# Importance-Weighted Offline Learning Done Right

**Germano Gabbianelli** GERMANO.GABBIANELLI@UPF.EDU
*Universitat Pompeu Fabra,Barcelona, Spain*

**Gergely Neu** GERGELY.NEU@GMAIL.COM
*Universitat Pompeu Fabra,Barcelona, Spain*

**Matteo Papini** MATTEO.PAPINI@POLIMI.IT
*Politecnico di Milano,Milan, Italy*

**Editors:** Claire Vernade and Daniel Hsu

## Abstract

We study the problem of offline policy optimization in stochastic contextual bandit problems, where the goal is to learn a near-optimal policy based on a dataset of decision data collected by a sub-optimal behavior policy. Rather than making any structural assumptions on the reward function, we assume access to a given policy class and aim to compete with the best comparator policy within this class. In this setting, a standard approach is to compute importance-weighted estimators of the value of each policy, and select a policy that minimizes the estimated value up to a "pessimistic" adjustment subtracted from the estimates to reduce their random fluctuations. In this paper, we show that a simple alternative approach based on the "implicit exploration" estimator of Neu (2015) yields performance guarantees that are superior in nearly all possible terms to all previous results. Most notably, we remove an extremely restrictive "uniform coverage" assumption made in all previous works. These improvements are made possible by the observation that the upper and lower tails importance-weighted estimators behave very differently from each other, and their careful control can massively improve on previous results that were all based on symmetric two-sided concentration inequalities. We also extend our results to infinite policy classes in a PAC-Bayesian fashion, and showcase the robustness of our algorithm to the choice of hyper-parameters by means of numerical simulations.

## 1. Introduction

Offline Policy Optimization (OPO) is the problem of learning a near-optimal policy based on a dataset of historical observations. This problem is of outstanding importance in real-world applications where experimenting directly with the environment is costly, but otherwise large volumes of offline data is available to learn from. Such settings include problems in healthcare (Murphy, 2003; Kim et al., 2011; Bertsimas et al., 2017; Rehg et al., 2017), advertising (Bottou et al., 2013; Farias and Li, 2019), or recommender systems (Li et al., 2011; Schnabel et al., 2016).

A popular approach for this setting is *importance-weighted offline learning*, where one optimizes an unbiased estimate of the expected reward, obtained through an appropriately reweighted average of the rewards in the dataset (Li et al., 2011; Bottou et al., 2013). To deal with unstable nature of these estimators, the influential work of Swaminathan and Joachims (2015) proposed an approach called "counterfactual risk minimization", which consists of adding a regularization term to the

optimization problem to down the fluctuations, thus preventing the optimizer to overfit to random noise. Their work has inspired a number of follow-ups that either refined the regularization terms to yield better theoretical guarantees (Jin et al., 2022; Wang et al., 2023), or developed practical methods with improved empirical performance in large-scale problems London and Sandler (2019); Sakhi et al. (2023). In this paper, we contribute to this line of work by studying a simple and robust variant of the standard importance-weighted reward estimators used in past work, and showing tight theoretical performance guarantees for it.

Our main contribution is showing that the so-called *implicit exploration* (IX) estimator (originally proposed by Kocák et al., 2014 and Neu, 2015 in the context of online learning) achieves a massive variance-reducing effect in our offline learning setting, and using this observation to derive performance guarantees that are both significantly tighter and easier to interpret than all previous results in the literature. In particular, we formally show that the regularization effect built into the IX estimator is strong enough so that no further regularizer is required to stabilize the performance of policy optimization. This result is perhaps surprising for the reader familiar with past work on the subject, especially since several of these works made use of IX-like variance reduced estimators without managing to drop the additional regularization. The key observation that allows us to prove our main results is that the tails of importance-weighted estimators are *asymmetric*, which allows us to tightly control the two tails separately via specialized concentration inequalities. This is to be contrasted with previous results that all rely on symmetric confidence intervals that turn out to be needlessly conservative. This new perspective not only allows us to obtain better results but also to simplify the analysis: both of the concentration inequalities we use for the two tails can be derived using elementary techniques in a matter of a few lines[1].

More concretely, our main result is a regret bound that scales with the degree of "overlap" between the comparator policy and the behavior policy, demonstrating better scaling against policies that are covered better by the observed data. Unlike virtually all previous work, our guarantees do not require the unrealistic condition that action-sampling probabilities be bounded away from zero for all contexts. Our algorithm can be implemented efficiently using a single call to a cost-sensitive classification oracle, thus effectively reducing the offline policy optimization problem to a standard supervised learning task (which feature is in high regard thanks to the influential works of Langford and Zhang, 2007; Dudík et al., 2011; Agarwal et al., 2014 in the broader area of contextual bandit learning). For simplicity of exposition, we prove our main result for finite policy classes and show that the regret scales logarithmically with the size of the class. We also provide some extensions to the simple algorithm achieving these results, namely a version that trades oracle-efficiency for a better scaling with the quantity measuring the mismatch between the target and behavior policies, and a "PAC-Bayesian" variant that can make use of prior information on the problem and also works for infinite policy classes. This extends the recent works of London and Sandler (2019); Sakhi et al. (2023); Flynn et al. (2023) by providing better generalization bounds and introducing a new family of PAC-Bayesian regret bounds that apparently have not existed so far in the literature. We also illustrate our theoretical findings with a set of experiments conducted on real data, and empirically verify the robustness of our method as compared to some natural baselines.

1. In fact, both results are readily available in the literature: one is the main result of Neu (2015) regarding the upper tail of the IX estimator, and another is stated as an exercise in Boucheron et al. (2013).

It is worth mentioning a parallel line of work on contextual bandits that starts from the assumption that the reward function belongs to a known function class, and thus a near-optimal policy can be learned by identifying the true reward function within the class up to sufficient accuracy. This perspective has been adopted by Jin et al. (2021) (as well as a sequence of follow-up works on offline reinforcement learning) who considered function classes that are linear in some low-dimensional features of the context-action pairs. These works provide simple algorithms with strong theoretical performance guarantees, but they are all limited by the strong assumptions that need to be made about the reward function (and it is unclear how sensitive they are to model misspecification). In contrast, the setting we consider assumes access to a policy class and allows the development of algorithms that perform nearly as well as the best policy within the class *without* requiring that the rewards have a simple parametric form. This setting comes with its own set of trade-offs: the statistical complexity of learning in this setting depends on the complexity of the policy class, and hard problems will evidently require large classes of policies to accommodate best-in-class policies with satisfying performance. Our results in this paper highlight some further open questions in this setting regarding computational-statistical trade-offs—the discussion of which we relegate to Section 7.

## 2. Preliminaries

We study the problem of offline learning in stochastic contextual bandits. A contextual bandit problem instance is defined by the tuple $(\mathcal{X}, \mathcal{A}, \nu, p)$, where $\mathcal{X}$ is a set of contexts, $\mathcal{A}$ is a set of actions with finite cardinality $K$, $\nu$ is an unknown distribution over the context space representing the probability of encountering each context, and $p : \mathcal{X} \times \mathcal{A} \to \mathbb{R}_{[0,1]}$ is a probability kernel mapping context-action pairs to rewards in the interval $[0, 1]$. The mean reward associated with context-action pair $x, a$ is denoted as $r(x, a)$.

A policy $\pi : \mathcal{X} \to \Delta_{\mathcal{A}}$ is defined as mapping from contexts to distributions over actions, with $\pi(a|x)$ denoting the probability of selecting action $a \in \mathcal{A}$ in context $x \in \mathcal{X}$ when following policy $\pi$. The expected reward of a policy $\pi$ is called *value* and is defined as $v(\pi) = \mathbb{E}\left[\sum_a \pi(a|X) r(X, a)\right]$. We are given access to a dataset $\mathcal{D} = (X_t, A_t, R_t)_{t=1}^n$, sampled by a fixed behavior policy $\mu : \mathcal{X} \to \Delta_{\mathcal{A}}$ according to the following protocol:

- $X_t$ is a context drawn i.i.d. from the unknown distribution $\nu$,
- $A_t$ is an action drawn from the behavior policy $\mu(\cdot|X_t)$ independently from all random variables other than $X_t$,
- $R_t$ is a random reward drawn from $p(\cdot|X_t, A_t)$, assumed to lie almost surely in $[0, 1]$, with mean given by the reward function $\mathbb{E}\left[R_t | X_t, A_t\right] = r(X_t, A_t)$.

For simplicity, we suppose that the behavior policy $\mu$ is fixed and known, and only note here that extension to adaptive behavior policies is straightforward.

