# PAC-Bayesian Offline Contextual Bandits With Guarantees

**Otmane Sakhi** [1 2] **Pierre Alquier** [3] **Nicolas Chopin** [2]

## Abstract

This paper introduces a new principled approach for off-policy learning in contextual bandits. Unlike previous work, our approach does not derive learning principles from intractable or loose bounds. We analyse the problem through the PAC-Bayesian lens, interpreting policies as mixtures of decision rules. This allows us to propose novel generalization bounds and provide tractable algorithms to optimize them. We prove that the derived bounds are tighter than their competitors, and can be optimized directly to confidently improve upon the logging policy *offline*. Our approach learns policies with guarantees, uses all available data and does not require tuning additional hyperparameters on held-out sets. We demonstrate through extensive experiments the effectiveness of our approach in providing performance guarantees in practical scenarios.

## 1. Introduction

Online industrial systems encounter sequential decision problems as they interact with the environment and strive to improve based on the received feedback. The contextual bandit framework formalizes this mechanism, and proved valuable with applications in recommender systems (Valko et al., 2014) and clinical trials (Villar et al., 2015). It describes a game of repeated interactions between a system and an environment, where the latter reveals a context that the system interacts with, and receives a feedback in return.

While the online solution to this problem involves strategies that find an optimal trade-off between exploration and exploitation to minimize the *regret* (Lattimore & Szepesvári, 2020), we are concerned with its offline formulation (Swaminathan & Joachims, 2015b), which is arguably better suited for real-life applications, where more control over the decision-maker, often coined *the policy*, is needed. The learning of the policy is performed offline, based on historical data, typically obtained by logging the interactions between an older version of the decision system and the environment. By leveraging this data, our goal is to discover new strategies of greater performance.

There are two main paths to address this learning problem. The *direct method* (Sakhi et al., 2020a; Jeunen & Goethals, 2021) attacks the problem by modelling the feedback and deriving a policy according to this model. This approach can be praised for its simplicity (Brandfonbrener et al., 2021), is well-studied in the offline setting (Nguyen-Tang et al., 2022) but will often suffer from a bias as the feedback received is complex and the efficiency of the method directly depends on our ability to understand the problem's structure.

We will consider the second path of off-policy learning, or IPS: *inverse propensity scoring* (Horvitz & Thompson, 1952) where we learn the policy directly from the logged data after correcting its bias with importance sampling (Owen & Zhou, 2000). As these estimators (Bottou et al., 2013; Swaminathan & Joachims, 2015a) can suffer from a variance problem as we drift away from the logging policy, the literature gave birth to different learning principles (Swaminathan & Joachims, 2015b; Ma et al., 2019; London & Sandler, 2019; Faury et al., 2020) motivating penalizations toward the logging policy. These principles are inspired by generalization bounds, but introduce a hyperparameter $\lambda$ to either replace intractable quantities (Swaminathan & Joachims, 2015b) or to tighten a potentially vacuous bound (London & Sandler, 2019). These approaches require tuning $\lambda$ on a held-out set and sometimes fail at improving the previous decision system (London & Sandler, 2019; Chen et al., 2019).

In this work, we analyse off-policy learning from the PAC-Bayesian perspective (McAllester, 1998; Catoni, 2007). We aim at introducing a novel, theoretically-grounded approach, based on the direct optimization of newly derived tight generalization bounds, to obtain guaranteed improvement of the previous system ***offline***, without the need for held-out sets nor hyperparameter tuning. We show that our approach is perfectly suited to this framework, as it naturally incorporates information about the old decision system and can confidently improve it.

[1]Criteo AI Lab, Paris, France [2]CREST, ENSAE, IPP, Palaiseau, France [3]ESSEC Business School, Asia-Pacific campus, Singapore. Correspondence to: Otmane Sakhi <o.sakhi@criteo.com>.

*Proceedings of the $40^{th}$ International Conference on Machine Learning*, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

## 2. Preliminaries

### 2.1. Setting

We use $x \in \mathcal{X}$ to denote a context and $a \in \mathcal{A} = [K]$ an action, where $K$ denotes the number of available actions. For a context $x$, each action is associated with a cost $c(x, a) \in [-1, 0]$, with the convention that better actions have smaller cost. The cost function $c$ is unknown. Our decision system is represented by its policy $\pi : \mathcal{X} \to \mathcal{P}(\mathcal{A})$ which given $x \in \mathcal{X}$, defines a probability distribution over the discrete action space $\mathcal{A}$ of size $K$. Assuming that the contexts are stochastic and follow an unknown distribution $\nu$, we define the *risk* of the policy $\pi$ as the expected cost one suffers when playing actions according to $\pi$:

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

The learning problem is to find a policy $\pi$ which minimizes the risk. This risk can be naively estimated by deploying the policy online and gathering enough interactions to construct an accurate estimate. Unfortunately, we do not have this luxury in most real-world problems as the cost of deploying bad policies can be extremely high. We can obtain instead an estimate by exploiting the logged interactions collected by the previous system. Indeed, the previous system is represented by a *logging policy* $\pi_0$ (e.g a previous version of a recommender system that we are trying to improve), which gathered interaction data of the following form:

$$\mathcal{D}_n = \{x_i, a_i \sim \pi_0(\cdot|x_i), c_i\}_{i\in[n]}, \quad \text{with } c_i = c(x_i, a_i).$$

Given this data, one can build various estimators, with the clipped IPS (Bottou et al., 2013) the most commonly used. It is constructed based on a clipping of the importance weights or the logging propensities to mitigate variance issues (Ionides, 2008). We are more interested in clipping the logging probabilities as we need objectives that are linear in the policy $\pi$ for our study. The cIPS estimator is given by:

$$\hat{R}_n^{\tau}(\pi) = \frac{1}{n}\sum_{i=1}^{n} \frac{\pi(a_i|x_i)}{\max\{\pi_0(a_i|x_i), \tau\}} c_i \tag{1}$$

with $\tau \in [0, 1]$ being the clipping factor.

Choosing $\tau \ll 1$ reduces the bias of cIPS. We recover the classical IPS estimator (Horvitz & Thompson, 1952) (unbiased under mild conditions) by taking $\tau = 0$.

Another estimator with better statistical properties is the doubly robust estimator (Ben-Tal et al., 2013), which uses the importance weights as control variates to reduce further the variance of the cIPS estimators. This estimator is asymptotically optimal (Farajtabar et al., 2018) (in terms of variance) amongst the class of unbiased and consistent off-policy estimators.

We consider a simplified version of this estimator, which replaces the use of a model $\hat{c}$ of the cost by one parameter $\xi \in [-1, 0]$ that can be chosen freely. We define the control variate clipped IPS, or cvcIPS as follows:

$$\hat{R}_n^{\tau,\xi}(\pi) = \xi + \frac{1}{n}\sum_{i=1}^{n} \frac{\pi(a_i|x_i)}{\max\{\pi_0(a_i|x_i), \tau\}}(c_i - \xi). \tag{2}$$

The cvcIPS estimator can be seen as a special case of the doubly robust estimator when the cost model $\hat{c} = \xi$ is constant and $\tau = 0$. cIPS is recovered by setting $\xi = 0$. This simple estimator is deeply connected to the *SNIPS* estimator (Swaminathan & Joachims, 2015a) and was shown to be more suited to off-policy learning as it mitigates the problem of propensity overfitting (Joachims et al., 2018).

### 2.2. Related Work: Learning Principles

The literature so far has focused on deriving new principles to learn policies with good online performance. The first line of work in this direction is CRM: Counterfactual Risk minimization (Swaminathan & Joachims, 2015b) which adopted SVP: Sample Variance Penalization (Maurer & Pontil, 2009) to favor policies with small empirical risk and controlled variance. The intuition behind it is that the variance of cIPS depends on the disparity between $\pi$ and $\pi_0$ making the estimator unreliable when $\pi$ drifts away from $\pi_0$. The analysis focused on the cIPS estimator and used uniform bounds based on empirical Bernstein inequalities (Maurer & Pontil, 2009), where intractable quantities were replaced by a tuning parameter $\lambda$, giving the following learning objective:

$$\arg\min_{\pi} \left\{ \hat{R}_n^{\tau}(\pi) + \lambda\sqrt{\frac{\hat{V}_n(\pi)}{n}} \right\} \tag{3}$$

with $\hat{V}_n(\pi)$ the empirical variance of the cIPS estimator. A majorization-minimization algorithm was provided in (Swaminathan & Joachims, 2015b) to solve Equation (3) for parametrized softmax policies.

In the same spirit, (Faury et al., 2020; Sakhi et al., 2020b) generalize SVP using the distributional robustness framework, showing that the CRM principle can be retrieved with a particular choice of the divergence and provide asymptotic coverage results of the true risk. Their objectives are competitive with SVP while providing simple ways to scale its optimization to large datasets.

Another line of research, closer to our work, uses PAC-Bayesian bounds to derive learning objectives in the same fashion as (Swaminathan & Joachims, 2015b). Indeed, (London & Sandler, 2019) introduce the Bayesian CRM, motivating the use of $L_2$ regularization towards the parameter $\theta_0$ of the logging policy $\pi_0$. The analysis uses (McAllester, 2003)'s bound, is conducted on the cIPS estimator and con-

trols the $L_2$ norm by a hyperparameter $\lambda$, giving the following learning objective for parametrized softmax policies:

$$\arg\min_{\theta} \left\{ \hat{R}_n^{\tau}(\pi_\theta) + \lambda ||\theta - \theta_0||^2 \right\}. \quad (4)$$

(London & Sandler, 2019) minimize a convex upper-bound of objective (4) (by taking a log transform of the policy) which is amenable to stochastic optimization, giving better results than (3) while scaling better to the size of the dataset.

**Limitations.** The principles found in the literature are mainly inspired by generalization bounds. However, the bounds from where these principles are derived either depend on intractable quantities (Swaminathan & Joachims, 2015b) or are not tight enough to be used as-is (London & Sandler, 2019). For example, (Swaminathan & Joachims, 2015b) derive a generalisation bound (see Theorem 1 in (Swaminathan & Joachims, 2015b)) for offline policy learning using the notion of covering number. This introduces the complexity measure $\mathcal{Q}_{\mathcal{H}}(n, \gamma)$ that cannot be computed (even for simple policy classes) making their bound intractable. This forces the introduction of a hyperparameter $\lambda$ that needs further tuning. Unfortunately, this approach suffers from numerous problems:

- **No Theoretical Guarantees.** Introducing the hyperparameter $\lambda$ in Equations (3) and (4) gives tractable objectives, but loses the theoretical guarantees given by the initial bounds. These objectives do not necessarily cover the true risk, and optimizing them can lead to policies worse than the logging $\pi_0$. Empirical evidence can be found in (Chen et al., 2019; London & Sandler, 2019) where the SVP principle in Equation (3) fails to improve on $\pi_0$.
- **Inconsistent Strategy.** These principles were first introduced to mitigate the suboptimality of learning with off-policy estimators, deemed untrustworthy for their potential high variance. The strategy minimizes the objectives for different values of $\{\lambda_1, ..., \lambda_m\}$, generating a set of policy candidates $\{\pi_{\lambda_1}, ..., \pi_{\lambda_m}\}$, from which we select the best policy $\pi_{\lambda_*}$ according to the same untrustworthy, high variance estimators on a held-out set. This makes the selection strategy used inconsistent with what these principles are claiming to solve.
- **Tuning requires additional care.** Tuning $\lambda$ needs to be done on a held-out set. This means that we need to train multiple policies (computational burden) on a fraction of the data (data inefficiency), and select the best policy among the candidates using off-policy estimators (variance problem) on the held-out set.

In this paper, we derive a coherent principle, that learns policies using all the available data and provides guarantees on their performance, without the introduction of new hyperparameters.

### 2.3. Learning With Guaranteed Improvements

Our first concern in most applications is to improve upon the actual system $\pi_0$. As $\mathcal{D}_n$ is collected by $\pi_0$, we have access to $R(\pi_0)$[1]. Given a new policy $\pi$, we want to be confident that the improvement $\mathcal{I}(\pi, \pi_0) = R(\pi_0) - R(\pi)$ is positive before deployment.

Let us suppose that we are restricted to a class of policies $\mathcal{H}$, and have access to a generalization bound that gives the following result with high probability over draws of $\mathcal{D}_n$:

$$R(\pi) \leq \mathcal{UB}_n(\pi) \quad \forall \pi \in \mathcal{H}.$$

with $\mathcal{UB}_n$ an empirical upper bound that depends on $\mathcal{D}_n$. For any $\pi$, we define the guaranteed improvement:

$$\mathcal{GI}_{\mathcal{UB}_n}(\pi, \pi_0) = R(\pi_0) - \mathcal{UB}_n(\pi).$$

We can be sure of improving $R(\pi_0)$ offline if we manage to find $\pi \in \mathcal{H}$ that achieves $\mathcal{GI}_{\mathcal{UB}_n}(\pi, \pi_0) > 0$ as the following result will hold with high probability:

$$\mathcal{I}(\pi, \pi_0) \geq \mathcal{GI}_{\mathcal{UB}_n}(\pi, \pi_0) > 0.$$

To obtain such a policy, we look for the minimizer of $\mathcal{UB}_n$ over the class of policies $\mathcal{H}$ as:

$$\pi^*_{\mathcal{UB}_n} \in \arg\min_{\pi \in \mathcal{H}} \mathcal{UB}_n(\pi) = \arg\max_{\pi \in \mathcal{H}} \mathcal{GI}_{\mathcal{UB}_n}(\pi, \pi_0).$$

We define the best guaranteed risk and the best guaranteed improvement follows:

$$\mathcal{GR}^*_{\mathcal{UB}_n} = \mathcal{UB}_n(\pi^*_{\mathcal{UB}_n})$$
$$\mathcal{GI}^*_{\mathcal{UB}_n}(\pi_0) = R(\pi_0) - \mathcal{GR}^*_{\mathcal{UB}_n}.$$

> A *theoretically-grounded* strategy to improve $\pi_0$ will be to deploy $\pi^*_{\mathcal{UB}_n}$ if we obtain a positive guaranteed improvement $\mathcal{GI}^*_{\mathcal{UB}_n}(\pi_0) > 0$, otherwise continue collecting data with the current system $\pi_0$.

