Sequential Off-Policy Learning with Logarithmic Smoothing

Maxime Haddouche*
INRIA - CNRS, École Normale Supérieure,
PSL University, France

Otmane Sakhi*
Criteo AI Lab
Paris, France

Abstract

Off-policy learning enables training policies from logged interaction data. Most prior work considers the batch setting, where a policy is learned from data generated by a single behavior policy. In real systems, however, policies are updated and redeployed repeatedly, each time training on all previously collected data while generating new interactions for future updates. This sequential off-policy learning setting is common in practice but remains largely unexplored theoretically. In this work, we present and study a simple algorithm for *sequential off-policy learning*, combining Logarithmic Smoothing (`LS`) estimation with online PAC-Bayesian tools. We further show that a principled adjustment to `LS` improves performance and accelerates convergence under mild conditions. The algorithms introduced generalize previous work: they match state-of-the-art offline approaches in the batch case and substantially outperform them when policies are updated sequentially. Empirical evaluations highlight both the benefits of the sequential framework and the strength of the proposed algorithms.

1 INTRODUCTION

Off-policy learning has become a ubiquitous task for practitioners in a wide range of applications such as recommender systems [Sakhi et al., 2023b], ad placement [Bottou et al., 2013], and personalized medicine [Kallus and Zhou, 2018]. This task involves learning optimal policies from offline data summarizing a *behavior policy*'s past interactions. These interactions, typically captured as tuples of contextual features, action played, and reward received, hold valuable insight into the underlying dynamics of the environment. Each tuple represents a single round of interaction, where the agent observes a context (including relevant features), takes an action according to its current, behavior policy, and receives a reward that depends on both the observed context and the taken action.

The off-policy learning literature focused on developing efficient approaches to learn from these logged interactions [Bottou et al., 2013, Swaminathan and Joachims, 2015, London and Sandler, 2019, Sakhi et al., 2023a]. At the core of these methods lies the development of tailored off-policy estimators with improved concentration properties [Bottou et al., 2013, Aouali et al., 2023, Metelli et al., 2021, Sakhi et al., 2024], combined with tight generalization bounds allowing efficient learning [Swaminathan and Joachims, 2015, Sakhi et al., 2023a, 2024]. The methodologies employed derive from the pessimism principle [Jin et al., 2021], which has been demonstrated to be optimal for learning in these scenarios. Particularly, analyzing this learning problem from the PAC-Bayesian learning theory lens [London and Sandler, 2019] proved effective, improving the seminal work of Swaminathan and Joachims [2015]. This direction was further explored by Sakhi et al. [2023a], Aouali et al. [2023, 2024], Gabbianelli et al. [2024], Sakhi et al. [2024], reducing assumptions and improving the pessimistic principle, *i.e.* controlling the worst-case scenario, with latest developments [Sakhi et al., 2024] resulting in state of the art approaches.

These methods are well understood and have been extensively studied in the one-step, batch setting. In practice, however, practitioners in production environments typically apply them *sequentially* [Zenati et al., 2023], repeatedly updating and redeploying policies by retraining on the full history of collected data while simultaneously generating new interactions for future updates. Intuitively, sequential off-policy learning is preferable, as it can yield improved policies at each de-

*Equal contribution.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

ployment compared to batch approaches. This, in turn, enhances the quality of the logged data and accelerates convergence toward the optimal policy. Although this sequential off-policy learning paradigm is ubiquitous in real-world applications, it remains largely underexplored from a theoretical perspective. Recent work has begun to address this gap [Bibaut et al., 2021, Zenati et al., 2023], but existing algorithms deviate from common practice: they rely on strong uniform coverage assumptions and often require unrealistic data collection strategies [Zenati et al., 2023]. In the conventional off-policy learning setting, some of these assumptions are avoided via the use of PAC-Bayes learning [Alquier, 2024] combined with novel off-policy estimators to reach both theory-driven learning algorithms and convergence guarantees [Sakhi et al., 2024]. However, only batch algorithms have been developed, leaving the existence of sequential off-policy PAC-Bayes algorithms as an open question. In this work, we close this gap and develop the first sequential, off-policy PAC-Bayes algorithms with associated theoretical guarantees.

