi(a_i|x_i)}{\pi_0(a_i|x_i)} c_i$$

Proof. The proof of this proposition is quite simple. The function:

$$f(x) = -\log(1 - \lambda x)$$

is increasing. This means that the lowest possible value of h_i ensures the tightest result. As our variables h_i verifies **(C1)**, we recover IPS as an optimal choice for this bound. $\square$

F.8 Comparison with the IX bound (Proposition 8)

We now attack the recently derived IX bound in Gabbianelli et al. [21] and show that our newly proposed bound dominates it in all scenarios.

> **Proposition** (Comparison with IX [21])**.** *Let $\pi \in \Pi$, $\delta \in]0,1]$ and $\lambda > 0$, the IX bound from [21] states that we have with at least probability $1 - \delta$:*
>
> $$R(\pi) \le \hat{R}_n^{\lambda\text{-IX}}(\pi) + \frac{\ln(1/\delta)}{\lambda n} \tag{53}$$
>
> *with:*
>
> $$\hat{R}_n^{\lambda\text{-IX}}(\pi) = \frac{1}{n}\sum_{i=1}^{n} \frac{\pi(a_i|x_i)}{\pi_0(a_i|x_i) + \lambda/2} c_i.$$
>
> *Let $U_{\text{IX}}^\lambda(\pi)$ be the IX upper bound defined above, we have for any $\lambda > 0$:*
>
> $$U_\infty^\lambda(\pi) \le U_{\text{IX}}^\lambda(\pi). \tag{54}$$

Proof. Let $\pi \in \Pi$, $\delta \in]0,1]$ and $\lambda > 0$. Recall that $U_\infty^\lambda(\pi) = \psi_\lambda\left(\hat{R}_n^\lambda(\pi) + \frac{\ln(1/\delta)}{\lambda n}\right)$. We have:

$$\psi_\lambda\left(\hat{R}_n^\lambda(\pi) + \frac{\ln(1/\delta)}{\lambda n}\right) \le \hat{R}_n^\lambda(\pi) + \frac{\ln(1/\delta)}{\lambda n} \quad (\forall x, \psi_\lambda(x) \le x)$$

$$\le -\frac{1}{n}\sum_{i=1}^{n} \frac{1}{\lambda} \log(1 - \lambda w_\pi(x_i, a_i)c_i) + \frac{\ln(1/\delta)}{\lambda n}.$$

Using the inequality $\log(1+x) \ge \frac{x}{1+x/2}$ for all $x > 0$, we get:

$$U_\infty^\lambda(\pi) \le -\frac{1}{n}\sum_{i=1}^{n} \frac{1}{\lambda} \log(1 - \lambda w_\pi(x_i, a_i)c_i) + \frac{\ln(1/\delta)}{\lambda n}$$

$$\le \frac{1}{n}\sum_{i=1}^{n} \frac{w_\pi(x_i, a_i)}{1 - \lambda w_\pi(x_i, a_i)c_i/2} c_i + \frac{\ln(1/\delta)}{\lambda n} \quad \left(\log(1+x) \ge \frac{x}{1+x/2}\right)$$

$$\le \frac{1}{n}\sum_{i=1}^{n} \frac{\pi(a_i|x_i)}{\pi_0(a_i|x_i) - \lambda\pi(a_i|x_i)c_i/2} c_i + \frac{\ln(1/\delta)}{\lambda n}$$

$$\le \frac{1}{n}\sum_{i=1}^{n} \frac{\pi(a_i|x_i)}{\pi_0(a_i|x_i) + \lambda/2} c_i + \frac{\ln(1/\delta)}{\lambda n} \quad (-\pi(a_i|x_i)c_i \le 1 \text{ and } c_i \le 0)$$

$$\le \hat{R}_n^{IX-\lambda}(\pi) + \frac{\ln(1/\delta)}{\lambda n} = U_{IX}^\lambda(\pi),$$

which ends the proof. $\square$

The result states the dominance of the LS bound compared to IX. The proof of this result also gives us insight on when the LS bound will be much tighter than IX. Indeed, to obtain the IX bound, LS bound is loosened through 3 steps:

1. The use of $\psi_\lambda(x) \leq x, \forall x$.
2. The use of $\log(1 + \lambda x) \geq \frac{\lambda x}{1 + \lambda x/2}, \forall x \geq 0$.
3. The use of $-\pi(a_i|x_i)c_i \leq 1, \forall i \in [n]$.

The two first inequalities are loose when $\lambda \sim 1/\sqrt{n}$ is not too small, which means that LS will be much better in problems with few samples. The third inequality is loose when π is not a peaked policy or the cost is way less than 1. Even if LS bound is always smaller than IX, LS will give way better result if the number of samples is small, and/or the policy evaluated is diffused.