This strategy will always produce policies that are at least as good as $\pi_0$, optimizes directly a bound over all data and does not require held-out sets nor new hyperparameters. However, the tightness of the bounds $\mathcal{UB}_n$ will play an important role. Indeed, If we fix $\mathcal{D}_n$ and $\pi_0$, $\mathcal{GI}^*_{\mathcal{UB}_n}(\pi_0)$ will only depend on the minimum of $\mathcal{UB}_n$, motivating the derivation of bounds that are tractable and tight enough to achieve the smallest minimum possible.

In this regard, we opt for the PAC-Bayesian framework to tackle this problem as it is proven to give tractable, non-vacuous bounds even in difficult settings (Dziugaite & Roy,

[1] up to a small $\mathcal{O}(1/\sqrt{n})$ approximation error.

2017). Its paradigm also fits our application as we can incorporate information about the previous system $\pi_0$ in the form of a prior; see (Alquier, 2021) for a recent review.

**Contributions.** We advocate for a theoretically grounded strategy that uses generalization bound to improve $\pi_0$ with guarantees. So far, the existing bounds are either intractable (Swaminathan & Joachims, 2015b) or can be proven to be suboptimal (London & Sandler, 2019). In this work,

- we derive new, tractable and tight generalization bounds using the PAC-Bayesian framework. These bounds are fully tractable unlike (Swaminathan & Joachims, 2015b)'s bound and are tighter than (London & Sandler, 2019)'s bound.
- we provide a way to optimize our bounds over a particular class of policies and show empirically that they can guarantee improvement over $\pi_0$ in practical scenarios.

## 3. Motivating PAC-Bayesian tools

As previously discussed, in the contextual bandit setting, we seek a policy that minimizes the expected cost:

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

The minimizer of this objective over the unrestricted space of policies is a deterministic decision rule defined by:

$$\forall x, a \quad \pi^*(a|x) = \mathbb{1}[\operatorname*{argmin}_{a'} c(x,a') = a].$$

Given a context $x$, the solution will always choose the action that has the minimum cost. However, as the function $c$ is generally unknown, we instead learn a parametric score function $f_\theta \in \mathcal{F}_\Theta = \{f_\theta : \mathcal{X}\times[K] \to \mathbb{R}, \theta\in\Theta\}$ that encodes the action's relevance to a context $x$. Given a function $f_\theta$, we define the decision rule $d_\theta$ by:

$$d_\theta(a|x) = \mathbb{1}[\operatorname*{argmax}_{a'} f_\theta(x,a') = a].$$

These parametric decisions rules will be the building blocks of our analysis. We view stochastic policies as smoothed decision rules, with smoothness induced by distributions over the space of score functions $\mathcal{F}_\Theta$. Given a context $x$, instead of sampling an action $a$ directly from a distribution on the action set, we sample a function $f_\theta$ from a distribution over $\mathcal{F}_\Theta$ and compute the action as $a = \operatorname{argmax}_{a'} f_\theta(x,a')$. With this interpretation, for any distribution $Q$ on $\mathcal{F}_\Theta$, the probability of an action, $a\in\mathcal{A}$, given a context $x\in\mathcal{X}$, is defined as the expected value of $d_\theta$ over $Q$, that is:

$$\begin{aligned}\pi_Q(a|x) &= \mathbb{E}_{\theta\sim Q}\left[d_\theta(a|x)\right]\\ &= \mathbb{P}_Q\left(\operatorname*{argmax}_{a'} f_\theta(x,a') = a\right).\end{aligned}$$

**Policies as mixtures of decision rules.** This perspective on policies was introduced in (Seldin et al., 2011) and later developed in (London & Sandler, 2019). Constructing policies as *mixtures of deterministic decision rules* does not restrict the class of policies our study applies to. Indeed, if the family $\mathcal{F}_\Theta$ is rich enough (e.g, neural networks), we give the following theorem that proves that any policy $\pi$ can be written as a mixture of deterministic policies.

**Theorem 3.1.** *Let us fix a policy $\pi$. Then there is a probability distribution $Q_\pi$ on the set of all functions $f : \mathcal{X}\times[K] \to \{0,1\}$ such that*

$$\forall x, a \quad \pi(a|x) = \mathbb{E}_{f\sim Q_\pi}\left[\mathbb{1}\left[\operatorname*{argmax}_{a'} f(x,a') = a\right]\right].$$

A formal proof of Theorem 3.1 is given in Appendix A.1. This means that adopting this perspective on policies does not narrow the scope of our study. For a policy $\pi_Q$ defined by a distribution $Q$ over $\mathcal{F}_\Theta$, we observe that by linearity, its true risk can be written as:

$$R(\pi_Q) = \mathbb{E}_{\theta\sim Q}[R(d_\theta)].$$

Similarly, clipping the logging propensities in our empirical estimators allows us to obtain a linear estimators in $\pi$. For instance, we can estimate empirically the risk of the policy $\pi_Q$ with cvcIPS (as it generalizes cIPS) and obtain:

$$\hat{R}_n^{\tau,\xi}(\pi_Q) = \mathbb{E}_{\theta\sim Q}[\hat{R}_n^{\tau,\xi}(d_\theta)].$$

By linearity, one can see that both the true and empirical risk of a policy $\pi_Q$ can also be interpreted as the average risk of decision rules drawn from the distribution $Q$. This duality is in the heart of our analysis and paves the way nicely to the PAC-Bayesian framework, which studies generalization properties of the average risk of randomized predictors (Alquier, 2021). If we fix a reference distribution $P$ over $\mathcal{F}_\Theta$ and define the KL divergence from $P$ to $Q$ as:

$$\mathcal{KL}(Q||P) = \begin{cases}\displaystyle\int \ln\left\{\frac{dQ}{dP}\right\} dQ & \text{if } Q \text{ is } P\text{-continuous},\\ +\infty & \text{otherwise},\end{cases}$$

we can construct with the help of PAC-Bayesian tools bounds holding for the average risk of decision rules over any distribution $Q$;

$$\mathbb{E}_{\theta\sim Q}[R(d_\theta)] \le \mathbb{E}_{\theta\sim Q}[\hat{R}_n^{\tau,\xi}(d_\theta)] + \mathcal{O}\left(\mathcal{KL}(Q||P)\right).$$

Our objective will be to find tight generalisation bounds of this form as this construction, coupled with the linearity of our objective and estimator, allows us to obtain tight bounds holding for any policy $\pi_Q$;

$$R(\pi_Q) \le \hat{R}_n^{\tau,\xi}(\pi_Q) + \mathcal{O}\left(\mathcal{KL}(Q||P)\right).$$

**The PAC-Bayesian Paradigm.** Before we dive deeper into the analysis, we want to emphasize the similarities between the PAC-Bayesian paradigm and the offline contextual bandit problem. This learning framework proceeds as follows: Given a class of functions $\mathcal{F}_\Theta$, we fix a prior (reference distribution) $P$ on $\mathcal{F}_\Theta$ before seeing the data, then, we receive some data $\mathcal{D}_n$ which help us learn a better distribution $Q$ over $\mathcal{F}_\Theta$ than our reference $P$. With the previous perspective on policies, the prior $P$, even if it can be any data-free distribution, will be our logging policy (i.e. $\pi_0 = \pi_P$), and we will use the data $\mathcal{D}_n$ to learn distribution $Q$, thus a new policy $\pi_Q$ that improves the logging policy $\pi_0$.

## 4. PAC-Bayesian Analysis

### 4.1. Bounds for clipped IPS

The clipped IPS estimator (Bottou et al., 2013) is often studied for offline policy learning (Swaminathan & Joachims, 2015b; London & Sandler, 2019) as it is easy to analyse, and have a negative bias (once the cost is negative) facilitating the derivation of learning bounds.

(London & Sandler, 2019) adapted (McAllester, 2003)'s bound to derive their learning objective. We state a slightly tighter version in Proposition 4.1 for the cIPS estimator. The proof of this bound cannot be adapted naively to the cvcIPS estimator, because once $\xi \neq 0$, the bias of the estimator, an intractable quantity that depends on the unknown distribution $\nu$, is no longer negative and needs to be incorporated in the bound, making the bound itself intractable.

**Proposition 4.1.** *Given a prior $P$ on $\mathcal{F}_\Theta$, $\tau \in (0,1]$, $\delta \in (0,1]$, the following bound holds with probability at least $1-\delta$ uniformly for all distribution $Q$ over $\mathcal{F}_\Theta$:*

$$R(\pi_Q) \leq \hat{R}_n^\tau(\pi_Q) + \frac{2(\mathcal{KL}(Q||P) + \ln\frac{2\sqrt{n}}{\delta})}{\tau n} + \sqrt{\frac{2[\hat{R}_n^\tau(\pi_Q) + \frac{1}{\tau}](\mathcal{KL}(Q||P) + \ln\frac{2\sqrt{n}}{\delta})}{\tau n}}.$$

The upper bound stated in the previous proposition will be denoted by $\mathcal{LS}_n^{P,\delta,\tau}(\pi_Q)$. When there is no ambiguity, we will also drop $P, \delta, \tau$ (all fixed) and only use $\mathcal{LS}_n(\pi_Q)$.

(McAllester, 2003)'s bound can give tight results in the $[0,1]$-bounded loss case when the empirical risk is close to 0 as one obtains fast convergence rates in $O(1/n)$. However, its use in the case of offline contextual bandits is far from being optimal. Indeed, to achieve a fast rate in this context, one needs $\hat{R}_n^\tau(\pi_Q) + \frac{1}{\tau} \approx 0$. This is hardly achievable in practice especially when $n$ is large and $\tau \ll 1$.

To defend our claim, let us suppose that for each context $x$, there is one optimal action $a_x^*$ for which $c(x, a_x^*) = -1$ and it is 0 otherwise. Let us write down the clipped IPS:

$$\hat{R}_n^\tau(\pi) = \frac{1}{n}\sum_{i=1}^n \frac{\pi(a_i|x_i)}{\max\{\pi_0(a_i|x_i), \tau\}} c_i \geq -\frac{1}{\tau}.$$

To get equality, we need:

$$\forall i \in [n], \quad c_i = -1, \quad \pi_0(a_i|x_i) \leq \tau, \quad \pi(a_i|x_i) = 1.$$

If $n$ is large, the first condition on the costs means that $\pi_0$ is near optimal and the played actions $a_i$ are optimal. For this, we get that $\forall i \in [n], \pi_0(a_i|x_i) \approx 1$. This, combined with the second condition on $\pi_0$ gives that $\tau \approx 1$. In practice, $\pi_0$ is never the optimal policy and $\tau \ll 1$ which makes the fast rate condition $\hat{R}_n^\tau(\pi) + \frac{1}{\tau} \approx 0$ unachievable. In the majority of scenarios, as we penalize $\pi_Q$ to stay close to $\pi_0$ through the KL divergence, we will have $\hat{R}_n^\tau(\pi_Q) \in [-1, 0]$, thus $\hat{R}_n^\tau(\pi_Q) + \frac{1}{\tau} \approx \frac{1}{\tau}$, giving a limiting behavior:

$$\mathcal{LS}_n(\pi_Q) = \hat{R}_n^\tau(\pi_Q) + \mathcal{O}\left(\frac{1}{\tau}\sqrt{\frac{\mathcal{KL}(Q||P)}{n}}\right).$$

If we want to get tighter results, we need to look for bounds with better dependencies on $\tau$ and $n$. In our pursuit of a tighter bound, we derive the following result:

**Proposition 4.2.** *Given a prior $P$ on $\mathcal{F}_\Theta$, $\tau \in (0,1]$, $\delta \in (0,1]$. The following bound holds with probability at least $1-\delta$ uniformly for all distribution $Q$ over $\mathcal{F}_\Theta$:*

$$R(\pi_Q) \leq \min_{\lambda > 0} \frac{1 - \exp\left(-\tau\lambda\hat{R}_n^\tau(\pi_Q) - \frac{1}{n}\left[\mathcal{KL}(Q||P) + \ln\frac{2\sqrt{n}}{\delta}\right]\right)}{\tau(e^\lambda - 1)}$$

We will denote by $\mathcal{C}_n^{P,\delta,\tau}(\pi_Q)$ the upper bound stated in this proposition. When there is no ambiguity, we will also drop $P, \delta, \tau$ (all fixed) and only use $\mathcal{C}_n(\pi_Q)$.

This is a direct application of (Catoni, 2007)'s bound to the bounded loss cIPS while exploiting the fact that its bias is negative. A full derivation can be found in Appendix A.2. Note that Proposition 4.2 cannot be applied to the cvcIPS estimator ($\xi \neq 0$) as its bias is non-negative and intractable. To be able to measure the tightness of this bound, we estimate its limiting behaviour to understand its dependency on $\tau$ and $n$. We derive in Appendix A.3 the following result:

$$\mathcal{C}_n(\pi_Q) = \hat{R}_n^\tau(\pi_Q) + \mathcal{O}\left(\frac{\mathcal{KL}(Q||P)}{\tau n}\right).$$

This shows that $\mathcal{C}_n$ has a better dependency on $n$ compared to $\mathcal{LS}_n$. Actually, we can prove that $\mathcal{C}_n$ will always give tighter results as the next theorem states that it is smaller than $\mathcal{LS}_n$ in all scenarios.

**Theorem 4.3.** *For any $\mathcal{D}_n \sim (\mu, \pi_0)^n$, any distributions $P, Q$, any $\tau \in (0,1], \delta \in (0,1]$, we have:*

$$\mathcal{C}_n^{P,\delta,\tau}(\pi_Q) \leq \mathcal{LS}_n^{P,\delta,\tau}(\pi_Q).$$

One can refer to Appendix A.4 for the full proof. This result confirms that the bound given by Proposition 4.2 is *theoretically* tighter than the bound in Proposition 4.1, making $\mathcal{LS}_n$ unusable if we seek tight guarantees on $R(\pi_Q)$.