The goal is to use the available data to produce a policy $\widetilde{\pi}_n$ achieving the highest possible expected reward. The performance will be measured in terms of *regret* (or *excess risk*) with respect to a comparator policy $\pi^*$:

$$\mathfrak{R}_n(\pi^*) = v(\pi^*) - v(\widetilde{\pi}_n).$$

We assume to have access to a policy class $\Pi \subseteq \{\pi : \mathcal{X} \to \Delta_{\mathcal{A}}\}$ and aim to provide regret bounds against all policies within the class. For most of our contributions, we will work with finite policy

classes and assume access to a computational oracle that can return optimal policies given an appropriately defined input dataset, Precisely, the oracle takes as input a dataset $\{x_t, g_t\}_{t=1}^n$ with contexts $x_t \in \mathcal{X}$ and gains $g_t \in \mathbb{R}^{\mathcal{A}}$, and returns

$$\mathrm{CSC}\big(\{x_t, g_t\}_{t=1}^n\big) = \operatorname*{argmax}_{\pi \in \Pi} \sum_{t=1}^n \sum_a \pi(a|x_t) g_t(a).$$

This definition is slightly more general than the one used in previous works such as Dudík et al. (2011), in that it allows $\Pi$ to include stochastic policies. The optimization problem solved by the oracle is easily seen to be equivalent to the task of *cost-sensitive classification* (CSC), when considering the rewards of each action as negative costs associated with misclassification errors. We will thus occasionally refer to the oracle as a *CSC oracle* (which also explains the notation used above). We are interested in developing algorithms that access the oracle a small constant number of times while providing formal performance guarantees on the quality of the output policy.

## 3. Pessimistic importance-weighted offline learning in contextual bandits

A natural approach for the offline learning setting we consider is to define an estimator $\widehat{v}_n(\pi)$ for the value function $v(\pi)$ of each policy $\pi$, and to return the policy $\widehat{\pi}_n \in \Pi$ which maximizes it. The simplest possible estimator one can think of is the *importance-weighted* (IW) value estimator (Horvitz and Thompson, 1952) defined for each policy $\pi$ as

$$\widehat{v}_n(\pi) = \frac{1}{n} \sum_{t=1}^n \frac{\pi(A_t|X_t)}{\mu(A_t|X_t)} \cdot R_t. \tag{1}$$

This estimator is unbiased in the sense that for any policy $\pi$ we have $\mathbb{E}[\widehat{v}_n(\pi)] = v(\pi)$, and the output policy $\widehat{\pi}_n$ can be computed with a single call to the computational oracle, by setting the reward vectors as $g_t(a) = \mathbb{1}\{A_t = a\} R_t / \mu(A_t|X_t)$. However, this estimator is notoriously heavy-tailed, and thus, with significant probability, may be very distant from its true value (Ionides, 2008). Specifically, actions that are sampled with very low probability by the behavior policy may falsely appear to yield huge rewards. This is especially detrimental in offline learning where fitting a policy to such sampling artifacts can lead to poor performance during deployment.

In recent years, a range of ideas have been proposed to tame the adverse behavior of the IW estimator. The most widely adopted approach, first proposed by Swaminathan and Joachims (2015) (and later elaborated on in a variety of contexts by works like London and Sandler, 2019; Jin et al., 2021; Rashidinejad et al., 2021; Li et al., 2022; Jin et al., 2022), involves implementing a "pessimistic" adjustment to reduce random fluctuations. Concretely, this method involves calculating an adjustment term $B_n(\pi)$ which, when subtracted from the IW estimator $\widehat{v}_n(\pi)$, ensures that the result is always smaller than the true value $v(\pi)$ for any policy $\pi$. Subsequently, the best pessimistic policy in $\Pi$ is identified and returned, by maximizing the expression $\widehat{v}_n(\pi) - B_n(\pi)$. The adjustment $B_n(\pi)$ is typically computed using standard concentration inequalities like Bernstein's inequality (see, e.g., Sections 2.7 and 2.8 in Boucheron et al., 2013), and generally tends to grow larger as the policy $\pi$ deviates further from the behavior policy. It is then easy to show that the regret of this method with respect to any comparator $\pi^*$ can be bounded by $2B_n(\pi^*)$, via a straightforward calculation which we reproduce in the proof of our Theorem 1.

This generic recipe for offline learning has been combined with the IW estimator defined above by Swaminathan and Joachims (2015), Jin et al. (2022) and Wang et al. (2023). This "pessimistic importance-weighted offline learning" approach, which we abbreviate as *PIWO learning*, has several downsides, depending on the choice of $B_n(\pi)$. First, as pointed out recently by Wang et al., 2023, $\widehat{v}_n - B_n$ may not be necessarily be of the form required by a practical optimization oracle. Even more concerningly, a conservatively chosen adjustment $B_n$ may not only result in loose theoretical guarantees, but also poor empirical performance. Indeed, notice that setting $B_n$ too large may overwhelm the data-dependent value estimates, thus resulting in a policy that effectively ignores the observed data from policies that are relatively poorly covered. In extreme cases, this approach may even favor policies that have never been observed to yield any reward whatsoever over policies with positive estimated reward but high estimated uncertainty.

The cleanest results for this PIWO learning approach have been derived by Wang et al. (2023), who used the adjustment $B_n(\pi) = \beta \sum_{t=1}^{n} \sum_a \frac{\pi(a|X_t)}{\mu(a|X_t)}$. Their regret bounds are stated in terms of the following quantity that measures the "overlap" between a given policy $\pi$ and the behavior policy $\mu$:

$$C(\pi) = \mathbb{E}\left[\sum_a \frac{\pi(a|X)}{\mu(a|X)}\right]. \tag{2}$$

We will refer to $C(\pi)$ as the *policy coverage ratio* between $\pi$ and $\mu$. The coverage ratio can be seen as a notion of similarity between $\pi$ and $\mu$: it is minimized when the two policies are equal, and otherwise grows to infinity as the two policies drift apart. Assuming that the likelihood ratio between the two policies is uniformly upper-bounded as $\sup_{x,a} \frac{\pi(a|x)}{\mu(a|x)} \leq \frac{1}{\alpha}$, Wang et al. (2023) obtain, for their oracle-efficient algorithm, a regret bound of the form

$$\mathfrak{R}_n(\pi^*) = \mathcal{O}\left(C(\pi^*)\sqrt{\frac{\log\left(|\Pi|/\delta\right)}{n}} + \frac{\log(|\Pi|/\delta)}{\alpha n}\right). \tag{3}$$

This bound has the appealing property that its leading term scales as $C(\pi^*)/\sqrt{n}$, thus guaranteeing good performance when the comparator policy $\pi^*$ is well-covered by the behavior policy. The bound can be improved to scale with $\sqrt{C(\pi^*)}$ instead of $C(\pi^*)$ if one has prior knowledge of the coverage ratio against the target policy $\pi^*$. On the negative side, the result effectively requires the strong *uniform coverage* condition $\inf_{x,a} \mu(a|x) \geq \alpha$ which ensures that all actions are sampled at least a constant $\alpha$ fraction of times in the data set. This condition is typically not met in realistic applications for reasonable values of $\alpha$, and in particular the bound becomes completely void of meaning if there exists one single context $x$ where some action $a$ is selected with zero probability.

The original algorithm by Swaminathan and Joachims (2015) suffered from the same issue. Recently, Jin et al. (2022) were able to relax this uniform-coverage condition by developing a sophisticated concentration inequality that only requires the third moment of the importance weights $\sum_a \frac{\pi^*(a|X_t)}{\mu(a|X_t)}$ to be bounded. Eventually, their bounds only apply to deterministic policies that map each context $x$ to a single action $\pi^*(x)$, and depend on the quantity $\alpha^* = \inf_x \mu(\pi^*(x)|x)$. Their most clearly stated result is Corollary 4.3, where they effectively show

$$\mathfrak{R}_n(\pi^*) = \mathcal{O}\left(\sqrt{\frac{\log\left(|\Pi|T\right)}{\alpha^* n}} \cdot \left(\log\left(\frac{1}{\delta}\right)\right)^{3/2}\right).$$

This bound still remains vacuous if there is one single context where $\mu(\pi^*(x)|x)$ is zero. A further downside of their method pointed out by Wang et al. (2023) is that the proposed algorithm is not directly implementable with a CSC oracle due to the form of the adjustment $B_n$ they use. In the following section, we will develop an algorithm that eliminates all these limitations.