G   Proofs of OPS and OPL

G.1   OPS: Proof of suboptimality bound (Proposition 9)

Proposition (Suboptimality of our selection strategy in (20)). *Let $\lambda > 0$ and $\delta \in (0, 1]$. Then, it holds with probability at least $1 - \delta$ that*

$$0 \leq R(\hat{\pi}_n^{\mathrm{S}}) - R(\pi_*^{\mathrm{S}}) \leq \lambda\mathcal{S}_\lambda(\pi_*^{\mathrm{S}}) + \frac{2\ln(2|\Pi_{\mathrm{S}}|/\delta)}{\lambda n},$$

where π_*^{S} and $\hat{\pi}_n^{\mathrm{S}}$ are defined in (19) and (20), and

$$\mathcal{S}_\lambda(\pi) = \mathbb{E}\left[(w_\pi(x,a)c)^2/(1 - \lambda w_\pi(x,a)c)\right].$$

In addition, our upper bound is always finite as:

$$\lambda\mathcal{S}_\lambda(\pi) = \lambda\mathbb{E}\left[\frac{(w_\pi(x,a)c)^2}{1 - \lambda w_\pi(x,a)c}\right] \leq \min\left\{|R(\pi)|, \lambda\mathbb{E}\left[(w_\pi(x,a)c)^2\right]\right\} \leq |R(\pi)|.$$

Proof. To prove this bound on the suboptimality of our selection method, we need both an upper bound and a lower bound on the true risk using the LS estimator. Luckily, we already have derived them in **??** . For a fixed λ, taking a union of the two bounds over the cardinal of the finite policy class $|\Pi_s|$, we get the following holding with probability at least $1 - \delta$ for all $\pi \in \Pi_s$:

$$R(\pi) - \hat{R}_n^\lambda(\pi) \leq \frac{\ln(2|\Pi_s|/\delta)}{\lambda n}, \qquad \text{and} \qquad \hat{R}_n^\lambda(\pi) - R(\pi) \leq \lambda\mathcal{S}_\lambda(\pi) + \frac{\ln(2|\Pi_s|/\delta)}{\lambda n}.$$

As $\hat{\pi}_n^{\mathrm{S}} \in \Pi_s$ and by definition of $\hat{\pi}_n^{\mathrm{S}}$ (minimizer of $\hat{R}_n^\lambda(\pi)$), we have:

$$R(\hat{\pi}_n^{\mathrm{S}}) \leq \hat{R}_n^\lambda(\hat{\pi}_n^{\mathrm{S}}) + \frac{\ln(2|\Pi_s|/\delta)}{\lambda n} \leq \hat{R}_n^\lambda(\hat{\pi}_*^{\mathrm{S}}) + \frac{\ln(2|\Pi_s|/\delta)}{\lambda n}.$$

Using the lower bound on the risk of $R(\hat{\pi}_*^{\mathrm{S}})$, we have:

$$R(\hat{\pi}_n^{\mathrm{S}}) \leq \hat{R}_n^\lambda(\hat{\pi}_*^{\mathrm{S}}) + \frac{\ln(2|\Pi_s|/\delta)}{\lambda n}$$
$$\leq R(\hat{\pi}_*^{\mathrm{S}}) + \lambda\mathcal{S}_\lambda(\hat{\pi}_*^{\mathrm{S}}) + \frac{2\ln(2|\Pi_s|/\delta)}{\lambda n}.$$

which gives us the suboptimality upper bound:

$$0 \leq R(\hat{\pi}_n^{\mathrm{S}}) - R(\pi_*^{\mathrm{S}}) \leq \lambda\mathcal{S}_\lambda(\pi_*^{\mathrm{S}}) + \frac{2\ln(2|\Pi_{\mathrm{S}}|/\delta)}{\lambda n}.$$

Note that:

$$\lambda\mathcal{S}_\lambda(\pi) = \lambda\mathbb{E}\left[\frac{(w_\pi(x,a)c)^2}{1 - \lambda w_\pi(x,a)c}\right] \leq \min\left\{|R(\pi)|, \lambda\mathbb{E}\left[(w_\pi(x,a)c)^2\right]\right\},$$

always ensuring a finite bound. $\square$

G.2 OPL: Proof of PAC-Bayesian LS-LIN bound (Proposition 10)

> **Proposition** (PAC-Bayes learning bound for $\hat{R}_n^{\lambda-\texttt{LIN}}$). *Given a prior $P \in \mathcal{P}(\Theta)$, $\delta \in (0, 1]$ and $\lambda > 0$, the following holds with probability at least $1 - \delta$:*
> $$\forall Q \in \mathcal{P}(\Theta), \quad R(\pi_Q) \leq \psi_\lambda \left(\hat{R}_n^{\lambda-\texttt{LIN}}(\pi_Q) + \frac{\mathcal{KL}(Q||P) + \ln \frac{1}{\delta}}{\lambda n} \right)$$