### 4.2. Going beyond clipped IPS

The cvcIPS estimator in Equation (2) generalizes cIPS, and can behave better as it achieves improved variance with a well chosen $\xi$. To study its learning properties, we derive a **novel** *Bernstein-type* PAC-Bayesian bound that holds for the cvcIPS estimator. Let $g$ be the function $g : u \to \frac{\exp(u)-1-u}{u^2}$. We also define the conditional bias $\mathcal{B}_n^\tau$ and the conditional second moment $\mathcal{V}_n^\tau$:

- $\mathcal{B}_n^\tau(\pi) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\pi(.|x_i)}\left[\mathbb{1}\left[\pi_0(a|x_i) < \tau\right]\left(1 - \frac{\pi_0(a|x_i)}{\tau}\right)\right]$
- $\mathcal{V}_n^\tau(\pi) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\pi(.|x_i)}\left[\frac{\pi_0(a|x_i)}{\max\{\pi_0(a|x_i), \tau\}^2}\right].$

With these definitions, we can state our proposition:

**Proposition 4.4.** *Given a prior $P$ on $\mathcal{F}_\Theta$, $\xi \in [-1, 0]$, $\tau \in (0, 1]$, $\delta \in (0, 1]$ and a set of strictly positive scalars $\Lambda = \{\lambda_i\}_{i\in[n_\Lambda]}$. The following bound holds with probability at least $1 - \delta$ uniformly for all distribution $Q$ over $\mathcal{F}_\Theta$:*

$$R(\pi_Q) \leq \hat{R}_n^{\tau,\xi}(\pi_Q) - \xi\mathcal{B}_n^\tau(\pi_Q) + \sqrt{\frac{\mathcal{KL}(Q||P) + \ln\frac{4\sqrt{n}}{\delta}}{2n}}$$
$$+ \min_{\lambda\in\Lambda}\left\{\frac{\mathcal{KL}(Q||P) + \ln\frac{2n_\Lambda}{\delta}}{\lambda n} + \lambda l_\xi g\left(\lambda b_\xi\right)\mathcal{V}_n^\tau(\pi_Q)\right\}$$

*with $l_\xi = \max\left[\xi^2, (1+\xi)^2\right]$, $b_\xi = (1+\xi)/\tau - \xi$.*

The choice of $\Lambda$ as well as the full proof of a more general version of Proposition 4.4 can be found in Appendix A.5. The upper bound given by Proposition 4.4 (with $P, \tau, \delta$ fixed) will be denoted by $\mathcal{CBB}_n^\xi(\pi_Q)$.

Proposition 4.4 covers the general cvcIPS estimator ($\xi \neq 0$) and its objective is decomposable into a sum, making it amenable to stochastic first order optimization (Robbins & Monro, 1951). However, minimizing it requires access to the logging policy $\pi_0$. This is reasonable as $\pi_0$ represents the currently deployed decision system, which we want to improve. In the limit, the bound is estimated to behave like:

$$\mathcal{CBB}_n^\xi(\pi_Q) = \hat{R}_n^{\tau,\xi}(\pi_Q) - \xi\mathcal{B}_n^\tau(\pi_Q)$$
$$+ \mathcal{O}\left(\left(\frac{1}{2\sqrt{2}} + \sqrt{l_\xi\mathcal{V}_n^\tau(\pi_Q)}\right)\sqrt{\frac{\mathcal{KL}(Q||P)}{n}}\right)$$

A derivation of this result can be found in Appendix A.5.3. We have $\mathcal{V}_n^\tau \leq 1/\tau$ and expect this bound to give tight results when $\mathcal{V}_n^\tau(\pi_Q) \ll 1/\tau$. However, once $\pi_0$ is uniform and $\tau = 1/K$, we can never have the previous condition as:

$$\forall\pi, \quad \mathcal{V}_n^\tau(\pi) = 1/\tau.$$

This means that the worst regime for $\mathcal{CBB}_n^\xi$ is when $\pi_0$ is uniform, and even in that case, this bound should be tighter than $\mathcal{LS}_n$ as it has a better dependency on $\tau$. We can also get an intuition about the impact of $\xi$, in particular:

- $\xi = 0$ recovers cIPS. This estimator has the best dependency on the bias (it nullifies the impact of $\mathcal{B}_n^\tau(\pi_Q)$) and the worst dependency on $\mathcal{V}_n^\tau(\pi_Q)$ as $l_\xi = 1$.
- $\xi = -0.5$ obtains the best dependency on $\mathcal{V}_n^\tau(\pi_Q)$ as $l_\xi$ reaches its minimum and make the bound suffer only half the bias $\mathcal{B}_n^\tau(\pi_Q)$.
- $\xi = -1$ in the other hand, has both the worst dependencies on the bias and the variance, and can be considered a bad choice. Actually, this value of $\xi$ shifts the costs to be always positive ($\forall i, c_i - \xi \geq 0$), which is known to make the off-policy risk estimators not suited for policy learning (Swaminathan & Joachims, 2015b).

These observations point to the importance of the choice of $\xi$, which can drastically change the behavior of the bound. We will empirically study the impact of two candidate values of $\xi \in \{0, -0.5\}$ on the tightness of the bound.

## 5. Restricting the Space of Policies

Our PAC-Bayesian bounds hold for any policy $\pi_Q$. However, to obtain policies of practical use, we ask for some desired properties that are summarized in the points below.

**Sampling.** Being able to efficiently sample actions from our policy is crucial as the decisions taken by our online system boil down to sampling. For a given context $x$, we have:

$$a \sim \pi_Q(\cdot|x) \iff a = \operatorname*{argmax}_{a'} f_\theta(x, a'), \theta \sim Q.$$

The complexity of sampling from $\pi_Q$ depends on how easy it is to sample from the distribution $Q$.

**Computing propensities.** Computing propensities for a given pair $(x, a)$ is essential for off-policy evaluation. A generic estimate can be obtained by:

$$\hat{\pi}_Q^{\text{naive}}(a|x) = \frac{1}{S}\sum_{i=1}^{S} d_{\theta_i}(a|x)$$

with $\{\theta_i\}_{i=1}^S$ samples from $Q$. This estimator will behave badly once we deal with large action spaces and/or high-dimensional distributions parameters. Ideally, we would like to exploit the family of distributions $Q$ considered and

the form of the function $f_\theta$ to come up with a better behaved estimator for the propensities.

**Numerical optimization.** If we restrict our study to a parametric family $\mathcal{Q}_\Psi = \{Q_\psi, \psi \in \Psi\}$, computing gradients will be essential to minimising the bounds. For a given pair $(x, a)$, we can compute for any parameterized distribution $Q_\psi$ the score function gradient estimator (Williams, 1992) of $\pi_{Q_\psi}(a|x)$:

$$\begin{aligned}\nabla_\psi \pi_{Q_\psi}(a|x) &= \nabla_\psi \mathbb{E}_{\theta\sim Q_\psi}[d_\theta(a|x)] \\ &= \mathbb{E}_{\theta\sim Q_\psi}[d_\theta(a|x)\nabla_\psi \log Q_\psi(\theta)].\end{aligned}$$

This gradient suffers from a variance problem (Xu et al., 2019) and we might need to choose a specific family of distributions $\mathcal{Q}_\Psi$, or/and specify a form of $f_\theta$ to obtain $\psi \to \pi_{Q_\psi}(a|x)$ with better behaved gradients.

(London & Sandler, 2019) restricted their study to Mixed Logit policies (Hensher & Greene, 2003). These policies are easy to sample from, and have easy to compute propensities and gradients. However, their learning properties are demonstrated to be sub-optimal (Mei et al., 2020). To this end, we adopt another class of policies that we deem better behaved for our objective. We discuss the reasons behind this choice in detail in Appendix A.7.

### 5.1. Linear Independent Gaussian Policies

As mentioned previously, even if our analysis is valid for all distributions $Q$ and any form of $f_\theta$, we need to restrict our space to obtain practical policies. We restrict $f_\theta$ to:

$$\forall x, a \quad f_\theta(x, a) = \phi(x)^T \theta_a \tag{5}$$

with $\phi$ a fixed transform[2] over the contexts. This form of $f_\theta$ is widely used in this context (Faury et al., 2020; Swaminathan & Joachims, 2015b). This results in a parameter $\theta$ 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.

With these choices of $f_\theta$ and $\mathcal{Q}$, the induced $\pi_{\boldsymbol{\mu},\sigma}$, that we call **LIG: Linear Independent Gaussian** policies, will provide fast sampling and easily computable propensities and gradients. Indeed, sampling from $\pi_{\boldsymbol{\mu},\sigma}$ will reduce to sampling from a normal distribution $\theta \sim Q_{\boldsymbol{\mu},\sigma}$ and computing $a = \text{argmax}_{a'} f_\theta(x, a')$. When it comes to estimating the propensity of $a$ given $x$, we can suggest another expression of $\pi_{\boldsymbol{\mu},\sigma}(a|x)$ that reduces the computation to a one dimensional integral:

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

with $\Phi$ the cumulative distribution function of the standard normal. See Appendix A.6 for a full derivation. The computation of $\pi_{\boldsymbol{\mu},\sigma}(a|x)$ becomes easier as one dimensional standard normal integrals can be well approximated. The gradient can also be derived from this new expression:

$$\nabla_{\boldsymbol{\mu},\sigma}\pi_{\boldsymbol{\mu},\sigma}(a|x) = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1)}\left[\nabla_{\mu,\sigma} G_{\mu,\sigma}(\epsilon, a, x)\right]$$

which can be seen as a one dimensional reparametrization trick gradient, and is known to behave better than the score function gradient estimator (Xu et al., 2019).

**Optimising the bounds.** For their practicality, we focus on the class of **LIG** policies to optimise the bounds. As these policies are built with Gaussian distributions $Q$, we also adopt Gaussian priors[3] $P = \mathcal{N}(\boldsymbol{\mu_0}, \sigma_0 I_d)$ to obtain an analytical expression for $\mathcal{KL}(Q||P)$. We state the bounds for **LIG** policies with Gaussian priors in Appendix A.8.

Optimizing for **LIG** policies, the best guaranteed risk defined in Section 2.3 and the minimizer $\pi^*_{\mathcal{UB}_n}$ for the different bounds $\mathcal{UB}_n \in \{\mathcal{LS}_n, \mathcal{C}_n, \mathcal{CBB}^\xi_n\}$ are given by:

$$\mathcal{GR}^*_{\mathcal{UB}_n} = \min_{\pi_{\boldsymbol{\mu},\sigma}} \mathcal{UB}_n(\pi_{\boldsymbol{\mu},\sigma}) \tag{6}$$

$$\pi^*_{\mathcal{UB}_n} = \arg\min_{\pi_{\boldsymbol{\mu},\sigma}} \mathcal{UB}_n(\pi_{\boldsymbol{\mu},\sigma}). \tag{7}$$

The best guaranteed improvement $\mathcal{GI}^*(\pi_0)$ with these bounds follows as the difference between $R(\pi_0)$ and $\mathcal{GR}^*$.

## 6. Experiments

We are interested in studying the effectiveness of the proposed bounds in providing guaranteed improvement of the previously deployed system $\pi_0$ after collecting $\mathcal{D}_n$.

**Experimental Setup.** For ease of exposition, we use a softmax logging policy of parameter $\mu_0 \in \mathbb{R}^{p\times K}$:

$$\pi^{\mathcal{S}}_{\boldsymbol{\mu_0}}(a|x) \propto \exp(\phi(x)^T \boldsymbol{\mu_0}_a).$$

$\mu_{0_a} \in \mathbb{R}^p$ is the parameter associated with action $a$. The policy $\pi^{\mathcal{S}}_{\boldsymbol{\mu_0}}$ is used to generate the logged interactions data $\mathcal{D}_n$ and its parameter $\boldsymbol{\mu_0}$ constructs the reference distribution $P = \mathcal{N}(\boldsymbol{\mu_0}, I_d)$ for all bounds.

We adopt the standard supervised-to-bandit conversion to generate logged data in all of our experiments (Swaminathan & Joachims, 2015b). We use two mutliclass datasets: **FashionMNIST** (Xiao et al., 2017) and **EMNIST-b** (Cohen et al.,

[2]Even if we assume that $\phi$ is fixed, the analysis can be naturally extended to the more general case where $\phi_\psi$ is a parameterized neural network that we learn.

[3]The prior uses the parameters of the logging policy $\pi_0$.

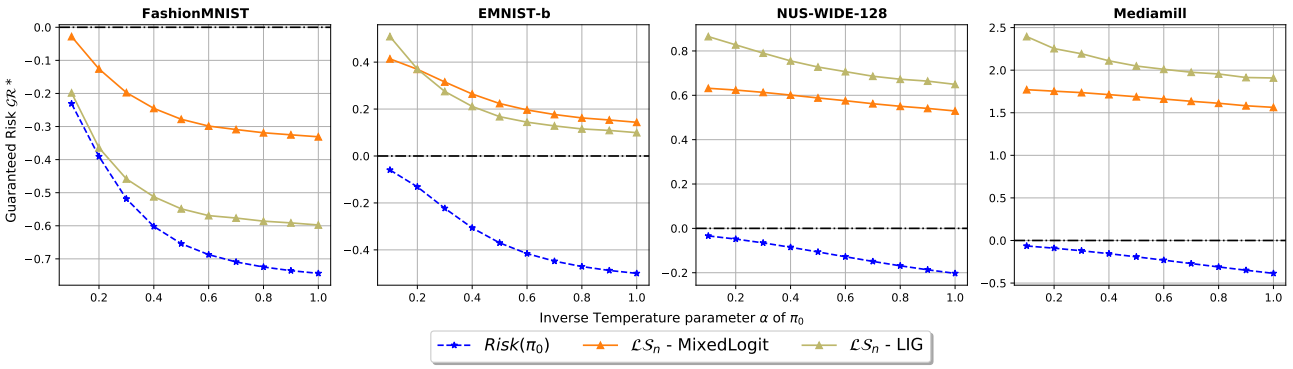


Figure 1: The Guaranteed Risk given by $\mathcal{LS}_n$ optimized over **Mixed Logit** and **LIG** policy classes while changing $\pi_0$. The $\mathcal{LS}_n$ bound fails to guarantee improvement ($\mathcal{GR}^*_{\mathcal{LS}_n} > R(\pi_0)$) in all scenarios considered.