## 4. Pessimism and Variance Reduction via Implicit Exploration

Our main contribution is addressing the limitations of the PIWO learning framework in the previous section by studying a very simple adjustment to the standard IW estimator. Concretely, we adapt the so-called "Implicit eXploration" (IX) estimator of Neu (2015) defined as

$$\widetilde{v}_n(\pi) = \frac{1}{n}\sum_{t=1}^{n} \frac{\pi(A_t|X_t)}{\mu(A_t|X_t)+\gamma} \cdot R_t, \tag{4}$$

where $\gamma \geq 0$ is a hyperparameter of the estimator that we will sometimes refer to as the "IX parameter". This adjustment implicitly acts like mixing the behavior policy with a uniform exploration policy, thus reducing the random fluctuations of the IW estimator (and justifying the name "implicit exploration"). The price of this stabilization effect is that the estimates are biased towards zero to an extent that can be controlled using the IX parameter $\gamma$. Indeed, as a simple calculation shows, the IX estimator satisfies $\mathbb{E}\left[\widetilde{v}_n(\pi)\right] = V(\pi) - \gamma C_\gamma(\pi)$, with the bias term $C_\gamma(\pi)$ given as

$$C_\gamma(\pi) = \mathbb{E}\left[\sum_a \frac{\pi(a|X)}{\mu(a|X)+\gamma} \cdot r(X,a)\right]. \tag{5}$$

Since the rewards are assumed to be non-negative, this bias can be interpreted as a *pessimistic* adjustment to an otherwise unbiased estimator, and it is thus reasonable to expect it to have the same effect as the adjustments used in the general PIWO framework[2].

Note that $C_\gamma(\pi)$ is closely related to the policy coverage ratio $C(\pi)$ as defined in Equation (2), up to the two differences that *i)* it replaces $\mu(X,a)$ by $\mu(X,a)+\gamma$ in the denominator and *(ii)* it is scaled with the rewards $r(X,a)$. Both of these adjustments make it strictly smaller than $C(\pi)$ as long as $\gamma > 0$, and notably it always remains bounded as $C_\gamma(\pi) \leq \frac{1}{\gamma}$, no matter how small $\mu(a|x)$ gets. Furthermore, due to the scaling with the rewards, $C_\gamma(\pi)$ is small for policies with low expected reward, and in particular it equals zero for a policy with zero expected reward. In what follows, we will refer to $C_\gamma$ as the *smoothed coverage ratio*.[3]

Our algorithm consists of simply selecting the policy that maximizes the IX value estimates:

$$\widehat{\pi}_n = \arg\max_{\pi\in\Pi} \widetilde{v}_n(\pi).$$

We refer to this algorithm as PIWO-IX, standing for "Pessimistic Importance-Weighted Offline learning with Implicit eXploration". Note that PIWO-IX can be implemented via a single call to the CSC oracle with the gain vectors defined as $g_t(a) = \mathbb{1}\{A_t = a\}R_t/(\mu(A_t|X_t)+\gamma)$. The following theorem states our main result regarding PIWO-IX.

2. In fact, the pessimistic bias of the IX estimators has been recently pointed out and utilized by Gabbianelli et al. (2023) in the vaguely related context of online learning with off-policy feedback.
3. We use this term in the sense of the Laplace smoothing of estimators, not to be confused with the smoothed analysis of algorithms (Spielman and Teng, 2001) applied to contextual bandits by Krishnamurthy et al. (2019).

**Theorem 1** *With probability at least* $1-\delta$*, the regret of* PIWO-IX *against any comparator policy* $\pi^* \in \Pi$ *satisfies*

$$\mathfrak{R}_n(\pi^*) \leq \frac{\log\left(2|\Pi|/\delta\right)}{\gamma n} + 2\gamma C_\gamma(\pi^*).$$

*Furthermore, by setting* $\gamma$ *to* $\sqrt{\frac{\log(2|\Pi|/\delta)}{n}}$*, the bound becomes*

$$\mathfrak{R}_n(\pi^*) \leq \left(2C_\gamma(\pi^*)+1\right)\sqrt{\frac{\log\left(2|\Pi|/\delta\right)}{n}}.$$

The bound improves on the results of Wang et al. (2023) stated as Equation (1) along several dimensions. Most importantly, our result removes the need for the behavior policy to be bounded away from zero, and as such completely does away with the uniform coverage assumptions needed by all previous work on the topic. Another improvement is that our bound tightens the dependence on the coverage ratio from $C(\pi^*)$ to the potentially much smaller $C_\gamma(\pi^*)$. A small practical improvement is that PIWO-IX calls the CSC oracle with a sparse input vector which can be computed slightly more efficiently than the dense inputs used by Wang et al. (2023). This sparsity also leads to the practical advantage that PIWO-IX does not output policies that have never been observed to yield nonzero rewards (as long as there are alternatives that do receive positive rewards). We provide further comments on the tightness of the bound above and other properties of PIWO-IX in Section 7.

The key idea behind the proof of Theorem 1 is noticing that the tails of the IX estimator are asymmetric: since $\widetilde{v}_n$ is a nonnegative random variable, its only extreme values are all going to be positive. More formally, this means that its lower tail will always be lighter than its upper tail, and thus a tight analysis needs to handle the two tails using different tools. Below, we state two lemmas that separately characterize the lower and upper tails of the IX estimator (4). The first of these bounds the upper tail along the lines of Lemma 1 (and Corollary 1) of Neu (2015):

**Lemma 2** *With probability at least* $1-\delta$*, the following holds simultaneously for all* $\pi \in \Pi$*:*

$$\widetilde{v}_n(\pi) - v(\pi) \leq \frac{\log\left(|\Pi|/\delta\right)}{2\gamma n}.$$

The proof is provided in Appendix A for completeness, but is otherwise lifted entirely from Neu (2015). The second lemma provides control of the lower tail of $\widetilde{v}_n$:

**Lemma 3** *With probability at least* $1-\delta$*, the following holds simultaneously for all* $\pi \in \Pi$*:*

$$v(\pi) - \widetilde{v}_n(\pi) \leq \frac{\log(|\Pi|/\delta)}{2\gamma n} + 2\gamma C_\gamma(\pi).$$

The proof follows from the observation that, since the rewards are non-negative, $\widetilde{v}_n$ is a non-negative random variable, and as such its lower tail is well-controlled by its second moment (see, e.g., Exercise 2.9 in Boucheron et al., 2013). The full proof is included in Section A for completeness. With the above two lemmas, we can easily prove our main theorem.

**Proof of Theorem 1** The statement follows from combining the two lemmas via a union bound, and exploiting the definition of the algorithm:

$$v(\widehat{\pi}_n) \geq \widetilde{v}_n(\widehat{\pi}_n) - \frac{\log\left(2|\Pi|/\delta\right)}{2\gamma n} \geq \widetilde{v}_n(\pi^*) - \frac{\log\left(2|\Pi|/\delta\right)}{2\gamma n} \geq v(\pi^*) - \frac{\log\left(2|\Pi|/\delta\right)}{\gamma n} - 2\gamma C_\gamma(\pi^*).$$

Concretely, the first of these inequalities follows from Lemma 2, the second one from the definition of the algorithm, and the third one from Lemma 3. This concludes the proof. ■

## 5. A PAC-Bayesian extension

Our previously stated results require the policy class $\Pi$ to be finite, and scale with $\log|\Pi|$. While this is a common assumption in past work on the subject (e.g., in Dudík et al., 2011; Agarwal et al., 2014; Wang et al., 2023), it is of course not satisfied in most practical scenarios of interest. Several extensions have been proposed in previous work, mostly based on the idea of replacing the union bound over policies by more sophisticated uniform-convergence arguments: for instance, Swaminathan and Joachims (2015) and Jin et al. (2022) respectively show bounds that depend on the covering number and the Natarajan dimension of the policy class. In this section, we provide an extension that makes use of so-called *PAC-Bayesian* generalization bounds (McAllester, 1998; Audibert, 2004; Catoni, 2007) that hold for arbitrary policy classes and often lead to meaningful performance guarantees even in large-scale settings of practical interest. We refer to the recent monograph of Alquier (2021) for a gentle introduction into the subject.

Before providing this extension, we will require some additional definitions. In this section, we will consider *randomized* algorithms that output a distribution $\widehat{Q}_n \in \Delta_\Pi$ over policies, and we will be interested in the performance guarantees that hold on expectation with respect to the random choice of $\widehat{\pi}_n \sim \widehat{Q}_n$, but still hold with high probablity with respect to the realization of the random data set. We overload our notation slightly by defining $v(Q) = \int v(\pi)\mathrm{d}Q(\pi)$, $\widetilde{v}_n(Q) = \int \widetilde{v}_n(\pi)\mathrm{d}Q(\pi)$, $C_\gamma(Q) = \int C_\gamma(\pi)\mathrm{d}Q(\pi)$, and $\mathfrak{R}_n(Q) = \int \mathfrak{R}_n(\pi)\mathrm{d}Q(\pi)$, which all capture relevant quantities evaluated on expectation under the distribution $Q \in \Delta_\Pi$.