Proof. To prove this proposition, we can either take the path of High Order Empirical moments as for Pessimistic OPE, or we can prove it directly. We provide here a simple proof of this proposition using ideas from Alquier [1, Corollary 2.5]. Let:

$$d_\theta(a|x) = \mathbb{1}\left[f_\theta(x) = a \right], \forall(x, a) \in \mathcal{X} \times \mathcal{A}, \tag{55}$$

it means that:

$$\pi_Q(a|x) = \mathbb{E}_{\theta \sim Q}\left[d_\theta(a|x) \right], \forall(x, a) \in \mathcal{X} \times \mathcal{A}.$$

Recall that to prove a PAC-Bayesian generalization bound, one can rely on the Change of measure Lemma (Lemma 14). For any $\lambda > 0$, the adequate function g to consider is:

$$
\begin{aligned}
g(\theta, \mathcal{D}_n) &= \sum_{i=1}^n \left(-\log(1 - \lambda R(d_\theta)) + \log\left(1 - \lambda \frac{d_\theta(a_i|x_i)c_i}{\pi_0(a_i|x_i)} \right) \right) \\
&= \sum_{i=1}^n \log\left(\frac{1 - \lambda \frac{d_\theta(a_i|x_i)c_i}{\pi_0(a_i|x_i)}}{1 - \lambda R(d_\theta)} \right).
\end{aligned}
$$

By exploiting the i.i.d. nature of the data and exchanging the order of expectations (P is independent of $\mathcal{D}_n$), we can naturally prove that:

$$
\begin{aligned}
\Psi_g &= \mathbb{E}_P \left[\prod_{i=1}^n \mathbb{E}\left[\exp\left(\log\left(\frac{1 - \lambda \frac{d_\theta(a_i|x_i)c_i}{\pi_0(a_i|x_i)}}{1 - \lambda R(d_\theta)} \right) \right) \right] \right] \\
&= \mathbb{E}_P \left[\prod_{i=1}^n \mathbb{E}\left[\frac{1 - \lambda \frac{d_\theta(a_i|x_i)c_i}{\pi_0(a_i|x_i)}}{1 - \lambda R(d_\theta)} \right] \right] \\
&= \mathbb{E}_P \left[\prod_{i=1}^n \frac{1 - \lambda R(d_\theta)}{1 - \lambda R(d_\theta)} \right] = 1.
\end{aligned}
$$

Injecting Ψ_g in Lemma 14, gives:

$$
\begin{aligned}
\mathbb{E}_{\theta \sim Q}\left[-\log(1 - \lambda R(d_\theta)) \right] &\leq \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\theta \sim Q}\left[-\log\left(1 - \lambda \frac{d_\theta(a_i|x_i)c_i}{\pi_0(a_i|x_i)} \right) \right] + \frac{\mathcal{KL}(Q||P) + \ln \frac{1}{\delta}}{n} \\
&\leq \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\theta \sim Q}\left[-d_\theta(a_i|x_i)\log\left(1 - \lambda \frac{c_i}{\pi_0(a_i|x_i)} \right) \right] + \frac{\mathcal{KL}(Q||P) + \ln \frac{1}{\delta}}{n} \\
&\leq -\frac{1}{n} \sum_{i=1}^n \pi_Q(a_i|x_i)\log\left(1 - \lambda \frac{c_i}{\pi_0(a_i|x_i)} \right) + \frac{\mathcal{KL}(Q||P) + \ln \frac{1}{\delta}}{n} \\
&\leq \lambda \hat{R}_n^{\lambda-\texttt{LIN}}(\pi_Q) + \frac{\mathcal{KL}(Q||P) + \ln \frac{1}{\delta}}{n}.
\end{aligned}
$$

From the convexity of $x \to -\log(1 + x)$, we have:

$$-\frac{1}{\lambda}\log\left(1 - \lambda R(\pi_Q) \right) \leq \frac{1}{\lambda}\mathbb{E}_{\theta \sim Q}\left[-\log(1 - \lambda R(d_\theta)) \right] \leq \hat{R}_n^{\lambda-\texttt{LIN}}(\pi_Q) + \frac{\mathcal{KL}(Q||P) + \ln \frac{1}{\delta}}{\lambda n}.$$

Applying the increasing function ψ_λ of Equation (5) to both sides concludes the proof. $\square$

G.3 OPL: Proof of PAC-Bayesian suboptimality bound (Proposition 11)