2017), alongside two multilabel datasets: **NUS-WIDE-128** (Chua et al., 2009) with 128-VLAD features (Spyromitros-Xioufis et al., 2014) and **Mediamill** (Snoek et al., 2006) to empirically validate our findings. The statistics of the datasets are described in Table 1 in Appendix B.1; $N$ the size of the training split, $K$ the number of actions and $p$ the dimension of the features $\phi(x)$. We take a small fraction (5%) of the training data that will only be used to learn $\boldsymbol{\mu_0}$ in a supervised manner.

With $\boldsymbol{\mu_0}$ obtained, we introduce an inverse temperature parameter $\alpha$ to our softmax logging policy $\pi^{\mathcal{S}}_{\alpha\boldsymbol{\mu_0}}$ giving a prior $P = \mathcal{N}(\alpha\boldsymbol{\mu_0}, I_d)$. Changing $\alpha$ allows us to cover logging policies with different entropies ($\alpha \approx 0$ gives a uniform $\pi_0$ and $\alpha \approx 1$ gives a peaked $\pi_0$). We run $\pi_{\alpha\boldsymbol{\mu_0}}$ on the rest ($n_c = 0.95N$) of the training data to generate $\mathcal{D}_n$. For a context $x$ and an action $a$, we define in our setting the cost as $c = -\mathbb{1}[a \in y]$ with $y$ the set of true labels for $x$.

Learning $\boldsymbol{\mu_0}$ on a split different than the one logged allows us to use the previous bounds as the parameter $\boldsymbol{\mu_0}$ does not depend on the logged interactions, making the reference distribution $P$ data-free.

We set the allowed uncertainty to $\delta = 0.05$. For all datasets, $\tau$ will be set to $\tau = 1/K$ to get no bias when the logging policy is close to uniform. We use Adam (Kingma & Ba, 2014) with a learning rate of $10^{-3}$ for 100 epochs to optimize the bounds w.r.t their parameters. More details on the training procedure can be found in Appendix B.2.

**$\mathcal{LS}_n$ is not tight enough.** We demonstrated in this work that the $\mathcal{LS}_n$ bound used by (London & Sandler, 2019) is suboptimal as it generally shows a worse dependence on $\tau$ and can be shown to be theoretically dominated by $\mathcal{C}_n$ (Theorem 4.3). However, we want to verify if it can guarantee the improvement of the logging policy $\pi_0$. Given logged data $\mathcal{D}_n$ generated by $\pi_0$, we optimize the $\mathcal{LS}_n$ bound with both the **LIG** (Equation (6)) and **Mixed Logit** (Theorem 3 in (London & Sandler, 2019)) policy classes.

In Figure 1, we plot $R(\pi_0)$, the true risk of $\pi_0$ alongside the best guaranteed risk $\mathcal{GR}^*_{\mathcal{LS}_n}$ given by the two policy classes, while changing the logging policies $\pi_0$ (going from uniform to peaked policies by changing $\alpha$).

We can observe, for the two policy classes and all scenarios considered, that this bound fails at providing guaranteed improvement as its guaranteed risk $\mathcal{GR}^*_{\mathcal{LS}_n}$ is always smaller than $R(\pi_0)$ and is vacuous ($\mathcal{GR}^*_{\mathcal{LS}_n} \geq 0$) for some datasets. This bound is not tight enough to be used with our strategy.

**$\mathcal{C}_n$ and $\mathcal{CBB}^{\xi}_n$ do guarantee improvement.** Given logged data $\mathcal{D}_n$, we optimize $\mathcal{C}_n$ and $\mathcal{CBB}^{\xi}_n$ for **LIG** policies (Equation (6)) and plot in Figure 2 the guaranteed risk $\mathcal{GR}^*$, the true risk of the minimizer $R(\pi^*)$ as well as the positive guaranteed improvement by the bound $\max(\mathcal{GI}^*, 0)$.

To answer our question, we are interested in the first row (best guaranteed risk $\mathcal{GR}^*$) and last row (best guaranteed improvement $\mathcal{GI}^*$) of Figure 2. For the $\mathcal{CBB}^{\xi}_n$ bound, we study particularly the two values of $\xi \in \{0, -\frac{1}{2}\}$.

We can observe that contrary to $\mathcal{LS}_n$, our bounds can guarantee improvement over $\pi_0$ in the majority of scenarios.

The $\mathcal{C}_n$ bound gives great results when $\pi_0$ is close to uniform ($\alpha \to 0$) but fails sometimes (when $\alpha \to 1$) at improving the logging policy and one can observe that in the context of **FashionMNIST** and **EMNIST-b**.

As for the $\mathcal{CBB}^{\xi}_n$ bound, we can observe that choosing $\xi = -\frac{1}{2}$ consistently give the best results as it reduces considerably the dependency on $\mathcal{V}^{\tau}_n$. Note that $\mathcal{CBB}^{\xi}$ with $\xi = -\frac{1}{2}$ never fails to produce guaranteed improvement across all settings. Having a uniform logging policy $\pi_0$ is the worst regime for this bound as $\mathcal{V}^{\tau}_n$ reaches its highest value $1/\tau$. This is empirically confirmed with our experiments. We can see in Figure 2 that, for small values of $\alpha$, the $\mathcal{CBB}^{\xi}_n$ bound suffers the most; $\mathcal{CBB}_n(\xi = 0)$ is always worse than $\mathcal{C}_n$ and $\mathcal{CBB}_n(\xi = -\frac{1}{2})$ produces worse guarantees than $\mathcal{C}_n$ in the **NUS-WIDE-128** dataset. Once we

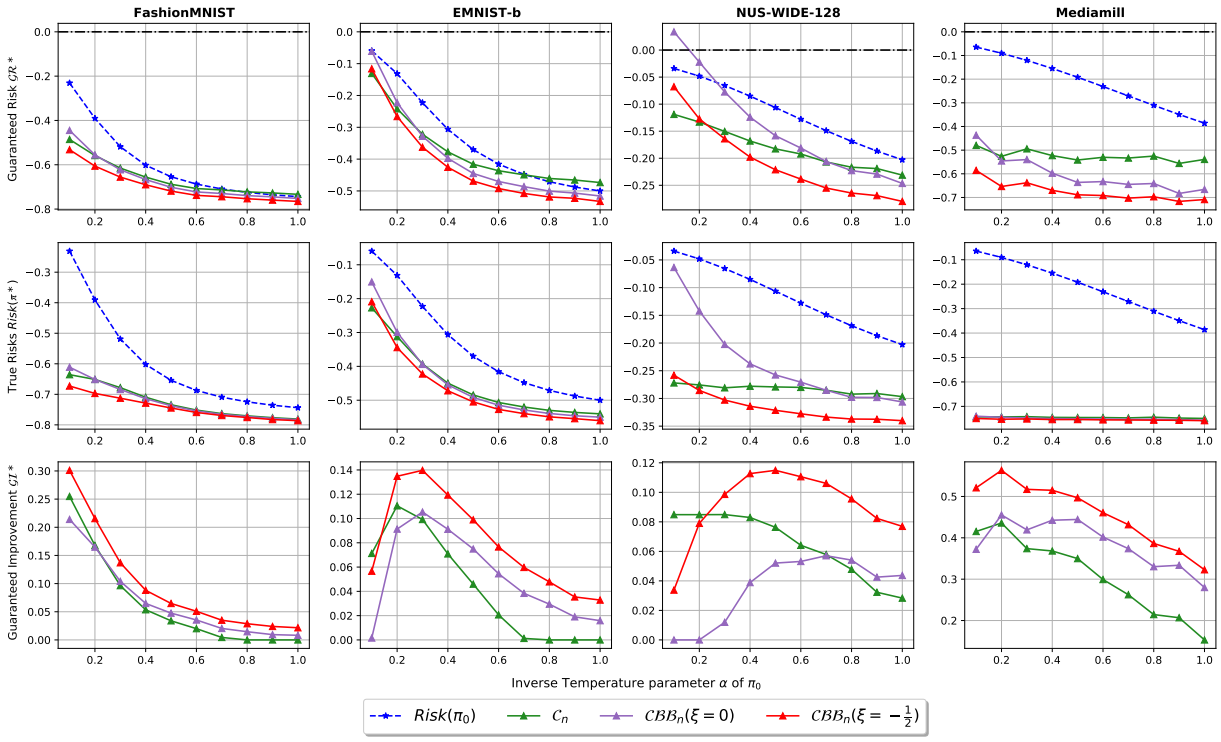


Figure 2: Behavior of the guaranteed risk $\mathcal{GR}^*$ (↓ is better), the risk of the minimizer $R(\pi^*)$ (↓ is better) and the guaranteed improvement $\mathcal{GI}^*$ (↑ is better) given by the bounds (optimized with **LIG** policies) while changing $\pi_0$. We can observe that the newly proposed bounds can efficiently improve on $\pi_0$, with $\mathcal{CBB}_n^{\xi}$ ($\xi = -\frac{1}{2}$) giving the best results.

drift away from uniform logging policies, the $\mathcal{CBB}^{\xi}$ bound, especially with $\xi = -\frac{1}{2}$ gives the best guarantees on all the datasets considered.

**Tighter bounds give the best true risk.** In real-world problems, we cannot have access to $R(\pi^*)$ before deployment. In our experiments, we can compute $R(\pi^*)$ on the test sets as we have access to the true labels[4]. We are interested in this quantity as we want to make sure that the bounds giving low guaranteed risk $\mathcal{GR}^*$ will produce policies $\pi^*$ with low true risk $R(\pi^*)$. Even if the limiting behaviors of the bounds give an intuition of how they compare, this will further confirm that the gap between the bounds is not linked to constants but to quantities valuable to learning $\pi^*$. The second row on Figure 2 confirms that the bounds with the best $\mathcal{GR}^*$ reliably give the best $R(\pi^*)$ in all settings.

**Take away.** These experiments confirm that the policies $\pi^*$ obtained by optimising our newly proposed bounds improve, with high confidence, the logging policies $\pi_0$. The results also suggest the use of the variance sensitive bound $\mathcal{CBB}_n(\xi = -1/2)$ for its consistent results across the different scenarios. However, if computing expectations under $\pi_0$ is difficult, one can adopt $\mathcal{C}_n$ as it showed great results.

[4] up to a small $\mathcal{O}(1/\sqrt{n_t})$ approximation error .

## 7. Conclusion

In this work, we introduce a new theoretically grounded strategy for offline policy optimization. This approach is based on generalization bounds, uses all the available data and does not require additional hyperparameters. Leveraging PAC-Bayesian tools, we provide novel generalization bounds tight enough to make our strategy viable, giving practitioners a principled way to confidently improve over the previous decision system offline. Our results can nicely be extended to learning efficient policies over slates (Swaminathan et al., 2017) or continuous action policies (Kallus & Zhou, 2018). We believe that our work brings us closer to offline policy learning with online performance certificates. In the future, we would like to investigate tighter bounds for this problem and loosen our assumptions; e.g. to remove the need for having access to the logging policy $\pi_0$.

## 8. Acknowledgments

The authors are grateful for all the fruitful discussions engaged with David Rohde, Louis Faury and Imad Aouali which helped improve this work. We would also want to thank the reviewers for their valuable feedback.

## References


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

Ben-Tal, A., Den Hertog, D., De Waegenaere, A., Melenberg, B., and Rennen, G. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

Bottou, L., Peters, J., Quiñonero-Candela, J., Charles, D. X., Chickering, D. M., Portugaly, E., Ray, D., Simard, P., and Snelson, E. Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising. *Journal of Machine Learning Research*, 14(65):3207–3260, 2013. URL http://jmlr.org/papers/v14/bottou13a.html.

Brandfonbrener, D., Whitney, W., Ranganath, R., and Bruna, J. Offline contextual bandits with overparameterized models. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1049–1058. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/brandfonbrener21a.html.

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

Chen, M., Gummadi, R., Harris, C., and Schuurmans, D. Surrogate objectives for batch policy optimization in one-step decision making. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/84899ae725ba49884f4c85c086f1b340-Paper.pdf.

Chua, T.-S., Tang, J., Hong, R., Li, H., Luo, Z., and Zheng, Y. Nus-wide: A real-world web image database from national university of singapore. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, CIVR '09, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605584805. doi: 10.1145/1646396.1646452. URL https://doi.org/10.1145/1646396.1646452.

Cohen, G., Afshar, S., Tapson, J., and van Schaik, A. EMNIST: an extension of MNIST to handwritten letters. *arXiv preprint arXiv:1702.05373*, 2017.

Dziugaite, G. K. and Roy, D. M. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2017.

Farajtabar, M., Chow, Y., and Ghavamzadeh, M. More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, pp. 1447–1456. PMLR, 2018.

Faury, L., Tanielan, U., Vasile, F., Smirnova, E., and Dohmatob, E. Distributionally Robust Counterfactual Risk Minimization. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

Hensher, D. A. and Greene, W. H. The mixed logit model: The state of practice. *Transportation*, 30(2):133–176, 2003. doi: 10.1023/A:1022558715350. URL https://doi.org/10.1023/A:1022558715350.

Horvitz, D. G. and Thompson, D. J. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. ISSN 01621459. URL http://www.jstor.org/stable/2280784.

Ionides, E. L. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008. doi: 10.1198/106186008X320456. URL https://doi.org/10.1198/106186008X320456.

Jeunen, O. and Goethals, B. Pessimistic reward models for off-policy learning in recommendation. In *Proceedings of the 15th ACM Conference on Recommender Systems*, RecSys '21, pp. 63–74, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384582. doi: 10.1145/3460231.3474247. URL https://doi.org/10.1145/3460231.3474247.

Joachims, T., Swaminathan, A., and de Rijke, M. Deep learning with logged bandit feedback. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=SJaP_-xAb.

Kallus, N. and Zhou, A. Policy evaluation and optimization with continuous treatments. In *AISTATS*, 2018.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kuzborskij, I., Vernade, C., Gyorgy, A., and Szepesvari, C. Confident off-policy evaluation and selection through self-normalized importance weighting. In Banerjee, A. and Fukumizu, K. (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine*

*Learning Research*, pp. 640–648. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/kuzborskij21a.html.