In the context of offline learning, several works have applied PAC-Bayesian techniques to provide concentration bounds for the importance-weighted estimator $\widehat{v}_n(Q)$, characterizing its deviations from its true mean $v(Q)$ uniformly for all "posteriors" $Q$—we refer to the recent work of Sakhi et al. (2023) and the survey of Flynn et al. (2023) for an extensive overview of such results. One common feature of these works is that they all provide concentration bounds derived from PAC-Bayesian versions of standard bounds like Hoeffding's or Bernstein's inequality, and as such suffer from the same limitations as the results described in Section 3. The biggest such limitation is that all bounds require a uniform coverage assumption $\inf_{x,a}\mu(a|x) \geq \alpha$, or work with biased estimates of $v(Q)$ without quantifying the effect of the bias on the learning performance. Instead of deriving regret bounds from the concentration bounds, the focus in these works is to derive implementable algorithms from the concentration bounds and test them extensively in large-scale settings.

Here, we provide a natural extension of PIWO-IX that is derived from PAC-Bayesian principles. For defining our algorithm, we let $P \in \Delta_\Pi$ be an arbitrary "prior" over the policy class $\Pi$ and define

the output distribution as

$$\widehat{Q}_n = \arg\max_{Q \in \Delta_\pi} \left\{ \widetilde{v}_n(Q) - \frac{\mathrm{KL}\left(Q \| P\right)}{\lambda} \right\},$$

where $\mathrm{KL}\left(Q\|P\right) = \int \log \frac{\mathrm{d}Q}{\mathrm{d}P} \mathrm{d}Q$ is the *Kullback–Leibler divergence* (or *relative entropy*) between the distributions $Q$ and $P$, and $\lambda > 0$ is a regularization parameter. It is well known that this distribution (often called the *Gibbs posterior*) has a closed-form expression with $\frac{\mathrm{d}\widehat{Q}_n}{\mathrm{d}P}(\pi) = \frac{e^{\lambda \widetilde{v}_n(\pi)}}{\int e^{\lambda \widetilde{v}_n(\pi')} \mathrm{d}P(\pi')}$. For practical purposes, we will simply choose $\lambda = 2\gamma n$ below. The following theorem establishes a regret guarantee for the resulting algorithm that we call *PAC-Bayesian* PIWO-IX.

**Theorem 4** *With probability at least $1-\delta$, the regret of PAC-Bayesian* PIWO-IX *against any distribution $Q^* \in \Delta_\Pi$ over comparator policies satisfies*

$$\mathfrak{R}_n(Q^*) \leq \frac{\mathrm{KL}\left(Q^* \| P\right) + \log(1/\delta)}{\gamma n} + 2\gamma C_\gamma(Q^*).$$

*Furthermore, by setting $\gamma = \sqrt{1/n}$, the bound becomes*

$$\mathfrak{R}_n(Q^*) \leq \frac{2C_\gamma(Q^*) + \mathrm{KL}\left(Q^* \| P\right) + \log(1/\delta)}{\sqrt{n}}.$$

This bound inherits the key strength of PAC-Bayesian generalization bounds: it holds *uniformly for all competitors $Q^*$* without requiring a union bound over policies. We warn the reader familiar with PAC-Bayesian bounds though that the role of $Q^*$ here is different from what they may expect: instead of being a data-dependent "posterior", it is a "comparator" distribution that the learner wishes to compete with. Thus, the bound expresses that distributions $Q^*$ that are closer to the "prior" $P$ in terms of relative entropy are "easier" to compete with. As before, the bound scales with the smoothed policy coverage ratio $C_\gamma(Q^*)$, only this time associated with the comparator distribution $Q^*$. Just like in the bound of Theorem 1, the bound requires no uniform coverage condition, and in particular continues to hold even if $\inf_{x,a} \mu(a|x)$ approaches zero. To our knowledge, this is the first regret bound for offline learning of such a PAC-Bayesian flavor, and in any case the first PAC-Bayesian bound for this setting that does not require uniform coverage.

The proof of Theorem 4 relies on the following generalizations of Lemmas 2 and 3:

**Lemma 5** *With probability at least $1-\delta$, the following holds simultaneously for all $Q \in \Delta_\Pi$:*

$$\widetilde{v}_n(Q) - v(Q) \leq \frac{\mathrm{KL}\left(Q\|P\right) + \log\left(1/\delta\right)}{2\gamma n}.$$

**Lemma 6** *With probability at least $1-\delta$, the following holds simultaneously for all $Q \in \Delta_\Pi$:*

$$v(Q) - \widetilde{v}_n(Q) \leq \frac{\mathrm{KL}\left(Q\|P\right) + \log\left(1/\delta\right)}{2\gamma n} + 2\gamma C_\gamma(Q).$$

The statements follow from combining the proofs of Lemmas 2 and 3 with a so-called "change-of-measure" trick commonly used in the PAC-Bayesian literature. We relegate the proofs to Appendix A.3 and only provide the very simple proof of Theorem 4 here.

**Proof of Theorem 4** The statement follows from combining the above two lemmas via a union bound, and exploiting the definition of the algorithm:

$$v(\widehat{Q}_n) \geq \widetilde{v}_n(\widehat{Q}_n) - \frac{\mathrm{KL}\big(\widehat{Q}_n \| P\big) + \log(1/\delta)}{2\gamma n} \geq \widetilde{v}_n(Q^*) - \frac{\mathrm{KL}\left(Q^* \| P\right) + \log(1/\delta)}{2\gamma n}$$
$$\geq v(Q^*) - \frac{\mathrm{KL}\left(Q^* \| P\right) + \log(1/\delta)}{\gamma n} - 2\gamma C_\gamma(Q^*).$$

Concretely, the first of these inequalities follows from Lemma 5, the second one from the definition of the algorithm, and the third one from Lemma 6. This concludes the proof. ■

## 6. Experiments

In this section we provide a set of simple experiments that illustrate our theoretical findings, and in particular to empirically validate the robustness of our algorithm to hyper-parameter selection. We compare our method (PIWO-IX) to the method of Wang et al. (2023) (here referred to as PIWO-PL), and follow an experimental setup that is directly inspired by theirs. Besides PIWO-PL, we also include a commonly used variant of our algorithm that uses the *clipped importance weights* (CIW) estimator defined as $\widehat{v}_n(\pi) = \frac{1}{n}\sum_{t=1}^{n} \frac{\pi(A_t|X_t)}{\max\{\mu(A_t|X_t),\gamma\}} \cdot R_t$. We refer to this metod as PIWO-CLIP.

We use the Letter (OpenML ID 247[4]) classification dataset to simulate an offline contextual bandit instance. The dataset contains one milion entries, each consisting of 16 features and a true label, representing one of the $K = 26$ letters of the alphabet. To simulate a contextual bandit instance, we consider the feature vectors as contexts and the true labels as the corresponding optimal actions. To simulate the rewards we build a reward matrix $M \in \mathbb{R}^{K\times K}$ with entries on the diagonal set to 1 and the rest of them uniformly sampled from the $[0, 1)$ interval, and we keep these random parameters fixed for all repetitions. We then set the reward distribution $p(\cdot|x, a)$ for each context-action pair $(x, a)$ as a Bernoulli distribution with parameter $M_{a,a*(x)}$, where $a^*(x)$ denotes the optimal action associated with context $x$.

The cost-sensitive classification oracle is implemented by fitting a multi-variate ridge regressor, with one target for each action[5]. Given any context $x$, the regressor can be queried to predict the reward for each arm, and a *max* or *softmax* can be used to construct a policy to select the best arm. In order to generate a range of behavior policies, we retain 10% of the data to train an estimator of the reward for each arm using the regressor described above with the true mean rewards as labels. We then use the predicted rewards to construct 20 softmax behavior policies, by varying the inverse temperature parameter as `logspace(-1, 3, 20)`.

We then collect an offline dataset using each of the behavior policies and train our method PIWO-IX, its variation PIWO-CLIP, and the algorithm of Wang et al. (2023), using the CSC oracle described above with an argmax to select the optimal action, and varying their hyper-parameter over a wide range (i.e. `logspace(-10, 0, 20)`). Finally we compute the expected reward for each combination of behavior policy and hyper-parameter, and show the result in Figure 1. It can be observed how most choices of hyper-parameters result in good performance for PIWO-IX and

4. https://www.openml.org/search?type=data&status=active&id=247
5. The choice of the regularization parameter $\alpha$ did not seem to impact significantly the result of the experiments.