Contributions. We study the sequential deployment of state-of-the-art off-policy learning methods, a paradigm extensively used in real-world applications [Bibaut et al., 2021, Zenati et al., 2023]. We extend in Sec. 3 the off-policy, logarithmic smoothing PAC-Bayesian strategy of Sakhi et al. [2024] to the sequential scenario, allowing the approach to handle successive interaction data, and progressively update the policy. This sequential strategy, detailed in Algorithm 1, is theoretically analyzed and shown to benefit from improved performance compared to its offline counterpart. Building on this, we identify a core limitation of our previous algorithm, inherited from Sakhi et al. [2024]. In Sec. 4, we address this issue by introducing a second original sequential algorithm (Algorithm 2) enjoying accelerated convergence rates to the optimal policy under reasonable assumptions. Our sequential algorithms encompass the offline scenario and strictly generalize the state-of-the-art offline PAC-Bayesian approach. Finally, we conduct in Sec. 5 numerical experiments to demonstrate the effectiveness of our approach, showing that they outperform both offline strategies and recent sequential strategies.

Outline. Sec. 2 introduces the sequential framework and the necessary background. Sec. 3 provably extends the logarithmic smoothing (LS) PAC-Bayesian off-policy learning approach to the sequential setting and studies its benefits and limitations. Sec. 4 provides a simple adjustment to the LS estimator, achieving an accelerated convergence rate under mild conditions. Empirical validation is conducted in Sec. 5 to highlight the favorable performance of our proposed method, and Sec. 6 provides concluding remarks.

2 FRAMEWORK

Contextual bandits. Let $\mathcal{X} \subset \mathbb{R}^d$ be the context space, which is a compact subset of $\mathbb{R}^d$, and let $\mathcal{A} = [K]$ be a finite action set. An agent's actions are guided by a *stochastic* policy $\pi : \mathcal{X} \to \mathcal{P}(\mathcal{A})$ within a policy space Π. Given a context $x \in \mathcal{X}, \pi(\cdot \mid x)$ is a probability distribution over the action set $\mathcal{A}; \pi(a \mid x)$ is the probability that the agent selects action a in context x.

Sequential framework. We consider that the agent interacts sequentially with a data stream: at time $k \geq 0$, the agent deploys a new behavior policy π_k and a new batch of contexts $\{x_j \mid j \in I_k\}$ observed by the agent is revealed, where $|I_k| = n_k$ and all $(I_k)_{k\geq 0}$ are disjoint. These contexts are drawn *i.i.d.* from the same, unknown distribution ν.

For each $j \in I_k$, the agent selects for context x_j an action $a_j \sim \pi_k(\cdot \mid x_j)$, where π_k is the current *behavior* policy of the agent at time k, and receives a stochastic cost $c_j \in [-1, 0]$ that depends on the observed context x_j and the taken action a_j. This cost c_j is sampled from an unknown cost distribution $p(\cdot|x_j, a_j)$ and we denote by $c(a, x)$ its expectation under $p(\cdot|x_j, a_j)$. After $|I_k| = n_k$ interactions, the agent sequentially updates its behavior to π_{k+1} *w.r.t.* previous behavior policies $(\pi_j)_{j=0\cdots k}$ and is allowed to use all available data $\{x_i, a_i, c_i\}_{i=1\cdots N_k}$, where $N_k := \sum_{i=0}^{k} n_i$.

Formally, we define the full (countable) dataset $S := \cup_{k\in\mathbb{N}} S_k$, where $S_k := \{(x_j, a_j, c_j) \mid j \in I_k\}$ is the set of logged data at time k. We also define $\mathcal{F}_k$ a filtration adapted to $\left(\cup_{j=0}^{k} S_j\right)_{j\in 0\cdots k}$ ($\mathcal{F}_{-1}$ is the trivial σ-algebra). We assume that, conditionally to $\mathcal{F}_{k-1}$, S_k is constituted of mutually independent random variables. This generalizes the pure *static* case, corresponding to $k = 0$. We denote as *sequential algorithm*, any algorithm `Alg` such that its output π_{k+1} at time k is $\mathcal{F}_k$-measurable. In other words, if we denote $S_{0:k} := \cup_{j=0}^{k} S_j$, then we have $\pi_{k+1} = \texttt{Alg}(S_{0:k})$.