Lattimore, T. and Szepesvári, C. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401.

Letarte, G., Germain, P., Guedj, B., and Laviolette, F. Dichotomize and generalize: PAC-Bayesian binary activated deep neural networks. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/7ec3b3cf674f4f1d23e9d30c89426cce-Paper.pdf.

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

Ma, Y., Wang, Y.-X., and Narayanaswamy, B. Imitation-regularized offline learning. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2956–2965. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/ma19b.html.

Maurer, A. and Pontil, M. Empirical Bernstein bounds and sample-variance penalization. In *COLT*, 2009.

McAllester, D. Simplified PAC-Bayesian margin bounds. In Schölkopf, B. and Warmuth, M. K. (eds.), *Learning Theory and Kernel Machines*, pp. 203–215, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg. ISBN 978-3-540-45167-9.

McAllester, D. A. Some PAC-Bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, pp. 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.

McDiarmid, C. Concentration. In *Probabilistic methods for algorithmic discrete mathematics*, pp. 195–248. Springer, 1998.

Mei, J., Xiao, C., Dai, B., Li, L., Szepesvari, C., and Schuurmans, D. Escaping the gravitational pull of softmax. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21130–21140. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f1cf2a082126bf02de0b307778ce73a7-Paper.pdf.

Nguyen-Tang, T., Gupta, S., Nguyen, A. T., and Venkatesh, S. Offline neural contextual bandits: Pessimism, optimization and generalization. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=sPIFuucA3F.

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

Robbins, H. and Monro, S. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407, 1951. doi: 10.1214/aoms/1177729586. URL https://doi.org/10.1214/aoms/1177729586.

Sakhi, O., Bonner, S., Rohde, D., and Vasile, F. 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*, KDD '20, pp. 783–793, New York, NY, USA, 2020a. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3403121. URL https://doi.org/10.1145/3394486.3403121.

Sakhi, O., Faury, L., and Vasile, F. Improving offline contextual bandits with distributional robustness. *arXiv preprint arXiv:2011.06835*, 2020b.

Seldin, Y., Auer, P., Shawe-taylor, J., Ortner, R., and Laviolette, F. PAC-Bayesian analysis of contextual bandits. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper/2011/file/58e4d44e550d0f7ee0a23d6b02d9b0db-Paper.pdf.

Seldin, Y., Cesa-Bianchi, N., Auer, P., Laviolette, F., and Shawe-Taylor, J. Pac-bayes-bernstein inequality for martingales and its application to multiarmed bandits. In Glowacka, D., Dorard, L., and Shawe-Taylor, J.

(eds.), *Proceedings of the Workshop on On-line Trading of Exploration and Exploitation 2*, volume 26 of *Proceedings of Machine Learning Research*, pp. 98–111, Bellevue, Washington, USA, 02 Jul 2012. PMLR. URL https://proceedings.mlr.press/v26/seldin12a.html.

Snoek, C. G. M., Worring, M., van Gemert, J. C., Geusebroek, J.-M., and Smeulders, A. W. M. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proceedings of the 14th ACM International Conference on Multimedia*, MM '06, pp. 421–430, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595934472. doi: 10.1145/1180639.1180727. URL https://doi.org/10.1145/1180639.1180727.

Spyromitros-Xioufis, E., Papadopoulos, S., Kompatsiaris, I. Y., Tsoumakas, G., and Vlahavas, I. A comprehensive study over VLAD and product quantization in large-scale image retrieval. *IEEE Transactions on Multimedia*, 16(6): 1713–1728, 2014. doi: 10.1109/TMM.2014.2329648.

Swaminathan, A. and Joachims, T. The Self-Normalized Estimator for Counterfactual Learning. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pp. 3231–3239, Cambridge, MA, USA, 2015a. MIT Press.

Swaminathan, A. and Joachims, T. Batch Learning from Logged Bandit Feedback through Counterfactual Risk Minimization. *Journal of Machine Learning Research*, 16(52):1731–1755, 2015b. URL http://jmlr.org/papers/v16/swaminathan15a.html.

Swaminathan, A., Krishnamurthy, A., Agarwal, A., Dudik, M., Langford, J., Jose, D., and Zitouni, I. Off-policy evaluation for slate recommendation. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/5352696a9ca3397beb79f116f3a33991-Paper.pdf.

Valko, M., Munos, R., Kveton, B., and Kocák, T. Spectral Bandits for Smooth Graph Functions. In *International Conference on Machine Learning*, pp. 46–54, 2014.

Villar, S. S., Bowden, J., and Wason, J. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199, 2015.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, 1992. doi: 10.1007/BF00992696.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Xu, M., Quiroz, M., Kohn, R., and Sisson, S. A. Variance reduction properties of the reparameterization trick. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2711–2720. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/xu19a.html.

# A. DISCUSSIONS AND PROOFS

## A.1. Policies as mixtures of deterministic decision rules

As described in the paper, a policy $\pi$ takes a context $x \in \mathcal{X}$ and defines a probability distribution over the $K$-dimensional simplex $\Delta_K$. In our work, we reinterpret policies as mixtures of deterministic decision rules.

Let $f$ be the function that encodes the relevance of the action to the context $x$. Given a distribution $Q$ over the functions $f \in \mathcal{F}_\Theta = \{f_\theta, \theta \in \Theta\}$, we define a policy as:

$$\forall x, a \quad \pi_Q(a|x) = \mathbb{E}_{f\sim Q}\left[\mathbb{1}\left[\operatorname*{argmax}_{a'} f(x, a') = a\right]\right].$$

A natural question is: can any policy $\pi$ be written in this form?

In general, the answer depends on the set $\mathcal{F}_\Theta = \{f_\theta, \theta \in \Theta\}$ we are considering. When the class $\mathcal{F}_\Theta$ is rich enough, answer is yes, as proven by the following theorem.

**Theorem A.1.** *Let us fix a policy $\pi$. Let*

$$\mathcal{G} = \{g : \mathcal{X} \times \mathcal{A} \to \{0, 1\} \textit{ such that } \forall x, \exists! a, g(x, a) = 1\}.$$

*Then, there is a $\sigma$-algebra $\mathcal{S}$ on $\mathcal{G}$ and a probability distribution $Q_\pi$ on $(\mathcal{G}, \mathcal{S})$ such that*

$$\forall x, a \quad \pi(a|x) = \mathbb{E}_{f\sim Q_\pi}\left[\mathbb{1}\left[\operatorname*{argmax}_{a'} f(x, a') = a\right]\right].$$

**Proof:** Fix a policy $\pi$. Define the set $\Omega = [K]^{\mathcal{X}}$. That is, an element $\omega$ of $\Omega$ is a family of elements of $[K]$ indexed by $\mathcal{X}$: $\omega = (\omega_x)_{x\in\mathcal{X}}$. Define the set of cylinders

$$\mathcal{C} = \left\{A \subset \Omega : A = \prod_{x\in\mathcal{X}} A_x \text{ and } \operatorname{card}(\{x : A_x \neq [K]\}) < \infty\right\}.$$

For such a set $A = \prod_{x\in\mathcal{X}} A_x$ we define

$$P_\pi(A) = \prod_{x:A_x\neq\Omega}\left[\sum_{a\in A_x} \pi(a|x)\right].$$

Note in particular that, for a fixed $x \in \mathcal{X}$ and $a \in [K]$, we have

$$P_\pi(\{\omega \in \Omega : \omega_x = a\}) = \pi(a|x). \tag{8}$$

Then, Kolmogorov extension theorem guarantees that there is a unique extension of $P_\pi$ to the $\sigma$-field $\mathcal{D}$ generated by $\mathcal{C}$, that is $\mathcal{D} = \sigma(\mathcal{C})$. We have thus built a probability space $(\Omega, \mathcal{D}, P_\pi)$.

Now, for any $\omega = (\omega_x)_{x\in\mathcal{X}}$, we define the function $f_\omega : \mathcal{X} \times \mathcal{A} \to \{0, 1\}$ by $f_\omega(x, a) = \mathbb{1}[\omega_x = a]$. Define, for any $C \in \mathcal{D}$, $S_C := \{f_\omega, \omega \in C\}$ and $Q_\pi(S_C) = P_\pi(C)$, and finally put $\mathcal{S} = \{S_C, C \in \mathcal{D}\}$. As the function $F : \omega \mapsto f_\omega$ is a bijection from $\Omega$ to $\mathcal{G}$, $\mathcal{S}$ is a $\sigma$-field and $Q_\pi$ is a probability distribution. We have thus equiped $\mathcal{G}$ with a $\sigma$-field $\mathcal{S}$ and a probability $Q_\pi$: $(\mathcal{G}, \mathcal{S}, Q_\pi)$ is a probability space.

Now, we check that

$$\begin{aligned}
\mathbb{E}_{f\sim Q_\pi}\left[\mathbb{1}\left[\operatorname*{argmax}_{a'} f(x, a') = a\right]\right] &= Q_\pi\left(\left\{f \in \mathcal{G} : \operatorname*{argmax}_{a'} f(x, a') = a\right\}\right) \\
&= P_\pi\left(\left\{\omega \in \Omega : \operatorname*{argmax}_{a'} f_\omega(x, a') = a\right\}\right) \\
&= P_\pi\left(\{\omega \in \Omega : \omega_x = a\}\right) \\
&= \pi(a|x)
\end{aligned}$$

thanks to (8). This ends the proof.

### A.2. Proof of Proposition 1

Proposition 4.2 is a direct application of (Catoni, 2007)'s bound (see Theorem 3 in (Letarte et al., 2019)) to the rescaled cIPS $0 \leq 1 + \tau \cdot \hat{R}_n^\tau(\cdot) \leq 1$ with deterministic decision functions $d_\theta$. Let us fix a prior $P$ over $\mathcal{F}_\Theta$ and $\tau \in (0, 1]$. For any $\delta \in (0, 1]$, we have with probability at least $1 - \delta$ over draws of $\mathcal{D}_n \sim (\nu, \pi_0)^n$: for any $Q$ that is $P$-continuous, any $\lambda > 0$:

$$1 + \tau \cdot \mathbb{E}_{\theta \sim Q} \left[ \mathbb{E}_{(\nu, \pi_0)} \left[ \hat{R}_n^\tau(d_\theta) \right] \right] \leq \frac{1}{(1 - e^{-\lambda})} \left( 1 - \exp \left[ -\lambda \cdot (1 + \tau \cdot \mathbb{E}_{\theta \sim Q}[\hat{R}_n^\tau(d_\theta)]) - \frac{\mathcal{KL}(Q||P) + \ln \frac{2\sqrt{n}}{\delta}}{n} \right] \right)$$

with $KL[Q||P] = \mathbb{E}_Q [\ln Q/P]$.

By linearity of the expectation and $\hat{R}_n^\tau(\cdot)$, we get:

$$1 + \tau \cdot \mathbb{E}_{(\nu, \pi_0)} \left[ \hat{R}_n^\tau(\pi_Q) \right] \leq \frac{1}{(1 - e^{-\lambda})} \left( 1 - \exp \left[ -\lambda \cdot (1 + \tau \cdot \hat{R}_n^\tau(\pi_Q)) - \frac{KL[Q||P] + \ln \frac{2\sqrt{n}}{\delta}}{n} \right] \right).$$

Rearranging the terms gives:

$$\begin{aligned} \mathbb{E}_{(\nu, \pi_0)} \left[ \hat{R}_n^\tau(\pi_Q) \right] &\leq \frac{e^{-\lambda}}{\tau(1 - e^{-\lambda})} \left( 1 - \exp \left[ -\lambda \cdot \tau \cdot \hat{R}_n^\tau(\pi_Q) - \frac{KL[Q||P] + \ln \frac{2\sqrt{n}}{\delta}}{n} \right] \right) \\ &\leq \frac{1}{\tau(e^{\lambda} - 1)} \left( 1 - \exp \left[ -\lambda \cdot \tau \cdot \hat{R}_n^\tau(\pi_Q) - \frac{KL[Q||P] + \ln \frac{2\sqrt{n}}{\delta}}{n} \right] \right). \end{aligned}$$

The last step is to exploit the fact that the bias of $\hat{R}_n^\tau(\cdot)$ is negative (because the cost $c \leq 0$), we have for any $\pi$:

$$\begin{aligned} \mathbb{E}_{x \sim \nu, a \sim \pi_0(\cdot|x)} \left[ \hat{R}_n^\tau(\pi) \right] &= \mathbb{E}_{x \sim \nu, a \sim \pi_0(\cdot|x)} \left[ \frac{\pi(a|x)}{\max(\pi_0(a|x), \tau)} c(x, a) \right] \\ &\geq \mathbb{E}_{x \sim \nu, a \sim \pi_0(\cdot|x)} \left[ \frac{\pi(a|x)}{\pi_0(a|x)} c(x, a) \right] = R(\pi) \end{aligned}$$

which gives the result stated in Proposition 4.2 by taking the minimum over $\lambda > 0$:

$$R(\pi_Q) \leq \min_{\lambda > 0} \frac{1}{\tau(e^{\lambda} - 1)} \left( 1 - \exp \left[ -\lambda \cdot \tau \cdot \hat{R}_n^\tau(\pi_Q) - \frac{KL[Q||P] + \ln \frac{2\sqrt{n}}{\delta}}{n} \right] \right).$$

### A.3. Limiting behavior of $\mathcal{C}_n$

To build an intuition of the dependency of $\mathcal{C}_n$ on both $n$ and $\tau$, we can linearize the bound by exploiting the well-known inequality $1 - \exp(-x) \leq x$ for $x \in [-1, 1]$. This gives:

$$\mathcal{C}_n(\pi_Q) \leq \min_{\lambda > 0} \frac{\lambda}{(e^{\lambda} - 1)} \hat{R}_n^\tau(\pi_Q) + \frac{KL[Q||P] + \ln \frac{2\sqrt{n}}{\delta}}{\tau(e^{\lambda} - 1)n}.$$

As the upper bound is a minimum over $\lambda > 0$, fixing $\lambda = \frac{1}{50}$ for example still gives a valid upper bound. This value leads to $\frac{\lambda}{(e^{\lambda} - 1)} \approx 1$ and $e^{\lambda} - 1 \approx \frac{1}{50}$, giving an approximated behaviour of the upper bound on $\mathcal{C}_n$ of:

$$\begin{aligned} \mathcal{C}_n(\pi_Q) &\leq \hat{R}_n^\tau(\pi_Q) + 50 \cdot \frac{KL[Q||P] + \ln \frac{2\sqrt{n}}{\delta}}{\tau n} \\ &\leq \hat{R}_n^\tau(\pi_Q) + \mathcal{O} \left( \frac{KL[Q||P] + \ln \frac{2\sqrt{n}}{\delta}}{\tau n} \right). \end{aligned}$$

This result shows that $\mathcal{C}_n$ improves the dependency on $n$ compared to $\mathcal{LS}_n$.