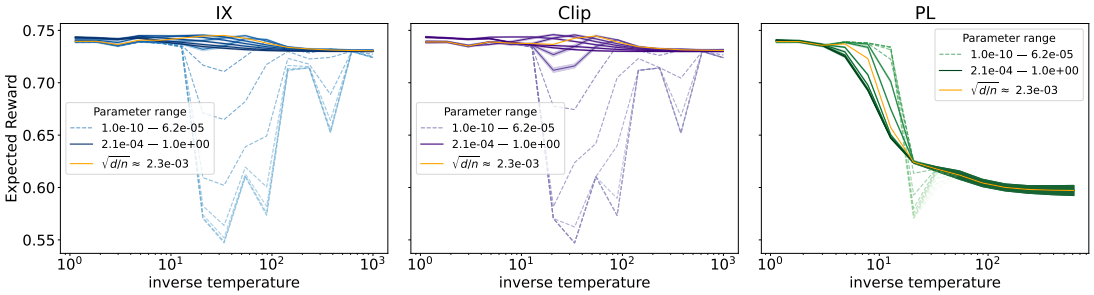


Figure 1: The performance of PIWO-IX, PIWO-CLIP, and the algorithm of Wang et al. (2023) as a function of the softmax parameter of the behavior policy. Different curves correspond to different hyperparameters for the algorithms, with lighter tones representing smaller hyperparameters and darker tones representing larger ones.

PIWO-CLIP, while the same cannot be said for PIWO-PL, which is very sensitive to small probabilities in the behavior policy and needs to compensate them with a very careful choice of its hyper-parameter. In particular, we note that in some experiments with large softmax parameters, $\mu(a|x)$ can be as low as $10^{-100}$ for some context-action pairs, and thus even a seemingly negligible regularization parameter like $\beta = 10^{-20}$ can result in massive pessimistic adjustments. In contrast, PIWO-IX is robust to the presence of such small observation probabilities and continues to work well for a broad range of hyperparameter choices. As expected, PIWO-CLIP performs very similarly to PIWO-IX due to the close similarity between these two methods. More details about the experiments are provided in Appendix C.

## 7. Discussion

We now provide some additional discussion on our results, related work, and open problems.

**No more uniform coverage.** The bounds we have proved are tighter than any that are known in the literature, and they have the particular strength that they do not require the action probabilities to be strictly bounded away from zero. Virtually all previous bounds require this "uniform-coverage" assumption, largely due to their excessive reliance on textbook concentration results like Bernstein's, Bennett's, or Freedman's inequalities. The only result we are aware of that does not explicitly suffer from this limitation is by Jin et al. (2022), who rely on a very sophisticated new proof technique which eventually does not yield easily interpretable performance bounds due to the appearance of some higher moments of the importance weights. The key to our stronger results is the observation that the tails of importance-weighted reward estimators are *asymmetric*: their lower tails are always light, and thus one only has to tame the upper tails via pessimistic adjustments. This simple observation allows us to derive very tight bounds using a few lines of elementary derivations. If there is any moral to this story, then it is that one should always avoid using two-sided concentration inequalities for importance-weighted estimators (at least as long as the rewards are positive).

**Implicit exploration and clipped importance weighting.** A perhaps more traditional way to control the tails of importance-weighted estimators is the clipped importance weighting (CIW) estimator we have defined in Section 6. Variants of this estimator have been studied at least since the work of Ionides (2008) and vigorously applied in the offline learning literature (Bottou et al., 2013;

Sakhi et al., 2023; Flynn et al., 2023). Interestingly, despite its broad usage, we are not aware of any work in this context that has worked out expressions for the bias of the CIW estimator, much less derived a regret bound for the resulting offline learning scheme. We believe that our results for the closely related IX estimator should essentially all apply to the CIW estimator and indeed our experiments show that they behave nearly identically in the settings we have tested. Nevertheless, we suspect that analyzing this estimator would end up being considerably more involved than our own analysis, but of course we would love to be proved wrong by future work.

**Reward-scaled coverage ratios.** A subtle improvement of our bounds as opposed to the ones of Wang et al. (2023) is that they depend on the *reward-scaled* version of the coverage ratio. This implies that bounds expressed in terms the scaled ratio $C_\gamma(\pi^*)$ can be much tighter than ones expressed in terms of $C(\pi^*)$ when the rewards of the comparator policy $\pi^*$ "tend to be small" in an appropriate sense. Note that this is a significant improvement in practical applications like online recommendation systems, where expected rewards correspond to clickthrough rates, which are very close to zero even for the very best ad campaigns. In the special case where rewards are negatively correlated with the importance weights (which may intuitively happen if the behavior policy is "reasonably good" in the sense that it puts larger weights on good actions), the coverage ratio against the optimal *deterministic* policy $\pi^*$ can be shown to satisfy $C_\gamma(\pi^*) \leq \upsilon(\pi^*)C(\pi^*)$, thus improving greatly over standard bounds that depend on $C(\pi^*)$. Bounds that improve for small expected rewards are known in the bandit literature at least since the work of Auer et al. (2002), and we are curious if guarantees like the above can be proved under more general conditions for offline learning as well.

**Lower bounds.** The "optimality" of pessimistic offline learning methods is a contentious topic that we prefer not to discuss here in much detail. In particular, even in the simplest case of offline learning in multi-armed bandits, Xiao et al. (2021) have shown that a large range of algorithms including pessimistic, greedy, and optimistic methods satisfies the standard notion of minimax optimality, and there is thus nothing special about pessimistic methods in these terms. Putting this alarming concern aside, pessimistic algorithms tend to have the property that their regret scales with the minimax sample complexity of *estimating* the value of the comparator policy (Xiao et al., 2021; Jin et al., 2021). In our case, it is not entirely clear if this statement continues to be true. In the special case of multi-armed bandits with binary rewards and a deterministic comparator policy, our bound matches the lower bound proved by Li et al. (2015) (up to a $\log K$ factor). That said, already in the case of stochastic comparator policies, our upper bounds no longer match the minimax sample complexity of estimation. Finding out if better algorithms with matching regret guarantees can be developed is a very interesting research question that we leave open for now.

**Computational-statistical tradeoffs.** As we show in this paper, it is possible to develop oracle-efficient algorithms with good statistical guarantees. However, these algorithms don't seem to demonstrate the correct scaling with the problem complexity unless prior knowledge of problem parameters is provided to the algorithm. This limitation can be bypassed by a more involved algorithm we describe in Appendix B, but the resulting method cannot apparently be implemented via a single call to the optimization oracle. Whether or not this computational-statistical tradeoff is inherent to the problem is unclear at this point and warrants further research.

**Further refinements.** Our results can be extended in a number of straightforward ways by building on previous developments in the literature. For instance, the dependence on $\log|\Pi|$ appearing in our main results can be most likely replaced by other complexity measures like covering num-

bers or the Natarajan dimension of the policy class, by adapting the techniques of either Swaminathan and Joachims (2015) or Jin et al. (2022). Similar bounds can be recovered by our PAC-Bayesian guarantees presented in Section 5 by building on techniques of Audibert (2004); Catoni (2007) (see also Grünwald et al., 2021). Another very simple generalization that our framework can readily handle is the case of adaptive behavior policies, where each sample point $(X_t, A_t, R_t)$ can be generated by a different behavior policy $\mu_t$ that may potentially depend on all past observations. The concentration bounds of Lemmas 2 and 3 can be very easily adapted to deal with such observations, and accordingly a version of our main result can be proved with the quantity $\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}\sum_{a}\frac{\pi^*(a|X_t)}{\mu_t(a|X_t)+\gamma}\cdot r(X_t,a)\right]$ taking the role of $C_\gamma(\pi^*)$. We hope that the simplicity of our techniques will enable further progress on the topic of importance-weighted offline learning, and in particular that further interesting extensions will be uncovered by future work.

## Acknowledgments

Matteo Papini is supported by FAIR (Future Artificial Intelligence Research) project, funded by the NextGenerationEU program within the PNRR-PE-AI scheme (M4C2, Investment 1.3, Line on Artificial Intelligence).

## References

Alekh Agarwal, Daniel J. Hsu, Satyen Kale, John Langford, Lihong Li, and Robert E. Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *ICML*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 1638–1646. JMLR.org, 2014.

Pierre Alquier. User-friendly introduction to PAC-Bayes bounds. *arXiv preprint arXiv:2110.11216*, 2021.

Jean-Yves Audibert. *PAC-Bayesian statistical learning theory*. PhD thesis, Université Paris VI, 2004.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multi-armed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002.