> **Proposition** (Suboptimality of the learning strategy in (27)). *Let $\lambda > 0$, $P \in \mathcal{L}(\Theta)$ and $\delta \in (0, 1]$. Then, it holds with probability at least $1 - \delta$ that*
>
> $$0 \leq R(\hat{\pi}_{Q_n}) - R(\pi_{Q^*}) \leq \lambda \mathcal{S}_\lambda^{\texttt{LIN}}(\pi_{Q^*}) + \frac{2\left(\mathcal{KL}(Q^*||P) + \ln(2/\delta)\right)}{\lambda n},$$
>
> *where*
>
> $$\mathcal{S}_\lambda^{\texttt{LIN}}(\pi) = \mathbb{E}\left[\frac{\pi(a|x)c^2}{\pi_0^2(a|x) - \lambda \pi_0(a|x)c}\right].$$
>
> *In addition, our upper bound is always finite as:*
>
> $$\lambda \mathcal{S}_\lambda^{\texttt{LIN}}(\pi) \leq \min\left\{|R(\pi)|, \lambda \mathbb{E}\left[\frac{\pi(a|x)c^2}{\pi_0^2(a|x)}\right]\right\} \leq |R(\pi)|.$$

Proof. To prove this bound on the suboptimality of our learning strategy, we need both a PAC-Bayesian upper bound and a lower bound on the true risk using the LS-LIN estimator. Luckily, we already have derived an upper bound in Proposition 10, that we linearize here as $\psi_\lambda(x) \leq x$:

$$\forall Q \in \mathcal{P}(\Theta), \quad R(\pi_Q) \leq \hat{R}_n^{\lambda-\texttt{LIN}}(\pi_Q) + \frac{\mathcal{KL}(Q||P) + \ln\frac{1}{\delta}}{\lambda n}.$$

For the lower bound, we rely a second time on the Change of measure Lemma (Lemma 14). For any $\lambda > 0$, we choose the following function g:

$$g(\theta, \mathcal{D}_n) = \sum_{i=1}^n \left(-\frac{1}{\lambda}\log\left(1 - \lambda\frac{d_\theta(a_i|x_i)c_i}{\pi_0(a_i|x_i)}\right) - R(d_\theta) - \lambda\mathcal{S}_\lambda^{\texttt{LIN}}(d_\theta)\right).$$

By exploiting the i.i.d. nature of the data and exchanging the order of expectations (P is independent of $\mathcal{D}_n$), we can prove that:

$$\Psi_g = \mathbb{E}_P\left[\prod_{i=1}^n \left(\exp\left(-\lambda(R(d_\theta) + \lambda\mathcal{S}_\lambda^{\texttt{LIN}}(d_\theta))\right) \mathbb{E}\left[\frac{1}{1 - \lambda\frac{d_\theta(a|x)c}{\pi_0(a|x)}}\right]\right)\right]$$

$$\leq \mathbb{E}_P\left[\prod_{i=1}^n \left(\exp\left(-\lambda(R(d_\theta) + \lambda\mathcal{S}_\lambda^{\texttt{LIN}}(d_\theta)) + \mathbb{E}\left[\frac{1}{1 - \lambda\frac{d_\theta(a|x)c}{\pi_0(a|x)}}\right] - 1\right)\right)\right]$$

$$\leq \mathbb{E}_P\left[\prod_{i=1}^n \left(\exp\left(-\lambda(R(d_\theta) + \lambda\mathcal{S}_\lambda^{\texttt{LIN}}(d_\theta)) + \mathbb{E}\left[\frac{\lambda d_\theta(a|x)c}{\pi_0(a|x) - \lambda d_\theta(a|x)c}\right]\right)\right)\right]$$

$$\leq \mathbb{E}_P\left[\prod_{i=1}^n \left(\exp\left(-\lambda(R(d_\theta) + \lambda\mathcal{S}_\lambda^{\texttt{LIN}}(d_\theta)) + \mathbb{E}\left[\frac{\lambda d_\theta(a|x)c}{\pi_0(a|x) - \lambda c}\right]\right)\right)\right] \quad (d_\theta \text{ is binary.})$$

$$\leq \mathbb{E}_P\left[\prod_{i=1}^n \left(\exp\left(-\lambda^2\mathcal{S}_\lambda^{\texttt{LIN}}(d_\theta) + \mathbb{E}\left[\frac{\lambda d_\theta(a|x)c}{\pi_0(a|x) - \lambda c} - \lambda\frac{d_\theta(a|x)c}{\pi_0(a|x)}\right]\right)\right)\right]$$

$$\leq \mathbb{E}_P\left[\prod_{i=1}^n \left(\exp\left(-\lambda^2\mathcal{S}_\lambda^{\texttt{LIN}}(d_\theta) + \lambda^2\mathcal{S}_\lambda^{\texttt{LIN}}(d_\theta)\right)\right)\right] \leq 1,$$

giving by rearranging terms, the following PAC-Bayesian bound:

$$\forall Q \in \mathcal{P}(\Theta), \quad \hat{R}_n^{\lambda-\texttt{LIN}}(\pi_Q) \leq R(\pi_Q) + \lambda\mathcal{S}_\lambda^{\texttt{LIN}}(\pi_Q) + \frac{\mathcal{KL}(Q||P) + \ln(2/\delta)}{\lambda n}.$$