### A.4. Proof of Theorem 1

We fix $\mathcal{D}_n \sim (\mu, \pi_0)^n, \tau \in (0, 1]$. To prove Theorem 4.3, we use the equality stated in Theorem 3 from (Letarte et al., 2019) applied to the rescaled cIPS $0 \leq \hat{\mathcal{L}}_n(\cdot) = 1 + \tau \cdot \hat{R}_n^\tau(\cdot) \leq 1$. For any distribution $P$, any distribution $Q$ that is $P$-continuous, $\delta \in (0, 1]$, we have:

$$\sup_{0 \leq p \leq 1} \left\{ p : kl(\hat{\mathcal{L}}_n(\pi_Q)||p) \leq \frac{KL[Q||P] + \ln \frac{2\sqrt{n}}{\delta}}{n} \right\} = 1 + \tau \cdot \mathcal{C}_n^{P,\delta,\tau}(\pi_Q)$$

with $\mathcal{C}_n^{P,\delta,\tau}(\pi_Q) := \min_{\lambda > 0} \frac{1 - e^{-\tau\lambda\Gamma_n^\tau(Q,\lambda,\delta)}}{\tau(e^\lambda - 1)}$, $\Gamma_n^\tau(Q, \lambda, \delta) = \hat{R}_n^\tau(\pi_Q) + \frac{KL[Q||P] + \ln \frac{2\sqrt{n}}{\delta}}{\tau\lambda n}$ and $kl(q||p) = q \log(\frac{q}{p}) + (1 - q) \log(\frac{1-q}{1-p})$, the KL divergence between two Bernoulli variables of parameters $p$ and $q$. This means that:

$$kl(\hat{\mathcal{L}}_n(\pi_Q)||1 + \tau \cdot \mathcal{C}_n^{P,\delta,\tau}(\pi_Q)) \leq \frac{KL[Q||P] + \ln \frac{2\sqrt{n}}{\delta}}{n}.$$

By leveraging the following inequality: $p \leq q + \sqrt{2qkl(q||p)} + 2kl(q||p)$ for $p \leq q$, we get:

$$1 + \tau\mathcal{C}_n^{P,\delta,\tau}(\pi_Q) \leq 1 + \tau\hat{R}_n^\tau(\pi_Q) + \sqrt{\frac{2[1 + \tau\hat{R}_n^\tau(\pi_Q)](KL[Q||P] + \ln \frac{2\sqrt{n}}{\delta})}{n}} + \frac{2(KL[Q||P] + \ln \frac{2\sqrt{n}}{\delta})}{n}.$$

Giving the result of Theorem 4.3:

$$\mathcal{C}_n^{P,\delta,\tau}(\pi_Q) \leq \hat{R}_n^\tau(\pi_Q) + \sqrt{\frac{2[\frac{1}{\tau} + \hat{R}_n^\tau(\pi_Q)](KL[Q||P] + \ln \frac{2\sqrt{n}}{\delta})}{\tau n}} + \frac{2(KL[Q||P] + \ln \frac{2\sqrt{n}}{\delta})}{\tau n}.$$

### A.5. Proposition 3 beyond the i.i.d. case

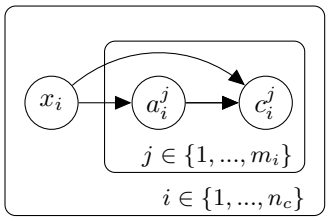


Figure 3: The "Multiple Interactions" Setting.

In a multitude of applications, the i.i.d. assumption made on $\{x_i, a_i, c_i\}_{i \in [n]}$ can be violated. Indeed, a decision system can interact with the same context $x_i$ multiples times, trying different actions and logging the feedbacks as represented in Figure 3. Let $m_i$ be the number of times the system interacted with context $x_i$. The logged dataset in this case can be represented by

$$\mathcal{D}_{n_c}^{\{m_i\}_{i \in [n_c]}} = \left\{ x_i, \{a_i^j, c_i^j\}_{j \in [m_i]} \right\}_{i \in [n_c]}$$

with $n_c$ representing the number of contexts and $n = \sum_i^{n_c} m_i$ the total number of datapoints. As soon as we have an $m_{i_0} > 1$, the i.i.d. assumption does not hold anymore as the samples $\{x_{i_0}, \{a_{i_0}^j, c_{i_0}^j\}_{j=1}^{m_{i_0}}\}$ share the same observation $x_{i_0}$ and thus are dependent. In this case, the cvcIPS estimator will be written as:

$$\hat{R}_n^{\tau,\xi}(\pi_Q) = \xi + \sum_{i=1}^{n_c} \sum_{j=1}^{m_i} \frac{\omega_{\pi_Q}^\tau(a_i^j|x_i)(c_i^j - \xi)}{n_c m_i}$$

We recover the i.i.d. case by taking $m_i = 1 \; \forall i$. Under this weaker assumption, (Catoni, 2007) or any classical PAC-Bayesian bound cannot be applied directly.

### A.5.1. Proof of Proposition 3

In this section, we begin by stating Proposition 4.4 for the more general case where we have a logged dataset $\mathcal{D}_{n_c}^{\{m_i\}_{i\in[n_c]}}$.

**Proposition A.2.** *Given a prior $P$ on $\mathcal{F}_\Theta$, $\xi \in [-1,0], \tau \in (0,1]$, $\delta \in (0,1]$ and a set of strictly positive scalars $\Lambda = \{\lambda_i\}_{i\in[n_\Lambda]}$. We have with probability at least $1-\delta$ over draws of $\mathcal{D}_{n_c}^{\{m_i\}_{i\in[n_c]}} \sim \prod_{i=1}^{n_c}(\nu, \pi_0^{m_i})$: For any $Q$ that is $P$-continuous, any $\lambda \in \Lambda$:*

$$R(\pi_Q) \leq \hat{R}_n^{\tau,\xi}(\pi_Q) - \xi\mathcal{B}_{n_c}^{\tau}(\pi_Q) + \sqrt{\frac{KL[Q||P] + \ln\frac{4\sqrt{n_c}}{\delta}}{2n_c}}$$
$$+ \frac{KL[Q||P] + \ln\frac{2n_\Lambda}{\delta}}{\lambda} + \frac{\lambda l_\xi}{n_c}\sum_{i=1}^{n_c}\frac{1}{m_i n_c} g\left(\frac{\lambda b_\xi}{m_i n_c}\right)\mathcal{V}^{\tau,i}(\pi_Q)$$

*with $g : u \to \frac{\exp(u)-1-u}{u^2}$, $l_\xi = \max\left(\xi^2, (1+\xi)^2\right)$, $b_\xi = \frac{1+\xi}{\tau} - \xi$, $\mathcal{V}^{\tau,i}(\pi) = \mathbb{E}_{\pi(.|x_i)}\left[\frac{\pi_0(a|x_i)}{\max(\tau,\pi_0(a|x_i))^2}\right]$ and $\mathcal{B}_{n_c}^{\tau}(\pi) = \frac{1}{n_c}\sum_{i=1}^{n_c}\mathbb{E}_{\pi(.|x_i)}\left[\mathbb{1}[\pi_0(a|x_i) < \tau]\left(1 - \frac{\pi_0(a|x_i)}{\tau}\right)\right]$.*

We use a decomposition similar to (Kuzborskij et al., 2021) and rewrite the difference $R(\pi_Q) - \hat{R}_n^{\tau,\xi}(\pi_Q) = D_1(\pi_Q) + D_2(\pi_Q) + D_3(\pi_Q)$ with:

$$D_1(\pi_Q) = R(\pi_Q) - \frac{1}{n_c}\sum_{i=1}^{n_c} R(\pi_Q|x_i)$$
$$D_2(\pi_Q) = \frac{1}{n_c}\sum_{i=1}^{n_c} R(\pi_Q|x_i) - \frac{1}{n_c}\sum_{i=1}^{n_c}\xi + \mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\omega_{\pi_Q}^{\tau}(a_i^j|x_i)(c(a,x_i)-\xi)\right]$$
$$D_3(\pi_Q) = \frac{1}{n_c}\sum_{i=1}^{n_c}\xi + \mathbb{E}_{\pi_0(\cdot|x_i)}\left[\omega_{\pi_Q}^{\tau}(a|x_i)(c(a,x_i)-\xi)\right] - \hat{R}_n^{\tau,\xi}(\pi_Q).$$

For the first difference $D_1$, we use (McAllester, 1998) bound for the $[0,1]$-bounded loss using $0 \leq 1 + R(\pi_Q|x_i) \leq 1$. We get with probability at least $1-\delta$, For any $Q$ that is $P$-continuous:

$$D_1(\pi_Q) \leq \sqrt{\frac{KL[Q||P] + \ln\frac{2\sqrt{n_c}}{\delta}}{2n_c}}. \tag{9}$$

The second difference quantifies the bias of our estimator given the contexts $\{x_i, ..., x_{n_c}\}$. Even if we cannot compute it, we can give an upper bound for $D_2$. We have:

$$\begin{aligned}
D_2(\pi_Q) &= \frac{1}{n_c}\sum_{i=1}^{n_c} R(\pi_Q|x_i) - \frac{1}{n_c}\sum_{i=1}^{n_c}\xi + \mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\omega_{\pi_Q}^{\tau}(a|x_i)(c(a,x_i)-\xi)\right] \\
&= \frac{1}{n_c}\sum_{i=1}^{n_c}\mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[(\omega_{\pi_Q}^{0}(a|x_i) - \omega_{\pi_Q}^{\tau}(a|x_i))(c(a,x_i)-\xi)\right] \\
&= \frac{1}{n_c}\sum_{i=1}^{n_c}\mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\mathbb{1}[\pi_0(a|x_i) < \tau](\frac{\pi_Q(a|x_i)}{\pi_0(a|x_i)} - \frac{\pi_Q(a|x_i)}{\tau})(c(a,x_i)-\xi)\right] \\
&= \frac{1}{n_c}\sum_{i=1}^{n_c}\mathbb{E}_{a\sim\pi_Q(\cdot|x_i)}\left[\mathbb{1}[\pi_0(a|x_i) < \tau](1 - \frac{\pi_0(a|x_i)}{\tau})(c(a,x_i)-\xi)\right](c \leq 0) \\
&\leq -\frac{\xi}{n_c}\sum_{i=1}^{n_c}\mathbb{E}_{a\sim\pi_Q(\cdot|x_i)}\left[\mathbb{1}[\pi_0(a|x_i) < \tau](1 - \frac{\pi_0(a|x_i)}{\tau})\right] = -\xi\mathcal{B}_{n_c}^{\tau}(\pi_Q).
\end{aligned}$$

We obtain $-\xi\mathcal{B}_{n_c}^{\tau}(\pi_Q)$, an empirical upper bound to $D_2(\pi_Q)$.

The last step is to control the difference $D_3$. Before doing this, we need to state two lemmas that will help us control the difference $D_3$.

**Lemma A.3.** *Change of measure: Let $f$ be a function of the parameter $\theta$ and data $S$, for any distribution $Q$ that is $P$ continuous, for any $\delta \in (0,1]$, we have with probability $1-\delta$ :*

$$\mathbb{E}_{\theta\sim Q}[f(\theta,S)] \leq KL[Q||P] + \ln \frac{\Psi_f}{\delta} \tag{10}$$

*with $\Psi_f = \mathbb{E}_S \mathbb{E}_{\theta\sim P}[e^{f(\theta,S)}]$.*

Lemma A.3 is the backbone of many PAC Bayes bounds. It is proven in many references, see for example (Alquier, 2021) or Lemma 1.1.3 in (Catoni, 2007). We will combine it with an inequality on the moment generating function to prove a Bernstein-like PAC-Bayes bound (Seldin et al., 2012).

**Lemma A.4.** *Let $W$ be a r.v with $\mathbb{E}[W^2] < \infty$, we suppose that $\mathbb{E}[W] - W \leq B$. Let $g : u \to \frac{\exp(u)-1-u}{u^2}$, we have for all $\eta \geq 0$:*

$$\mathbb{E}[\exp(\eta(\mathbb{E}[W] - W) - \eta^2 g(\eta B)\mathbb{V}[W])] \leq 1. \tag{11}$$

Lemma A.4 is stated and proven in (McDiarmid, 1998).