To ease notations, we denote by $P(\pi) := \nu \times \pi \times p$ the distribution of (x, a, c) tuples generated by policy π. The performance of any policy π is assessed by its *risk*, which is defined as the averaged incurred cost of π over the draw of new context:

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

Our goal is to come up with sequential learning algorithms that guarantee the convergence to the optimal policy $\pi_\star$, defined as: $\pi_\star = \arg\min_{\pi\in\Pi} R(\pi)$. For any context x, the optimal action is defined as $a_\star(x) = \arg\min_{a'\in\mathcal{A}} c(a', x)$ with ties broken arbitrarily. Hence, the optimal policy $\pi_\star$ chooses deterministically the optimal action $a_\star(x)$ for every context x:

$$\forall (x, a) \in \mathcal{X} \times \mathcal{A}, \quad \pi_\star(a|x) = \mathbb{1}[a = a_\star(x)] .$$

PAC-Bayesian off-policy learning. State of the art off-policy learning algorithms [Sakhi et al., 2023a, Gabbianelli et al., 2024, Sakhi et al., 2024] employ the pessimism principle, shown to be optimal in the batch, off-policy learning setting [Jin et al., 2021]. These methods aim at minimizing an empirical, tight upper bound on the risk $R(\pi)$ over policies $\pi \in \Pi$. These upper bounds leverage PAC-Bayes learning [Alquier, 2024] which controls the generalization guarantee of a *posterior* distribution Q over a predictor space Θ (namely $Q \in \mathcal{P}(\Theta)$). To make this consistent with controlling the risk of a policy π, we exploit a specific parametrization of policies $\pi \in \Pi$. For any $\theta \in \Theta$, we define the deterministic policy: $\forall x, d_\theta(x) := \mathbb{1}\{a = f_\theta(x)\}$, where $f_\theta : \mathcal{X} \to \mathcal{A}$, a parametrized predictor. Then, each distribution $Q \in \mathcal{P}(\Theta)$ defines a policy π_Q by the following mapping:

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

This expression is general as any policy can be represented in this form [Sakhi et al., 2023a, Theorem 3.1]. With this duality, Sakhi et al. [2024] gets with the PAC-Bayesian toolbox, by fixing a reference distribution P, and setting $\delta \in (0,1]$ a learning bound of the following general form, holding with probability $1-\delta$, simultaneously for all Q:

$$\forall Q \in \mathcal{C}(\Theta)\,, R(\pi_Q) \leq \hat{R}(\pi_Q) + \mathcal{O}\left(\mathrm{KL}(Q||P)\right)\,,$$

upper bounding the true risk of π_Q by a risk estimator $\hat{R}(\pi_Q)$ and $\mathrm{KL}(Q||P) := \mathbb{E}_Q\left[\log\left(\frac{dQ}{dP}\right)\right]$, the Kullback-Leibler divergence between P and Q. In the batch setting, these upper bounds are used as learning objectives and the convergence guarantees of their learned policies are studied [Gabbianelli et al., 2024, Sakhi et al., 2023a]. We extend this paradigm to the sequential framework, and begin by defining off-policy estimators using the interaction data collected.

Regularized IPS. Given the successive arrival of data batches $(S_k)_{k\geq 0}$ and output of sequential algorithm $(\pi_k)_{k\geq 0}$, we define for any $\pi \in \Pi$ the *Inverse Propensity Scoring* (IPS) estimator [Horvitz and Thompson, 1952] as: $\hat{R}^{\mathrm{IPS}}_{0:k}(\pi) = \frac{1}{N_k}\sum_{j=0}^{k}\sum_{i\in I_j}\frac{\pi(a_i|x_i)}{\pi_j(a_i|x_i)}c_i$, encompassing the static case ($k=0$). This estimator is unbiased under the common support assumption [Owen, 2013]. However, as in the static case, this estimator has poor concentration properties [Metelli et al., 2021] and regularized variants, trading some bias for lower variance are preferred [Bottou et al., 2013, Su et al., 2020]. The majority of these regularized estimators stems from specific choices of a regularizer $h : [0,1]^2 \times [-1,0] \to \mathbb{R}^-$ in the following definition:

$$\hat{R}^h_{0:k}(\pi) := \frac{1}{N_k}\sum_{j=0}^{k}\sum_{i\in I_j} h(\pi(a_i|x_i), \pi_j(a_i|x_i), c_i)\,,$$

with h respecting $\forall (p,q,c) \in [0,1]^2 \times [-1,0]\,, \quad pc/q \leq h(p,q,c) \leq 0$. $h(p,q,c) = pc/q$ recovers the IPS estimator while other choices recover known, well behaved estimators such as clipped IPS [Bottou et al., 2013], power-mean estimators [Metelli et al., 2021] and Implicit-Exploration [Gabbianelli et al., 2024] to name a few. When $k=0$, we denote $\hat{R}^h_{0:0}$ as $\hat{R}^h_0$.