Dimitris Bertsimas, Nathan Kallus, Alexander M Weinstein, and Ying Daisy Zhuo. Personalized diabetes management using electronic medical records. *Diabetes care*, 40(2):210–217, 2017.

Léon Bottou, Jonas Peters, Joaquin Quiñonero Candela, Denis Xavier Charles, Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Y. Simard, and Ed Snelson. Counterfactual reasoning and learning systems: the example of computational advertising. *J. Mach. Learn. Res.*, 14(1): 3207–3260, 2013.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities - A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Olivier Catoni. PAC-Bayesian supervised classification. *Lecture Notes-Monograph Series. IMS*, 1277, 2007.

Miroslav Dudík, Daniel J. Hsu, Satyen Kale, Nikos Karampatziakis, John Langford, Lev Reyzin, and Tong Zhang. Efficient optimal learning for contextual bandits. In *UAI*, pages 169–178. AUAI Press, 2011.

Vivek F. Farias and Andrew A. Li. Learning preferences with side information. *Manag. Sci.*, 65(7): 3131–3149, 2019.

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*, 2023.

Germano Gabbianelli, Gergely Neu, and Matteo Papini. Online learning with off-policy feedback. In *ALT*, volume 201 of *Proceedings of Machine Learning Research*, pages 620–641. PMLR, 2023.

Peter Grünwald, Thomas Steinke, and Lydia Zakynthinou. PAC-Bayes, MAC-Bayes and conditional mutual information: Fast rate bounds that handle general VC classes. In *COLT*, volume 134 of *Proceedings of Machine Learning Research*, pages 2217–2247. PMLR, 2021.

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.

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

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 5084–5096. PMLR, 2021.

Ying Jin, Zhimei Ren, Zhuoran Yang, and Zhaoran Wang. Policy learning "without" overlap: Pessimism and generalized empirical Bernstein's inequality. *arXiv preprint arXiv:2212.09900*, 2022.

Edward S. Kim, Roy S. Herbst, Ignacio I. Wistuba, J. Jack Lee, Jr. Blumenschein, George R., Anne Tsao, David J. Stewart, Marshall E. Hicks, Jr. Erasmus, Jeremy, Sanjay Gupta, Christine M. Alden, Suyu Liu, Ximing Tang, Fadlo R. Khuri, Hai T. Tran, Bruce E. Johnson, John V. Heymach, Li Mao, Frank Fossella, Merrill S. Kies, Vassiliki Papadimitrakopoulou, Suzanne E. Davis, Scott M. Lippman, and Waun K. Hong. *Cancer discovery*, 1(1):44–53, 2011.

Tomáš Kocák, Gergely Neu, Michal Valko, and Rémi Munos. Efficient learning by implicit exploration in bandit problems with side observations. In *NIPS*, pages 613–621, 2014.

Akshay Krishnamurthy, John Langford, Aleksandrs Slivkins, and Chicheng Zhang. Contextual bandits with continuous actions: Smoothing, zooming, and adapting. In *COLT*, volume 99 of *Proceedings of Machine Learning Research*, pages 2025–2027. PMLR, 2019.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *NeurIPS*, pages 817–824. Curran Associates, Inc., 2007.

Gene Li, Cong Ma, and Nati Srebro. Pessimism for offline linear contextual bandits using $\ell_p$ confidence sets. In *NeurIPS*, 2022.

Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *WSDM*, pages 297–306. ACM, 2011.

Lihong Li, Rémi Munos, and Csaba Szepesvári. Toward minimax off-policy value estimation. In *AISTATS*, volume 38 of *JMLR Workshop and Conference Proceedings*. JMLR.org, 2015.

Ben London and Ted Sandler. Bayesian counterfactual risk minimization. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 4125–4133. PMLR, 2019.

David A. McAllester. Some PAC-Bayesian theorems. In *COLT*, pages 230–234. ACM, 1998.

Susan A Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 65(2):331–355, 2003.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *NeurIPS*, pages 3168–3176, 2015.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. In *NeurIPS*, pages 11702–11716, 2021.

James M Rehg, Susan A Murphy, and Santosh Kumar. Mobile health. Springer, 2017.

Otmane Sakhi, Pierre Alquier, and Nicolas Chopin. PAC-Bayesian offline contextual bandits with guarantees. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 29777–29799. PMLR, 2023.

Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *ICML*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1670–1679. JMLR.org, 2016.

Daniel A. Spielman and Shang-Hua Teng. Smoothed analysis of algorithms: why the simplex algorithm usually takes polynomial time. In *STOC*, pages 296–305. ACM, 2001.

Adith Swaminathan and Thorsten Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *J. Mach. Learn. Res.*, 16:1731–1755, 2015.

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

Chenjun Xiao, Yifan Wu, Jincheng Mei, Bo Dai, Tor Lattimore, Lihong Li, Csaba Szepesvári, and Dale Schuurmans. On the optimality of batch policy optimization algorithms. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 11362–11371. PMLR, 2021.

## Appendix A. Omitted proofs

In this section, we prove our main technical lemmas. To facilitate this effort, we introduce the shorthand notations

$$\widehat{r}_t(\pi) = \frac{\pi(A_t|X_t)}{\mu(A_t|X_t)} R_t, \qquad \text{and} \qquad \widetilde{r}_t(\pi) = \frac{\pi(A_t|X_t)}{\mu(A_t|X_t)+\gamma} R_t,$$

and note that $\widehat{v}_n = \frac{1}{n}\sum_{t=1}^n \widehat{r}_t(\pi)$ and $\widetilde{v}_n = \frac{1}{n}\sum_{t=1}^n \widetilde{r}_t(\pi)$, and also recall that $\mathbb{E}[\widehat{r}_t(\pi)] = v(\pi)$ and $\mathbb{E}[\widetilde{r}_t(\pi)] = v(\pi) - \gamma C_\gamma(\pi)$ holds for all $t$.

### A.1. The proof of Lemma 2

Fix an arbitrary $\pi \in \Pi$. We start by using the elementary inequality $\log(1+y) \geq \frac{y}{1+y/2}$ that holds for all $y \geq 0$ to show that

$$\widetilde{r}_t(\pi) = \frac{\pi(A_t|X_t)R_t}{\mu(A_t|X_t)+\gamma} \leq \frac{\pi(A_t|X_t)R_t}{\mu(A_t|X_t)+\gamma\pi(A_t|X_t)R_t} = \frac{1}{2\gamma}\cdot\frac{2\gamma\widehat{r}_t(\pi)}{1+\gamma\widehat{r}_t(\pi)} \leq \frac{\log\left(1+2\gamma\widehat{r}_t(\pi)\right)}{2\gamma}.$$

This implies that

$$\mathbb{E}[e^{2\gamma\widetilde{r}_t(\pi)}] \leq \mathbb{E}[1+2\gamma\widehat{r}_t(\pi)] = 1+2\gamma v(\pi) \leq e^{2\gamma v(\pi)},$$

where the last step follows from the inequality $e^y \geq 1+y$ that holds for all $y \in \mathbb{R}$. Using the independence of all observations, this implies $\mathbb{E}[e^{2\gamma\sum_{t=1}^n(\widetilde{r}_t(\pi)-v(\pi))}] \leq 1$, and thus an application of Markov's inequality yields

$$\mathbb{P}\left[\sum_{t=1}^n(\widetilde{r}_t(\pi)-v(\pi)) \geq \varepsilon\right] = \mathbb{P}\left[e^{2\gamma\sum_{t=1}^n(\widetilde{r}_t(\pi)-v(\pi))} \geq e^{2\gamma\varepsilon}\right] \leq e^{-2\gamma\varepsilon}$$

for any $\varepsilon \geq 0$. Setting $\varepsilon = \frac{\log(|\Pi|/\delta)}{2\gamma}$ and taking a union bound over all policies concludes the proof. ■

### A.2. The proof of Lemma 3

Fix an arbitrary $\pi \in \Pi$. We start by noting that for any nonnegative random variable $Y$, and for any positive $\lambda$, we have

$$\mathbb{E}[e^{-\lambda Y}] \leq \mathbb{E}[1-\lambda Y+\lambda^2 Y^2/2] \leq e^{-\lambda\mathbb{E}[Y]+\lambda^2\mathbb{E}[Y^2]/2},$$

where the first inequality follows from $e^{-y} \leq 1-y+y^2/2$ that holds for all $y \geq 0$ and the second from $e^y \geq 1+y$ that holds for all $y \in \mathbb{R}$. Apply this inequality with $Y = \widetilde{r}_t(\pi)$ and note that

$$\begin{aligned}\mathbb{E}\Big[(\widetilde{r}_t(\pi))^2\Big] &= \mathbb{E}\left[\frac{(\pi(A_t|X_t))^2}{(\mu(A_t|X_t)+\gamma)^2}\cdot R_t^2\right] \leq \mathbb{E}\left[\sum_a \mathbb{I}_{\{A_t=a\}}\frac{\pi(a|X_t)}{(\mu(a|X_t)+\gamma)^2}\cdot r(X_t,a)\right] \\ &\leq \mathbb{E}\left[\sum_a \frac{\pi(a|X_t)}{\mu(a|X_t)+\gamma}\cdot r(X_t,a)\right] = C_\gamma(\pi),\end{aligned}$$

where the first inequality used the boundedness of the rewards to show $\mathbb{E}\left[R_t^2\right] \leq \mathbb{E}\left[R_t\right] = \mathbb{E}\left[r(X_t, a)\right]$ and $\left(\pi(a|X_t)\right)^2 \leq \pi(a|X_t)$, and the second inequality used that $\mathbb{E}\left[\mathbb{I}_{\{A_t=a\}} \middle| X_t\right] = \mu(a|X_t)$.