Combining both lemmas allows us to control the difference $D_3$ with a conditional Bernstein PAC-Bayesian bound:

**Corollary A.5.** *Conditional Bernstein PAC-Bayesian Bound: Let's fix a $\lambda > 0$ and a prior $P$, for any distribution $Q$ that is $P$ continuous, for any $\delta \in (0,1]$, we have with probability at least $1-\delta$:*

$$D_3(\pi_Q) \leq \frac{KL[Q||P] + \ln\frac{1}{\delta}}{\lambda} + \frac{\lambda l_\xi}{n_c}\sum_{i=1}^{n_c}\frac{1}{m_i n_c} g\left(\frac{\lambda b_\xi}{m_i n_c}\right)\mathcal{V}^{\tau,i}(\pi_Q) \tag{12}$$

*with $l_\xi = \max\left(\xi^2, (1+\xi)^2\right)$, $b_\xi = \frac{1+\xi}{\tau} - \xi$, $\mathcal{V}^{\tau,i}(\pi) = \mathbb{E}_{\pi(.|x_i)}\left[\frac{\pi_0(a|x_i)}{\max(\tau,\pi_0(a|x_i))^2}\right]$.*

**Proof:** Let us fix a context $x_i$ and an action $a_i^j$ and let $\theta \sim P$. We have:

$$D_i^j(\theta) = \mathbb{E}_{\pi_0(\cdot|x_i)}\left[\omega_{d_\theta}^\tau(a|x_i)(c(a,x_i)-\xi)\right] - \omega_{d_\theta}^\tau(a_i^j|x_i)(c(a_i^j,x_i)-\xi) \leq b_\xi = \frac{1+\xi}{\tau} - \xi.$$

We fix a $\lambda$ and choose:

$$\begin{aligned} f(\theta,S) &= \sum_{i=1}^{n_c}\sum_{j=1}^{m_i}\left[\frac{\lambda}{m_i n_c}D_i^j(\theta) - (\frac{\lambda}{m_i n_c})^2 g\left(\frac{\lambda b_\xi}{m_i n_c}\right)\mathbb{E}_{\pi_0(\cdot|x_i)}[D_i(\theta)^2]\right] \\ &= \sum_{i=1}^{n_c}\sum_{j=1}^{m_i}\left[\Delta_i^j(\theta)\right]. \end{aligned}$$

From Lemma 2 and because the prior $P$ does not depend on the data, we have:

$$\begin{aligned} \Psi_f = \mathbb{E}_{\prod_i \pi_0(\cdot|x_i)}\mathbb{E}_{\theta\sim P}[e^{f(\theta,S)}] &= \mathbb{E}_{\theta\sim P}\mathbb{E}_{\prod_i \pi_0(\cdot|x_i)}[e^{f(\theta,S)}] \\ &= \mathbb{E}_{\theta\sim P}\prod_i(\mathbb{E}_{\pi_0(\cdot|x_i)}[e^{\Delta_i^0(\theta)}])^{m_i} \leq 1. \end{aligned}$$

It means that $\ln \Psi_f \leq 0$. Using this in Lemma 1, we get:

$$\begin{aligned} D_3(\pi_Q) &\leq \frac{KL[Q||P] + \ln\frac{1}{\delta}}{\lambda} + \sum_{i=1}^{n_c}\sum_{j=1}^{m_i}\frac{\lambda}{(m_i n_c)^2} g\left(\frac{\lambda b_\xi}{m_i n_c}\right)\mathbb{E}_{\pi_0(\cdot|x_i)}\left[\mathbb{E}_{\theta\sim Q}[D_i(\theta)^2]\right] \\ &= \frac{KL[Q||P] + \ln\frac{1}{\delta}}{\lambda} + \frac{\lambda}{n_c}\sum_{i=1}^{n_c}\frac{1}{m_i n_c} g\left(\frac{\lambda b_\xi}{m_i n_c}\right)\mathbb{E}_{\theta\sim Q}\left[\mathbb{E}_{\pi_0(\cdot|x_i)}\left[D_i(\theta)^2\right]\right]. \end{aligned}$$

we also use the following inequality to upper bound $\mathbb{E}_{\pi_0(\cdot|x_i)}[D_i(\theta)^2]$:

$$\begin{aligned}\mathbb{E}_{\pi_0(\cdot|x_i)}[D_i(\theta)^2] &\leq \mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{d_\theta(a|x_i)}{\max(\pi_0(a|x_i),\tau)^2}(c(a,x_i)-\xi)^2\right] \\ &\leq \max(\xi^2,(1+\xi)^2)\mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{d_\theta(a|x_i)}{\max(\pi_0(a|x_i),\tau)^2}\right] \quad \text{because both } c,\xi\in[-1,0] \\ &= l_\xi \mathcal{V}^{\tau,i}(d_\theta).\end{aligned}$$

As the quantity $\mathcal{V}^{\tau,i}$ is linear in $d_\theta$, the result in Corollary 1 follows:

$$D_3(\pi_Q) \leq \frac{KL[Q||P]+\ln\frac{1}{\delta}}{\lambda} + \frac{\lambda l_\xi}{n_c}\sum_{i=1}^{n_c}\frac{1}{m_i n_c} g\left(\frac{\lambda b_\xi}{m_i n_c}\right)\mathcal{V}^{\tau,i}(\pi_Q).$$

Finally, We take a union bound of Corollary 1 over $\Lambda$, a discrete set with cardinal $n_\Lambda$, and combine its result with the bound giving (9) through another union bound to obtain Proposition A.2.

### A.5.2. Choice of $\Lambda$ when $m_i = m$

When the number of interactions $m$ is constant across all contexts, the result in Corollary 1 becomes for a fixed $\lambda$:

$$D_3(\pi_Q) \leq \frac{KL[Q||P]+\ln\frac{1}{\delta}}{\lambda n} + \lambda l_\xi g\left(\lambda b_\xi\right)\mathcal{V}^{\tau}_{n_c}(\pi_Q)$$

where $\mathcal{V}^{\tau}_{n_c}(\pi_Q)$ was defined in Proposition 4.4.

We would like to choose a $\lambda$ that minimizes the bound on $D_3$. Unfortunately, we cannot do it because the minimizer $\lambda^*$ depends on $Q$. Instead, we build an interval in which $\lambda^*$ can be found.

The function $g : u \to \frac{\exp(u)-1-u}{u^2}$ behaves like $\frac{\exp(u)}{u^2}$ when $u$ is big enough, meaning that we should control the values of $g$, and thus $\lambda$ by an upper bound. Choosing $\lambda \leq b = \frac{2n}{b_\xi}$ allows us to control the function $g\left(\frac{\lambda b_\xi}{n}\right) \leq g(2) \leq 1.1$.

Now that an upper bound is found, we still need to find the lowest possible value for $\lambda^*$. Of course, choosing the interval $[0,b]$ can be enough but we want to do more than that. $\lambda^*$ verifies the following equality:

$$\lambda^* = \sqrt{\frac{KL[Q||P]+\ln\frac{1}{\delta}}{\frac{l_\xi}{n}g\left(\frac{\lambda^* b_\xi}{n}\right)\mathcal{V}^{\tau}_{n_c}(\pi_Q) + \frac{\lambda^* l_\xi b_\xi}{n^2}g'\left(\frac{\lambda^* b_\xi}{n}\right)\mathcal{V}^{\tau}_{n_c}(\pi_Q)}}.$$

Let's assume that $\lambda^\star \leq b$. (If not, we can still restrict to $\lambda \in [a,b]$, with the value of $a$ found below.) We have that $KL[Q||P] \geq 0$, and $\mathcal{V}^{\tau}_{n_c} \leq \frac{1}{\tau}$. As the function $g$ is increasing and convex ($g'$ increasing), we get the following inequality:

$$\lambda^* \geq \sqrt{\frac{n\tau\ln\frac{1}{\delta}}{l_\xi g(2) + 2l_\xi g'(2)}}.$$

Using the fact that $g'(2) = 1/2$ and $g(2)+1 \leq 5/2$, we get:

$$\lambda^* \geq \sqrt{\frac{n\tau\ln\frac{1}{\delta}}{l_\xi g(2)+l_\xi}} \geq \sqrt{\frac{2n\tau\ln\frac{1}{\delta}}{5l_\xi}} = a.$$

We now have an interval $\lambda^* \in [a,b]$. One can observe that the optimal $\mathcal{O}(\sqrt{n}) \leq \lambda^* \leq \mathcal{O}(n)$.

We choose the set $\Lambda$ to be a linear discretization of $[a,b]$ giving $\Lambda = \{a+i(b-a)\}_{i\in[n_\Lambda]}$.

### A.5.3. Dependencies of the bound

The bound for the i.i.d. case can be written as:

$$R(\pi_Q) \leq \hat{R}_n^{\tau,\xi}(\pi_Q) - \xi\mathcal{B}_n^{\tau}(\pi_Q) + \sqrt{\frac{\mathcal{KL}(Q||P) + \ln\frac{4\sqrt{n}}{\delta}}{2n}} + \min_{\lambda\in\Lambda}\left\{\frac{\mathcal{KL}(Q||P) + \ln\frac{2n_\Lambda}{\delta}}{\lambda n} + \lambda l_\xi g\left(\lambda b_\xi\right)\mathcal{V}_n^{\tau}(\pi_Q)\right\}$$

with $\Lambda = \{\frac{a}{b} + i\frac{(b-a)}{n}\}_{i\in[n_\Lambda]}$. We know that the biggest value of $\lambda \in \Lambda$ is $b = \frac{2}{b_\xi}$ and that $g(2) \approx 1$. This gives:

$$\begin{aligned}
\min_{\lambda\in\Lambda} A(\lambda) &= \min_{\lambda\in\Lambda} \frac{\mathcal{KL}(Q||P) + \ln\frac{2n_\Lambda}{\delta}}{\lambda n} + \lambda l_\xi g\left(\lambda b_\xi\right)\mathcal{V}_n^{\tau}(\pi_Q) \\
&\leq \min_{\lambda\in\Lambda} \frac{\mathcal{KL}(Q||P) + \ln\frac{2n_\Lambda}{\delta}}{\lambda n} + \lambda l_\xi g\left(2\right)\mathcal{V}_n^{\tau}(\pi_Q) \\
&\lessapprox \min_{\lambda\in\mathbb{R}^+} \frac{\mathcal{KL}(Q||P) + \ln\frac{2n_\Lambda}{\delta}}{\lambda n} + \lambda l_\xi \mathcal{V}_n^{\tau}(\pi_Q) = 2\sqrt{\frac{l_\xi\mathcal{V}_n^{\tau}(\pi_Q)\left(\mathcal{KL}(Q||P) + \ln\frac{2n_\Lambda}{\delta}\right)}{n}}
\end{aligned}$$

We make the hypothesis that our $\Lambda$ is well built to have a value of $\lambda$ close to the true minimizer most of the time. This gives the following limiting behavior:

$$\mathcal{CBB}_n^{\xi}(\pi_Q) = \hat{R}_n^{\tau,\xi}(\pi_Q) - \xi\mathcal{B}_n^{\tau}(\pi_Q) + \mathcal{O}\left(\left(\frac{1}{2\sqrt{2}} + \sqrt{l_\xi\mathcal{V}_n^{\tau}(\pi_Q)}\right)\sqrt{\frac{\mathcal{KL}(Q||P)}{n}}\right).$$

## A.6. Linear Independent Gaussian Policies

To obtain these policies, we restrict $f_\theta$ to:

$$\forall x, a \quad f_\theta(x,a) = \phi(x)^T\theta_a \tag{13}$$

with $\phi$ a fixed transform over the contexts. This results in a parameter $\theta$ 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 $\Phi$ the cumulative distribution function of the standard normal.

**Proof:** We rewrite the definition of $\pi_{\boldsymbol{\mu},\sigma}$ as a probability and exploit the stability of the Gaussian distribution.

$$\begin{aligned}
\pi_{\boldsymbol{\mu},\sigma}(a|x) &= \mathbb{E}_{\theta\sim\mathcal{N}(\boldsymbol{\mu},\sigma^2 I_d)}\left[\mathbb{1}[\text{argmax}_{a'}\,\phi(x)^T\theta_{a'} = a]\right] \\
&= \mathbb{E}_{S\sim\mathcal{N}(\phi(x)^T\boldsymbol{\mu},\sigma^2||\phi(x)||^2 I_K)}\left[\mathbb{1}[\text{argmax}_{a'}\,S_{a'} = a]\right] \\
&= \mathbb{P}_{S\sim\mathcal{N}(\phi(x)^T\boldsymbol{\mu},\sigma^2||\phi(x)||^2 I_K)}\left(\text{argmax}_{a'}\,S_{a'} = a\right) \\
&= \mathbb{P}_{S\sim\mathcal{N}(\phi(x)^T\boldsymbol{\mu},\sigma^2||\phi(x)||^2 I_K)}\left(S_a \geq S_{a'}, \quad \forall a' \neq a\right) \\
&= \mathbb{P}_{Z\sim\mathcal{N}(0_K,I_K)}\left(Z_a + \frac{\phi(x)^T(\boldsymbol{\mu}_a - \boldsymbol{\mu}_{a'})}{\sigma||\phi(x)||} \geq Z_{a'}, \quad \forall a' \neq a\right).
\end{aligned}$$

We condition on $Z_a$ to obtain independent events as for all $a$, the random variables $Z_a$ are independent.

$$\begin{aligned}
\pi_{\boldsymbol{\mu},\sigma}(a|x) &= \mathbb{P}_{Z\sim\mathcal{N}(0_K,I_K)}\left(Z_a + \frac{\phi(x)^T(\boldsymbol{\mu}_a - \boldsymbol{\mu}_{a'})}{\sigma||\phi(x)||} \geq Z_{a'}, \quad \forall a' \neq a\right) \\
&= \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1)}\left[\mathbb{P}_{Z\sim\mathcal{N}(0_K,I_K)}\left(\epsilon + \frac{\phi(x)^T(\boldsymbol{\mu}_a - \boldsymbol{\mu}_{a'})}{\sigma||\phi(x)||} \geq Z_{a'}, \quad \forall a' \neq a | Z_a = \epsilon\right)\right] \\
&= \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1)}\left[\prod_{a'\neq a}\mathbb{P}_{z\sim\mathcal{N}(0,1)}\left(z \leq \epsilon + \frac{\phi(x)^T(\boldsymbol{\mu}_a - \boldsymbol{\mu}_{a'})}{\sigma||\phi(x)||}\right)\right] \\
&= \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1)}\left[\prod_{a'\neq a}\Phi\left(\epsilon + \frac{\phi(x)^T(\boldsymbol{\mu}_a - \boldsymbol{\mu}_{a'})}{\sigma||\phi(x)||}\right)\right] \quad \square.
\end{aligned}$$

### A.7. Why not Mixed Logit Policies?

(London & Sandler, 2019) used in their analysis Mixed Logit Policies to derive a learning principle for softmax policies. Mixed Logit Policies can be written as:

$$\forall (a,x), \quad \pi^{ML}_{\boldsymbol{\mu},\sigma}(a|x) = \mathbb{E}_{\theta\sim\mathcal{N}(\boldsymbol{\mu},\sigma^2 I_d)}[\text{softmax}_K(\phi(x)^T\theta_a)].$$

Even if these policies can behave properly (reparametrization trick gradient for instance), they are not ideal for learning with guarantees in the context of Offline Contextual Bandits. Indeed, we know that the solution of the contextual bandit problem is a deterministic decision function $d^*$, always choosing the action with the minimum cost. Let us suppose that there exists a parameter $\mu^*$ such that:

$$\forall (a,x), \quad d^*(a|x) = d_{\mu^*}(a|x) = \mathbb{1}[\text{argmax}_{a'\in\mathcal{A}}(\phi(x)^T\mu^*_{a'}) = a]$$

We also suppose that we have access to its parameter $\mu^*$. To recover $d_{\mu^*}$ with **LIG** policies, we need to have the scale parameter small enough $\sigma \to 0$ as :

$$\pi_{\boldsymbol{\mu}^*,\sigma}(a|x) \xrightarrow[\sigma\to 0]{} d_{\mu^*}(a|x) \quad \forall x, a.$$

For **Mixed Logit** policies however, having $\sigma \to 0$ is not enough as:

$$\pi^{ML}_{\boldsymbol{\mu}^*,\sigma}(a|x) \xrightarrow[\sigma\to 0]{} \text{softmax}_K(\phi(x)^T\mu^*_a) \quad \forall x, a.$$

One should also increase the norm of $\mu^*$ enough ($||\mu^*|| \to \infty$) to obtain $d_{\mu^*}$.