Now we take a union of the the two bounds, for them to hold with probability at least $1 - \delta$ for all Q. By definition of $\hat{\pi}_{Q_n}$ (minimizer of the upper bound), we have:

$$R(\hat{\pi}_{Q_n}) \leq \hat{R}_n^{\lambda-\texttt{LIN}}(\hat{\pi}_{Q_n}) + \frac{\mathcal{KL}(Q_n||P) + \ln(2/\delta)}{\lambda n} \leq \hat{R}_n^{\lambda-\texttt{LIN}}(\pi_{Q^*}) + \frac{\mathcal{KL}(Q^*||P) + \ln(2/\delta)}{\lambda n}.$$

Using the lower bound on the risk of $R(\pi_{Q^*})$, we have:

$$R(\hat{\pi}_{Q_n}) \leq \hat{R}_n^{\lambda-\texttt{LIN}}(\pi_{Q^*}) + \frac{\mathcal{KL}(Q^*||P) + \ln(2/\delta)}{\lambda n}$$
$$\leq R(\pi_{Q^*}) + \lambda\mathcal{S}_\lambda^{\texttt{LIN}}(\pi_{Q^*}) + \frac{\mathcal{KL}(Q^*||P) + \ln(2/\delta)}{\lambda n}.$$

which gives us the PAC-Bayesian suboptimality upper bound:

$$0 \leq R(\hat{\pi}_{Q_n}) - R(\pi_{Q^*}) \leq \lambda\mathcal{S}_\lambda^{\texttt{LIN}}(\pi_{Q^*}) + \frac{2\left(\mathcal{KL}(Q^*||P) + \ln(2/\delta)\right)}{\lambda n}.$$

Concluding the proof. $\qquad\square$

H Experimental design and detailed experiments

All our experiments were conducted on a machine with 16 CPUs. The PAC-Bayesian learning experiments require a moderate amount of computation due to the handling of medium-sized datasets. However, our experiments remain reproducible with minimal computational resources.

H.1 Off-policy evaluation and selection

H.1.1 Datasets

For both our OPE and OPS experiments, we use 11 UCI datasets with different sizes, action spaces and number of features. The statistics of all these datasets are described in Table 3.

Table 3: OPE and OPS: 11 Datasets used from OpenML [8].

Datasets	N	K	p
ecoli	336	8	7
arrhythmia	452	13	279
micro-mass	571	20	1300
balance-scale	625	3	4
eating	945	7	6373
vehicle	846	4	18
yeast	1484	10	8
page-blocks	5473	5	10
optdigits	5620	10	64
satimage	6430	6	36
kropt	28 056	18	6

H.1.2 (OPE) Tightness of the bounds

Additional details. For these experiments, as we only use oracle policies (faulty policies to log data and we evaluate ideal policies), we use the full 11 datasets without splitting them. The faulty policies are defined exactly as described in the experiments of Kuzborskij et al. [32]. For each datapoint, the behavior (faulty) policy plays an action and we record a cost. The triplets datapoint, action and cost constitute our logged bandit dataset, with which we can compute our estimates and bounds. As we have access to the true label, the original dataset can be used to compute the true risk of any policy.

Detailed results. Evaluating the worst case performance of a policy is done through evaluating risk upper bounds [10, 32]. This means that a better evaluation will solely depend on the tightness of the bounds used. To this end, given a policy π, we are interested in bounds with a small relative radius $|U(\pi)/R(\pi) - 1|$. We compare our newly derived bounds (cIPS-L=1 for U_1^λ and LS for U_∞^λ both with $\lambda = 1/\sqrt{n}$) to SNIPS-ES: the Efron Stein bound for Self Normalized IPS [32], cIPS-EB: Empirical Bernstein for Clipping [55] and the recent IX: Implicit Exploration bound [21]. We use all 11 datasets, with different behavior policies ($\tau_0 \in \{0.2, 0.25, 0.3\}$) and different noise levels ($\epsilon \in \{0., 0.1, 0.2\}$) to evaluate ideal policies with different temperatures ($\tau \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$), defining ~ 500 different scenarios to validate our findings. In addition to the cumulative distribution of the relative radius of the considered bounds of Figure 2. We give two tables in the following: the average relative

Table 4: OPE: Average relative radius for each datasets

Datasets	SN-ES	cIPS-EB	IX	cIPS-L=1	LS
ecoli	1.00	1.00	<u>0.676</u>	0.752	**0.573**
arrhythmia	1.00	1.00	<u>0.677</u>	0.707	**0.548**
micro-mass	0.962	0.840	0.394	<u>0.346</u>	**0.311**
balance-scale	1.00	0.950	<u>0.469</u>	0.550	**0.422**
eating	0.930	0.734	<u>0.318</u>	0.337	**0.265**
vehicle	0.981	0.867	<u>0.409</u>	0.482	**0.358**
yeast	0.861	0.660	<u>0.307</u>	0.311	**0.254**
page-blocks	0.760	0.547	<u>0.371</u>	0.447	**0.312**
optdigits	0.468	0.323	0.148	<u>0.139</u>	**0.113**
satimage	0.506	0.336	<u>0.171</u>	0.184	**0.140**
kropt	0.224	0.161	0.087	<u>0.066</u>	**0.060**

radius of our bounds for each dataset, compiled in Table 4, and the average relative radius of our bounds for each policy evaluated, compiled in Table 5. One can observe that LS always gives the best results no matter the projection. However, the cIPS-L=1 bound is sometimes better than IX, especially when it comes to evaluating diffused policies, see Table 5.