Using the independence of all observations, this implies $\mathbb{E}[e^{\lambda \sum_{t=1}^n (\mathbb{E}[\widetilde{r}_t(\pi)] - \widetilde{r}_t(\pi) - \lambda C_\gamma(\pi)/2)}] \leq 1$. Recalling that $\mathbb{E}[\widetilde{r}_t(\pi)] = v(\pi) - \gamma C_\gamma(\pi)$, an application of Markov's inequality yields

$$\mathbb{P}\left[\sum_{t=1}^n \left(v(\pi) - \widetilde{r}_t(\pi) - (\gamma + \lambda/2)\, C_\gamma(\pi)\right) \geq \varepsilon\right] \leq e^{-2\gamma\varepsilon}.$$

Setting $\lambda = 2\gamma$ and $\varepsilon = \frac{\log(|\Pi|/\delta)}{2\gamma}$, and finally taking a union bound over all policies concludes the proof. ■

### A.3. The proofs of Lemmas 5 and 6

To prove Lemma 5, let us first fix an arbitrary $Q \in \Delta_\Pi$, and recall from the proof of Lemma 2 that $\mathbb{E}[e^{2\gamma\widetilde{r}_t(\pi)}] \leq e^{2\gamma v(\pi)}$, holds for all fixed $\pi$. Thus, since $P$ is independent of the random observations, we also have

$$\mathbb{E}\left[\int e^{2\gamma \sum_{t=1}^n (\widetilde{r}_t(\pi) - v(\pi))} \mathrm{d}P(\pi)\right] \leq 1.$$

Now, let us introduce the notation $\rho_\pi(Q, P) = \log \frac{\mathrm{d}Q}{\mathrm{d}P}(\pi)$ and write

$$\begin{aligned}
&\mathbb{P}\left[\int \left(\sum_{t=1}^n (\widetilde{r}_t(\pi) - v(\pi)) - \frac{\rho_\pi(Q,P)}{2\gamma}\right) \mathrm{d}Q(\pi) \geq \varepsilon\right] \\
&\quad \leq \mathbb{E}\left[e^{2\gamma \int \left(\sum_{t=1}^n (\widetilde{r}_t(\pi) - v(\pi)) - \frac{\rho_\pi(Q,P)}{2\gamma}\right) \mathrm{d}Q(\pi)}\right] e^{-2\gamma\varepsilon} \\
&\quad \leq \mathbb{E}\left[\int e^{2\gamma\left(\sum_{t=1}^n (\widetilde{r}_t(\pi) - v(\pi)) - \frac{\rho_\pi(Q,P)}{2\gamma}\right)} \mathrm{d}Q(\pi)\right] e^{-2\gamma\varepsilon} \\
&\quad = \mathbb{E}\left[\int e^{2\gamma\left(\sum_{t=1}^n (\widetilde{r}_t(\pi) - v(\pi))\right)} \frac{\mathrm{d}P}{\mathrm{d}Q}(\pi) \mathrm{d}Q(\pi)\right] e^{-2\gamma\varepsilon} \\
&\quad = \mathbb{E}\left[\int e^{2\gamma\left(\sum_{t=1}^n (\widetilde{r}_t(\pi) - v(\pi))\right)} \mathrm{d}P(\pi)\right] e^{-2\gamma\varepsilon} \leq e^{-2\gamma\varepsilon}.
\end{aligned}$$

Here, the first step follows from Markov's inequality, the second from Jensen's inequality for the convex function $y \mapsto e^{2\gamma y}$, the third from the definition of $\rho_\pi(Q, P)$, the fourth from the definition of the Radon–Nykodim derivative $\frac{\mathrm{d}P}{\mathrm{d}Q}$, and the last step from the inequality that we have established above. Noticing that $\int \rho_\pi(Q, P) \mathrm{d}Q(\pi) = \mathrm{KL}\left(Q \| P\right)$ and setting $\varepsilon = \frac{\log(1/\delta)}{2\gamma}$ concludes the proof of Lemma 5. The proof of Lemma 6 then follows analogously by recalling from the proof of Lemma 3 that $\mathbb{E}[e^{2\gamma(v(\pi) - \widetilde{r}_t(\pi) - 2\gamma C_\gamma(\pi))}] \leq 1$, and then following the same steps as above. ■

## Appendix B. Adaptivity to the coverage

One shortcoming of the result in Theorem 1 is that it scales linearly with $C_\gamma(\pi^*)$ even though prior results suggest that a scaling with $\sqrt{C_0(\pi^*)}$ should be possible (Swaminathan and Joachims, 2015; Wang et al., 2023). This improvement can be trivially achieved by setting $\gamma = \sqrt{\frac{\log(|\Pi|/\delta)}{C_0(\pi^*)n}}$, but this requires prior knowledge of $C_0(\pi^*)$ which is of course unavailable in practice (at least in the most interesting case where $\pi^*$ is the optimal policy).

This limitation can be addressed by defining the following *non-uniformly scaled* version of the IX estimator:

$$\widetilde{v}_n^\dagger(\pi) = \frac{1}{n}\sum_{t=1}^{n} \frac{\pi(A_t|X_t)}{\mu(A_t|X_t)+\gamma_\pi} \cdot R_t - \frac{\log\left(|\Pi|/\delta\right)}{2\gamma_\pi}. \tag{6}$$

Here, $\gamma_\pi > 0$ is a *policy-dependent* IX parameter that is potentially different for each policy $\pi$. Using this estimator, we define a variant of our main algorithm called *coverage-scaled PIWO-IX* that outputs

$$\widehat{\pi}_n = \operatorname*{arg\,min}_{\pi\in\Pi} \widetilde{v}_n^\dagger(\pi).$$

Notice that, unlike PIWO-IX, this algorithm cannot be directly implemented using a standard optimization oracle due to the policy-dependent IX parameters $\gamma_\pi$. The following theorem is straightforward to prove using our previously established Lemmas 2 and 3:

**Theorem 7** *With probability at least $1-\delta$, the regret of coverage-scaled* PIWO-IX *against any comparator policy $\pi^* \in \Pi$ satisfies*

$$\mathfrak{R}_n(\pi^*) \le \frac{\log\left(2|\Pi|/\delta\right)}{\gamma_{\pi^*} n} + 2\gamma_{\pi^*} C_{\gamma_{\pi^*}}(\pi^*).$$

*Furthermore, by setting $\gamma_\pi = \sqrt{\frac{\log(2|\Pi|/\delta)}{2C_0(\pi)n}}$ for each $\pi$, the bound becomes*

$$\mathfrak{R}_n(\pi^*) \le \sqrt{\frac{8C_0(\pi^*)\log\left(2|\Pi|/\delta\right)}{n}}.$$

**Proof** First observe that the statements of Lemmas 2 and 3 can be trivially adjusted to show that the bounds

$$0 \le v(\pi) - \widetilde{v}_n^\dagger(\pi) \le \frac{\log(2|\Pi|/\delta)}{\gamma_\pi n} + 2\gamma C_{\gamma_\pi}(\pi).$$

hold simultaneously for all policies with probability at least $1-\delta$. Then, by the definition of the algorithm, we obtain

$$v(\widehat{\pi}_n) \ge \widetilde{v}_n^\dagger(\widehat{\pi}_n) \ge \widetilde{v}_n^\dagger(\pi^*) \ge v(\pi^*) - \frac{\log\left(2|\Pi|/\delta\right)}{\gamma_{\pi^*} n} - 2\gamma_{\pi^*} C_{\gamma_{\pi^*}}(\pi^*).$$

This concludes the proof of the first claim. The second claim can be verified by noticing that $C_\gamma(\pi^*) \le C_0(\pi^*)$ for all $\gamma > 0$ and plugging in the choice of $\gamma_\pi$ stated in the theorem. ■

## Appendix C. Further details on the experiments

In this section we give more detail on all the experiments we ran. The first step we performed was to use $10\%$ of the data to fit a multivariate ridge regressor $\texttt{reg}(x, a)$ to predict the expected reward of each action, given any context. For each context $x$ and each corresponding optimal action $a^*$ in the data, we selected $M_{\cdot,a^*}$ as the label vector (having one entry for each possible action).