Let us suppose that we start with the same prior $P = \mathcal{N}(\boldsymbol{\mu}^*, I_d)$ in our bounds. The price to pay in terms of complexity $KL(Q_{\boldsymbol{\mu},\sigma}||P)$ to obtain the solution; a deterministic policy, will be much higher for **Mixed Logit** policies (as we should decrease $\sigma$ and increase the norm of $\boldsymbol{\mu}$) than **LIG** policies (only decrease $\sigma$ and let $\boldsymbol{\mu} = \boldsymbol{\mu}^*$). This means that for a fixed number of samples $n$, we will always get better results with **LIG** policies than **Mixed Logit** policies.

### A.8. The bounds stated for LIG policies

In this section, we want to state the previous Propositions 4.2 and 4.4 (valid for any policy) for the class of **LIG** policies. This class of policies uses Independent Gaussian distributions with shared scale so we will begin by stating the KL divergence between $P = \mathcal{N}(\boldsymbol{\mu}_0, \sigma_0 I_d)$ and $Q = \mathcal{N}(\boldsymbol{\mu}, \sigma I_d)$. We have:

$$KL[Q||P] = D[\boldsymbol{\mu}, \sigma, \boldsymbol{\mu_0}, \sigma_0] = \frac{||\mu - \mu_0||^2}{2\sigma_0^2} + d\left(\frac{\sigma^2}{2\sigma_0^2} + \ln\frac{\sigma_0}{\sigma} - \frac{1}{2}\right).$$

We write the bounds slightly differently by taking the minimimum over the considered $\lambda$ (if the bound is true for any $\lambda$, it is true for the minimum of the bound over $\lambda$).

We state Catoni's bound for **LIG** policies:

**Corollary A.6.** ***LIG*** *policies with Catoni's bound*

*Given a Gaussian prior* $P = \mathcal{N}(\boldsymbol{\mu}_0, \sigma_0 I_d)$, $\tau \in (0,1]$, $\delta \in (0,1]$. *We have with probability* $1-\delta$ *over draws of* $\mathcal{D}_n \sim (\nu, \pi_0)^n$*:*

$\forall \boldsymbol{\mu} \in \mathbb{R}^d, \sigma > 0$*:*

$$R(\pi_{\boldsymbol{\mu},\sigma}) \leq \min_{\lambda>0} \frac{1}{\tau(e^{\lambda}-1)} \left[ 1 - \exp\left( -\tau\lambda \hat{R}_n^{\tau}(\pi_{\boldsymbol{\mu},\sigma}) + \frac{D[\boldsymbol{\mu}, \sigma, \boldsymbol{\mu_0}, \sigma_0] + \ln \frac{2\sqrt{n}}{\delta}}{n} \right) \right]$$

We call $\mathcal{C}_n(\pi_{\boldsymbol{\mu},\sigma})$ the upper bound stated by this corollary. We get:

$$\begin{aligned} \mathcal{GR}_{\mathcal{C}}^* &= \min_{\pi_{\boldsymbol{\mu},\sigma}} \mathcal{C}_n(\pi_{\boldsymbol{\mu},\sigma}) \\ \pi_{\mathcal{C}}^* &= \arg\min_{\pi_{\boldsymbol{\mu},\sigma}} \mathcal{C}_n(\pi_{\boldsymbol{\mu},\sigma}) \\ \mathcal{GI}_{\mathcal{C}}^* &= R(\pi_0) - \mathcal{GR}_{\mathcal{C}}^*. \end{aligned}$$

Similarly, we state our variance sensitive bound for **LIG** policies:

**Corollary A.7.** ***LIG*** *policies variance sensitive bound.*

*Given a Gaussian prior* $P = \mathcal{N}(\boldsymbol{\mu}_0, \sigma_0 I_d)$, $\xi \in [-1,0], \tau \in (0,1]$, $\delta \in (0,1]$ *and a set of strictly positive scalars* $\Lambda = \{\lambda_i\}_{i \in [n_\Lambda]}$. *We have with probability at least* $1-\delta$ *over draws of* $\mathcal{D}_{n_c}^m \sim \prod_{i=1}^{n_c} (\nu, \pi_0^m)$*:*

$\forall \boldsymbol{\mu} \in \mathbb{R}^d, \sigma > 0$*:*

$$\begin{aligned} R(\pi_{\boldsymbol{\mu},\sigma}) \leq \hat{R}_n^{\tau,\xi}(\pi_{\boldsymbol{\mu},\sigma}) - \xi \mathcal{B}_{n_c}^{\tau}(\pi_{\boldsymbol{\mu},\sigma}) + \sqrt{\frac{D[\boldsymbol{\mu}, \sigma, \boldsymbol{\mu_0}, \sigma_0] + \ln \frac{4\sqrt{n_c}}{\delta}}{2n_c}} \\ + \min_{\lambda \in \Lambda} \left\{ \frac{D[\boldsymbol{\mu}, \sigma, \boldsymbol{\mu_0}, \sigma_0] + \ln \frac{2n_\lambda}{\delta}}{\lambda} + \frac{\lambda l_\xi}{n} g\left(\frac{\lambda b_\xi}{n}\right) \mathcal{V}_{n_c}^{\tau}(\pi_{\boldsymbol{\mu},\sigma}) \right\} \end{aligned}$$

We call $\mathcal{CBB}_n(\pi_{\boldsymbol{\mu},\sigma}, \xi, m)$ the upper bound stated by this corollary. Similarly we get:

$$\begin{aligned} \mathcal{GR}_{\mathcal{CBB}(\xi,m)}^* &= \min_{\pi_{\boldsymbol{\mu},\sigma}} \mathcal{CBB}_n(\pi_{\boldsymbol{\mu},\sigma}, \xi, m) \\ \pi_{\mathcal{CBB}(\xi,m)}^* &= \arg\min_{\pi_{\boldsymbol{\mu},\sigma}} \mathcal{CBB}_n(\pi_{\boldsymbol{\mu},\sigma}, \xi, m) \\ \mathcal{GI}_{\mathcal{CBB}(\xi,m)}^* &= R(\pi_0) - \mathcal{GR}_{\mathcal{CBB}(\xi,m)}^*. \end{aligned}$$

# B. EXPERIMENTS

## B.1. Detailed Statistics of the dataset splits used

As described in the experiments section, we use the supervised to bandit conversion to simulate logged data as previously adopted in the majority of the literature (Swaminathan & Joachims, 2015b;a; London & Sandler, 2019; Faury et al., 2020; Sakhi et al., 2020b). In this procedure, you need a split $D_l$ (of size $n_l$) to train the logging policy $\pi_0$, another split $D_c$ (of size $n_c$) to generate the logging feedback with $\pi_0$, and finally a test split $D_{test}$ (of size $n_{test}$) to compute the true risk $R(\pi)$ of any policy $\pi$. 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 1.

| Datasets | $N$ | $n_l$ | $n_c$ | $n_{test}$ | $K$ | $p$ |
|---|---|---|---|---|---|---|
| **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 |
| **Mediamill** | 30 993 | 1549 | 29 444 | 12 914 | 101 | 120 |

Table 1: Detailed statistics of the splits used.

### B.2. 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 $\pi_0$.** $\pi_0$ is trained on $D_l$ (supervised manner) with the following parameters:

- We use $L_2$ regularization of $10^{-6}$. This is used to prevent the logging policy $\pi_0$ from being close to deterministic, allowing efficient learning with importance sampling.
- We use Adam (Kingma & Ba, 2014) with a learning rate of $10^{-1}$ for 10 epochs.

**Optimising the bounds.** All the bounds are optimized with the following parameters:

- The clipping parameter $\tau$ is fixed to $1/K$ with $K$ the action size of the dataset.
- We use Adam (Kingma & Ba, 2014) with a learning rate of $10^{-3}$ for 100 epochs.
- For the bounds optimized over **LIG** policies, the gradient 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 $\mathcal{C}_n$, we treat $\lambda$ as a parameter and we look for the minimum of the bound with respect to $\mu, \sigma$ and $\lambda$.
- For $\mathcal{CBB}_n^{\xi}$, we choose the size of $\Lambda$ to be $n_\Lambda = 100$ 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 $\mu$ and $\sigma$ with $\lambda_j$.

### B.3. Impact of changing the number of interactions $m$

The bound proposed in Proposition 4.4 can work beyond the i.i.d. setting and applies to the "multiple interactions" case. Intuitively, adding more interactions with the contexts $x$ allows us to reduce the uncertainty on the cost and thus learn better policies. We want to explore this in Figure 4. We construct with $\pi_0$ a logged dataset with the number of interactions $m \in \{1, 2, 4, 8\}$ using both **FashionMNIST** and **Mediamill** datasets. Once $m > 1$, we can only use the $\mathcal{CBB}$ bound. We stick to the values of $\xi$ previously used $\xi \in \{0, -1/2\}$.

We can observe that increasing the number of $m$ consistently give better results, in terms of guarantees and also the quality of the policy $\pi^*$ minimizing the bounds. We can also observe that even though $m$ reduces the gap between the two estimators ($\xi = 0$ compared to $\xi = -1/2$), the cvcIPS estimator with $\xi = -1/2$ still gives the best results.

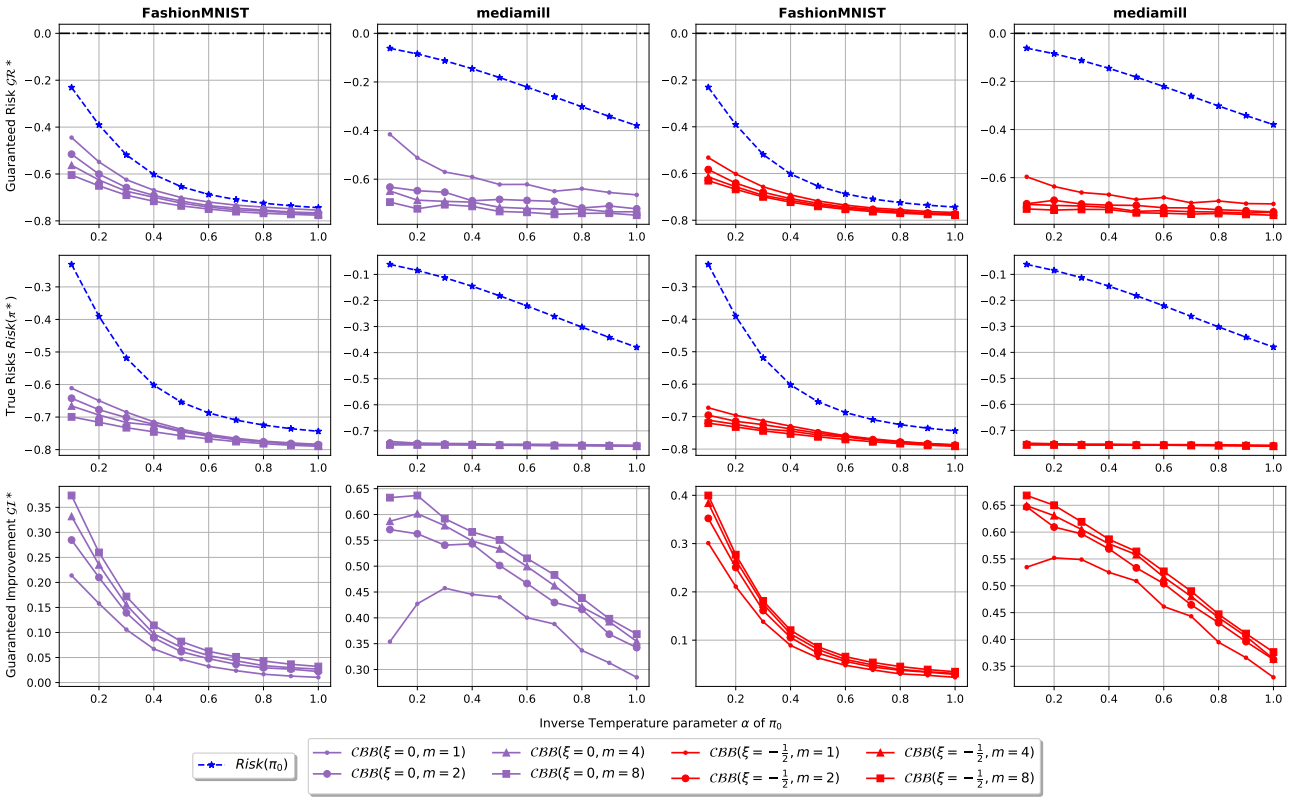


Figure 4: Behavior of the guaranteed risk $\mathcal{GR}^*$ ($\downarrow$ is better), the risk of the minimizer $R(\pi^*)$ ($\downarrow$ is better) and the guaranteed improvement $\mathcal{GI}^*$ ($\uparrow$ is better) given by changing the number of interactions $m$ and $\pi_0$.