Table 5: OPE: Average relative radiuses for each target policies (ideal policies with different τ)

τ	SN-ES	cIPS-EB	IX	cIPS-L=1	LS
$\tau = 0.1$	0.783	0.630	<u>0.332</u>	0.400	**0.308**
$\tau = 0.2$	0.781	0.630	<u>0.326</u>	0.390	**0.295**
$\tau = 0.3$	0.782	0.668	<u>0.353</u>	0.389	**0.297**
$\tau = 0.4$	0.793	0.706	<u>0.385</u>	0.385	**0.301**
$\tau = 0.5$	0.810	0.735	0.432	<u>0.397</u>	**0.323**

H.1.3 (OPS) Find the best, avoid the worst policy

Policy selection aims at identifying the best policy among a set of finite candidates. In practice, we are interested in finding policies that improve on π_0 and avoid policies that perform worse than π_0. To replicate real world scenarios, we design an experiment where π_0 is a faulty policy ($\tau = 0.2$), that collects noisy ($\epsilon = 0.2$) interaction data, some of which is used to learn $\pi_{\theta^{\text{IPS}}}, \pi_{\theta^{\text{SN}}}$, and that we add to our discrete set of policies $\Pi_{k=4} = \{\pi_0, \pi^{\text{ideal}}, \pi_{\theta^{\text{IPS}}}, \pi_{\theta^{\text{SN}}}\}$. The splits for these experiments are the following: 70% of the data is used to create bandit feedback (20% is used to train $\pi_{\theta^{\text{IPS}}}, \pi_{\theta^{\text{SN}}}$ and 50% is used to evaluate policies based on estimators/upper bounds.) the rest is used to evaluate the true value of the policies. The goal is to measure the ability of our selection strategies to choose from $\Pi_{k=4}$, better performing policies than π_0. We thus define three possible outcomes: a strategy can select *worse* performing policies, *better* performing or the *best* policy. We compare selection strategies based on upper bounds to the commonly used estimators IPS and SNIPS. The hyperparameters of all bounds (the clipping parameter M and λ) are set to $1/\sqrt{n}$. The comparison is conducted on the 11 datasets with 10 different seeds resulting in 110 scenarios. In addition to the plot in Figure 2, we collect the number of times each method selected the best policy (π_*^{S}), a better (**B**) or a worse (**W**) policy than π_0 for all datasets in Table 6. We can see that risk estimators can be unreliable, especially in small sample datasets, as they can choose worse performing policies than π_0, a catastrophic outcome in highly sensitive applications. Selecting policies based on upper bounds is more conservative, as it avoids completely poor performing policies. In addition, the tighter the bound, the better its percentage of time it selects the best policy: LS upper bound is less conservative and can find best policies more than any other bound, while never selecting poor performing policies.

Table 6: OPS: Number of times the worst, better or best policy was selected for each dataset.

Dataset	IPS			SNIPS			SN-ES			cIPS-EB			IX			cIPS-L=1			LS		
	W	B	π_*^s	W	B	π_*^s	W	B	π_*^s	W	B	π_*^s	W	B	π_*^s	W	B	π_*^s	W	B	π_*^s
ecoli	2	6	2	4	1	5	0	10	0	0	10	0	0	7	3	0	10	0	0	6	4
arrhythmia	0	10	0	0	10	0	0	10	0	0	10	0	0	7	3	0	10	0	0	5	5
micro-mass	3	0	7	1	0	9	0	10	0	0	10	0	0	0	10	0	0	10	0	0	10
balance-scale	0	3	7	0	2	8	0	10	0	0	10	0	0	4	6	0	10	0	0	3	7
eating	3	2	5	2	1	7	0	10	0	0	10	0	0	4	6	0	8	2	0	4	6
vehicle	3	0	7	1	1	8	0	10	0	0	10	0	0	5	5	0	10	0	0	3	7
yeast	0	2	8	2	0	8	0	10	0	0	10	0	0	2	8	0	7	3	0	2	8
page-blocks	0	0	10	0	0	10	0	10	0	0	10	0	0	0	10	0	10	0	0	0	10
optdigits	0	1	9	0	0	10	0	10	0	0	10	0	0	1	9	0	3	7	0	1	9
satimage	0	0	10	0	0	10	0	10	0	0	10	0	0	0	10	0	7	3	0	0	10
kropt	0	0	10	0	0	10	0	10	0	0	0	10	0	0	10	0	0	10	0	0	10