We then used the reamining $90\%$ of the data to perform two sets of experiments. In the first set, which is the one described in the main text (Section 6), we considered 20 softmax behavior policies, varying their inverse temperature parameter $\eta$ as $\texttt{logspace(-1, 3, 20)}$. That is,

$$\pi_\eta(a|x) \propto \exp(\eta\, \texttt{reg}(x, a)).$$

We repeated each set of experiments 10 times ($i \in [10]$), using a 10-fold validation procedure. That is, the data was first partitioned into 10 non overlapping folds. On each repetition $i$, 9 folds are used to generate the training data for the algorithms, by simulating the interaction of each behavior policy $\pi_\eta$ and the bandit instance. The resulting training dataset $\mathcal{D}_{\eta,i}$ was used to train each algorithm for each possible hyper-parameter choice $h \in \texttt{logspace(-10, 0, 20))}$. Finally, each trained algorithm $\mathfrak{A}_{\eta,i,h}$ is evaluated using the data in the remaining fold, by computing the expected regret using the true mean rewards.

This set of experiments was then repeated for a different set of "bad" behavior policies, which were defined as

$$\pi_\eta(a|x) \propto \exp(-\eta\, \texttt{reg}(x, a)).$$

The results for the two sets of experiments are shown respectively in Figures 2 and 3. On each figure, the first row of plots shows the expected reward as a function of the inverse temperature parameter $\eta$. Each plot on the row is for one of the three different algorithms, and it contains a line for each possible hyper-parameter. The lines are colored using a gradient from lighter to darker to represent increasing hyper-parameter values. In orange we highlighted the learning rate corresponding to $\sqrt{d/n}$, which we use as a crude approximation of the hyper-parameter recommended by theory, $\sqrt{\log|\Pi|/n}$. In addition, values of the hyper-parameters much smaller than $\sqrt{d/n}$ are represented with a dashed line. All lines (excluding for clarity of the representation the dashed ones) have a shaded region representing the standard deviation over the 10 runs. The second row of plots shows the expected regret as a function of the hyper-parameter $h$. Thus, we can observe a line for each different behavior policy parameter $\eta$. Here the lines are lighter for smaller values of $\eta$, and darker for bigger values of $\eta$.

From the plots, we can infer that PIWO-IX performs well when the behavior policy is "good" and $\gamma$ is set in a broad proximity of its theoretically recommended value. This behavior appears to be robust as we vary the degree of "goodness" of the policy modulated by the softmax parameter $\eta$, and in particular performance stays good even as $\eta$ approaches its higher extremes and the behavior policy gets more and more deterministic. As expected, PIWO-CLIP behaves comparably. In comparison PIWO-PL is a lot less robust in this case and its performance decays as $\eta$ increases, most likely due to the more and more extreme values of the importance weights arising from some sampling probabilities approaching zero. We note that the the case of "good" behavior policies is the most practical use case, and our experiments suggest that our algorithm performs excellently in this scenario for a wide range of hyperparameters.

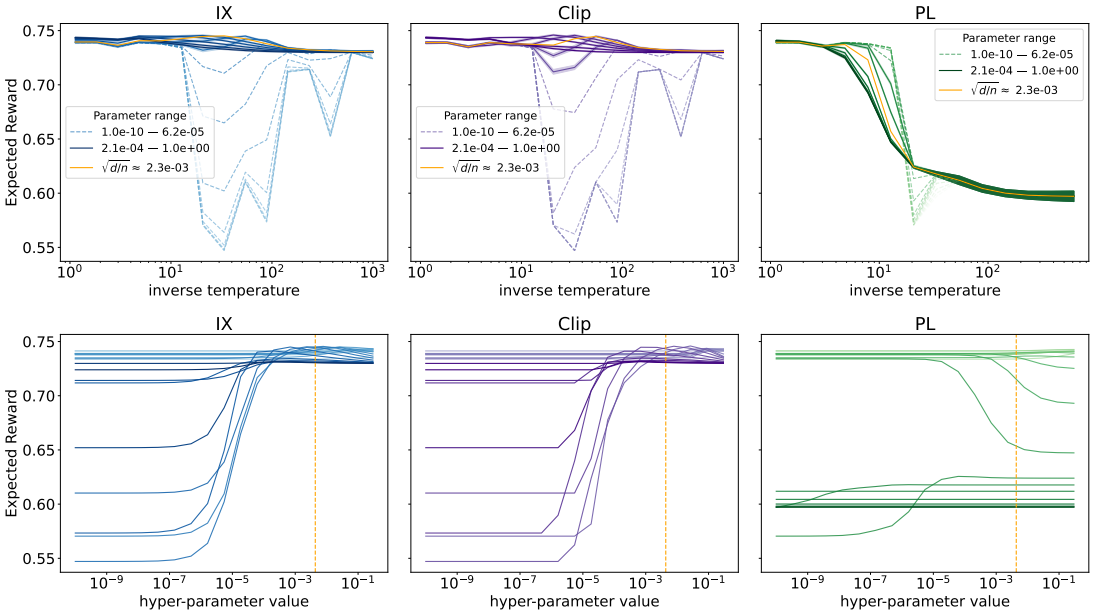


Figure 2: Results of PIWO-IX, PIWO-CLIP, and PIWO-PL with good behavior policies.

In comparison, the picture changes when considering the case of "bad" behavior policies. In this case, PIWO-IX and performs worse and worse as $\gamma$ is increased, especially for large values of $\eta$ corresponding to particularly bad behavior policies. This is not surprising given that the policy coverage ratio blows up in this extreme, as less and less mass is put on well-performing actions. Also notice that increasing the regularization parameter $\gamma$ forces the algorithm to be more and more pessimistic and thus stay closer and closer to the behavior policy, which again results in decaying performance. The performance of PIWO-PL is less consistent in this case, and it is hard to read out patterns that are well-predicted by theory.

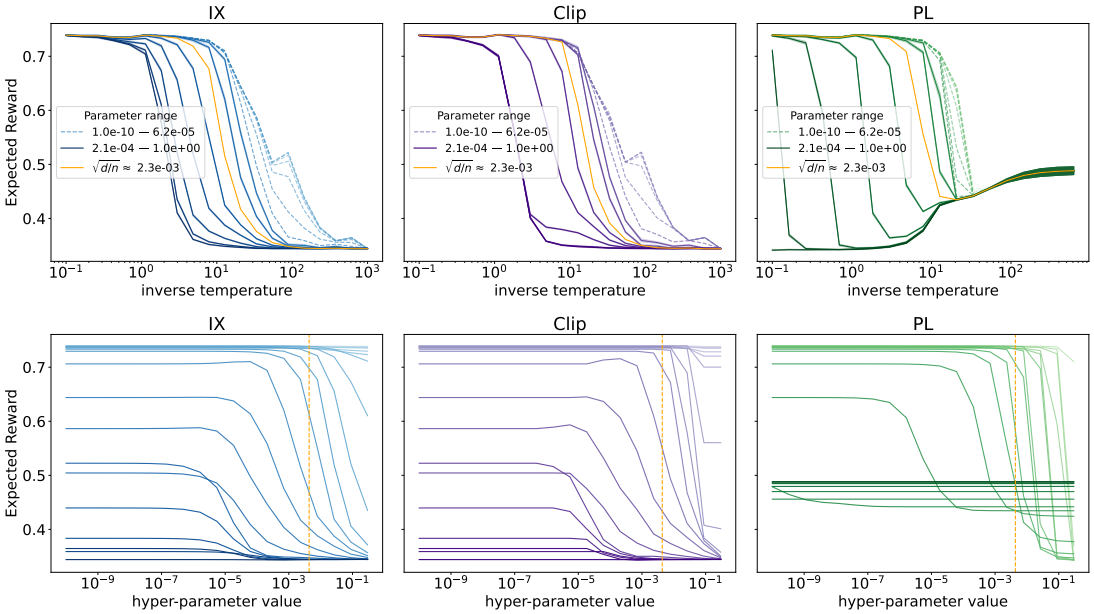


Figure 3: Results of PIWO-IX, PIWO-CLIP, and PIWO-PL with bad behavior policies.