H.2 Off-policy learning

H.2.1 Datasets

As described in the experiments section, we follow exactly the experimental design of Sakhi et al. [49], Aouali et al. [5] to conduct our PAC-Bayesian Off-Policy learning experiments. We however take the time to explain it in details. In this procedure, we need three splits: D_l (of size n_l) to train the logging policy π_0, another split D_c (of size n_c) to generate the logging feedback with π_0, and finally a test split D_{test} (of size n_{test}) to compute the true risk $R(\pi)$ of any policy π. In our experiments, we split the training split D_{train} (of size N) of the four datasets considered into D_l ($n_l = 0.05N$) and D_c ($n_c = 0.95N$) and use their test split D_{test}. The detailed statistics of the different splits can be found in Table 7. Recall that K is the number of actions and p the number of features.

Table 7: OPL: Detailed statistics of the splits used.

Datasets	N	n_l	n_c	n_{test}	K	p
MNIST	60 000	3000	57 000	10 000	10	784
FashionMNIST	60 000	3000	57 000	10 000	10	784
EMNIST-b	112 800	5640	107 160	18 800	47	784
NUS-WIDE-128	161 789	8089	153 700	107 859	81	128

H.2.2 Policy class

In the PAC-Bayesian Learning paradigm, we are interested in the definition of policies as mixtures of decision rules:

$$\pi_Q(a|x) = \mathbb{E}_{f_\theta \sim Q}\left[\mathbb{1}\left[f_\theta(x) = a\right]\right], \qquad \forall(x,a) \in \mathcal{X} \times \mathcal{A}. \qquad (56)$$

We use the Linear Gaussian Policy of Sakhi et al. [49]. To obtain these policies, we restrict f_θ to:

$$\forall x \in \mathcal{X}, \quad f_\theta(x) = \underset{a' \in \mathcal{A}}{\arg\max}\left\{x^t\theta_{a'}\right\} \qquad (57)$$

This results in a parameter θ of dimension $d = p \times K$ with p the dimension of the features $\phi(x)$ and K the number of actions. We also restrict the family of distributions $\mathcal{Q}_{d+1} = \{Q_{\boldsymbol{\mu},\sigma} = \mathcal{N}(\boldsymbol{\mu}, \sigma^2 I_d), \boldsymbol{\mu} \in \mathbb{R}^d, \sigma > 0\}$ to independent Gaussians with shared scale. Estimating the propensity of a given x reduces the computation to a one dimensional integral:

$$\pi_{\boldsymbol{\mu},\sigma}(a|x) = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,1)}\left[\prod_{a' \neq a} \Phi\left(\epsilon + \frac{\phi(x)^T(\boldsymbol{\mu}_a - \boldsymbol{\mu}_{a'})}{\sigma||\phi(x)||}\right)\right]$$

with Φ the cumulative distribution function of the standard normal.

H.2.3 Detailed hyperparameters

Contrary to previous work, our method does not require tuning any loss function hyperparameter over a hold out set. We do however need to choose parameters to optimize the policies.

The logging policy π_0. π_0 is trained on D_l (supervised manner) with the following parameters: We use L_2 regularization of 10^{-4}. This is used to prevent the logging policy π_0 from being close to deterministic, allowing efficient learning with importance sampling. We use Adam [30] with a learning rate of 10^{-1} for 10 epochs.

Parameters of the bounds. cIPS and cvcIPS: The clipping parameter τ is fixed to $1/K$ with K the action size of the dataset and cvcIPS is used with $\xi = -0.5$ (the values used in Sakhi et al. [49]). ES: The exponential smoothing parameter α is fixed to $1 - 1/K$.

Optimizing the bounds. We use Adam [30] with a learning rate of 10^{-3} for 100 epochs. The gradient of **LIG** policies is a one dimensional integral, and is approximated using $S = 32$ samples.

$$
\pi_{\boldsymbol{\mu},\sigma}(a|x) = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,1)} \left[\prod_{a' \neq a} \Phi \left(\epsilon + \frac{\phi(x)^T(\boldsymbol{\mu}_a - \boldsymbol{\mu}_{a'})}{\sigma ||\phi(x)||} \right) \right]
$$

$$
\approx \frac{1}{S} \sum_{s=1}^{S} \prod_{a' \neq a} \Phi \left(\epsilon_s + \frac{\phi(x)^T(\boldsymbol{\mu}_a - \boldsymbol{\mu}_{a'})}{\sigma ||\phi(x)||} \right) \quad \epsilon_1, ..., \epsilon_S \sim \mathcal{N}(0,1).
$$

For all bounds, instead of fixing λ, we take a union bound over a discretized space of possible parameters Λ of size $n_\Lambda = 100$ and for each iteration j of the optimization procedure, we take $\lambda_j \in \Lambda$ that minimizes the estimated bound and proceed to compute the gradient w.r.t μ and σ with λ_j.

H.2.4 Detailed results

In addition to the results of Table 2, we also provide a more detailed view of the results here. For each α and dataset, we average both $\{\mathcal{GR}, R\}$ over the 10 seeds and plot them in Figure 6 and Figure 5. Note that the error bars are too small $\sigma/\sqrt{10} \approx 0.001$ and all our results in these graphs are significant. We observe that the LS PAC-Bayesian bound improves substantially on its competitors in terms of the guaranteed risk, especially on MNIST and FashionMNIST and also obtains the best performing policies, on par with the `IX` bound in the majority of scenarios.
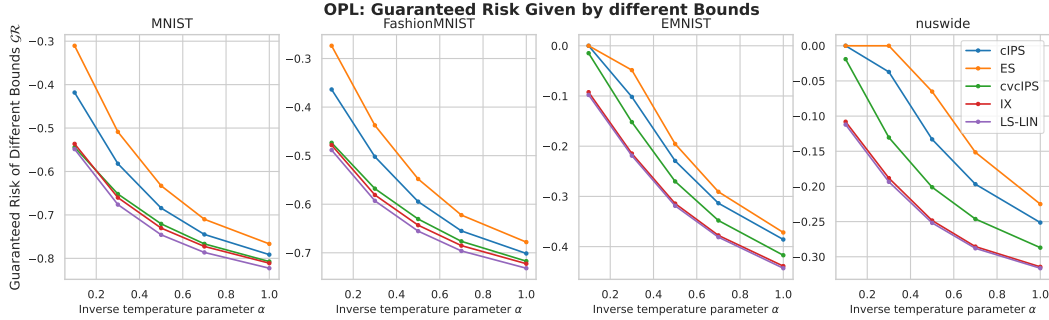

Figure 5: OPL: Guaranteed Risk given by the different bounds. We observe that our `LS-LIN` dominates all other bounds. `IX` comes close, especially on EMNIST and nuswide
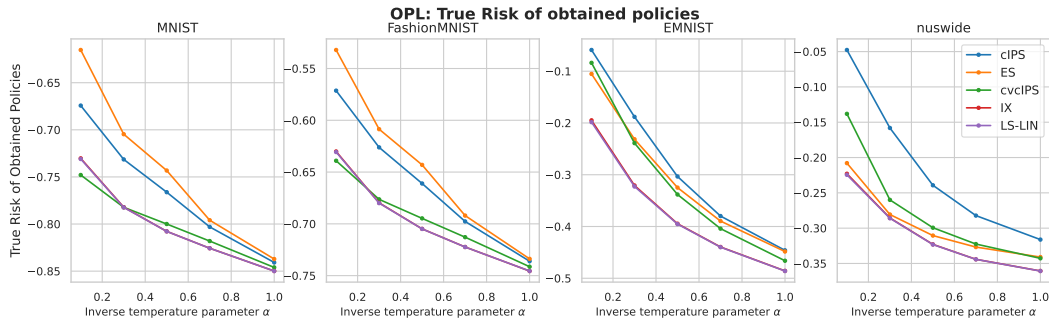

Figure 6: OPL: True risk of obtained policies after minimizing the PAC-Bayesian bounds. We observe that `LS-LIN` and `IX` are hardly distinguishable, they both give the best policies in the majority of scenarios.

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: The abstract and introduction scopes our work in the offline contextual bandit setting, describes our method and claims its superiority compared to existing work. We provide both strong theoretical and empirical evidence in the paper to defend the claim.

   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: We provide the limitations of our method in Appendix A and discuss ways to mitigate them in future work.

   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 our results are proven in the paper and all assumptions are discussed.

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: Our paper follows the classical off-policy experimental design, and all details to reproduce them are given in Appendix H.

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: All data used is accessible UCI Repository, the code is also given in the supplementary material to reproduce all experimental results.

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 experimental settings are detailed in Section 5 and Appendix H.

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 our experiments are run with multiple seeds and for different scenarios and datasets. Some graphs do not need error bars (cumulative distributions or selection strategies), for the other results, we have very small error bars (Appendix H.2), making our results significant.

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: Our experiments can be conducted in small machines and do not require heavy compute, this is detailed in Appendix H.

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: Our paper is of theoretical nature, presents ideas to increase safety in decision making, uses publicly available data for the experiments and conforms to 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 paper discusses both the positive and negative impacts in Appendix B.

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: Our paper tackles theoretical questions for decision-making, the data used for the experiments is openly accessible in UCI repository. We do not believe that our work poses such risks.

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: Our experimental design is inspired from the code base of some papers that we cite, and all data used is openly accessible in UCI repository.

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: [NA]

Justification: We do not release new assests, but we give the code to reproduce the experiments in the supplementary material.

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: Our 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: Our 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.