3 WARM-UP ON PESSIMISTIC LOGARITHMIC SMOOTHING

3.1 Background: Offline algorithms with theoretical guarantees

PAC-Bayesian Logarithmic Smoothing recently appeared as a fruitful theoretical approach to obtain both concrete learning algorithms and generalization guarantees for contextual bandits. In Sakhi et al. [2024], PAC-Bayesian bounds have been developed for off-policy contextual bandits, and optimizing them provide novel learning algorithms with optimality guarantees. We first recall the definition of the logarithmic smoothing estimator (`LS`) as well as its associated regularizer. For any $\lambda > 0$ we define `LS` for $k \geq 0$,

$$h^{\lambda-\mathrm{Lin}}_{\mathrm{LS}}(p,q,c) := -\frac{p}{\lambda}\log\left(1 - \frac{\lambda c}{q}\right)\,,$$
$$\hat{R}^{\lambda-\mathrm{LS}}_{0:k} := \hat{R}^{h^{\lambda-\mathrm{Lin}}_{\mathrm{LS}}}_{0:k}\,.$$

Using these notions, Sakhi et al. [2024] obtained results for static off-policy learning *i.e.* data are gathered in a single batch S_0 of size n_0. We recall them below.

Proposition 3.1 (Propositions 10 and 11 of Sakhi et al., 2024)**.** *Given a prior $P \in \mathcal{P}(\Theta), \delta \in (0,1]$ and $\lambda > 0$, the following holds with probability at least $1-\delta$ over $S_0, \forall Q \in \mathcal{P}(\Theta)$, :*

$$R\left(\pi_Q\right) \leq \hat{R}^{\lambda-\mathrm{LS}}_0\left(\pi_Q\right) + \frac{\mathrm{KL}(Q\|P) + \log\frac{1}{\delta}}{\lambda n_0}.$$

Furthermore, for any $\mathcal{C}(\Theta) \subset \mathcal{P}(\Theta)$, denoting by $\hat{Q} \in \mathrm{argmin}_{Q\in\mathcal{C}(\Theta)}\hat{R}^{\lambda-\mathrm{LS}}_0\left(\pi_Q\right) + \frac{\mathrm{KL}(Q\|P)}{\lambda n_0}$ and $Q^\star \in \mathrm{argmin}_{Q\in\mathcal{C}(\Theta)} R(\pi_Q)$, we have:

$$R(\pi_{\hat{Q}}) - R(\pi_{Q^\star}) \leq \lambda\mathcal{S}(\pi_{Q^\star}) + \frac{2\,\mathrm{KL}(Q^\star\|P) + \log\frac{2}{\delta}}{\lambda n_0},$$

where $\mathcal{S}(\pi) := \mathbb{E}_{(x,a,c)\sim P(\pi_0)}\left[\frac{\pi(a|x)c^2}{\pi_0^2(a|x)}\right]$ is a pseudo-variance term.

Those results provide two distinct insights: *(i)*, the learning procedure yielding $\hat{Q}$ (and exploited in practice in Sakhi et al. [2024]) minimizes a generalization bound and *(ii)*, once $\mathcal{S}(\pi_{Q^\star})$ finite, taking $\lambda = 1/\sqrt{n_0}$ gives a convergence rate of $1/\sqrt{n_0}$ to the optimal value of

the risk. However, Proposition 3.1 and its associated learning algorithm suffer from theoretical and practical limitations.

1. A batch procedure cannot handle a progressive arrival of data $(S_k)_{k\geq 0}$. In particular, Proposition 3.1 does not allow to retrain our policy *w.r.t* all past data when a new data batch arrives.
2. The pseudo-variance term of Proposition 3.1 does not go to zero, even in the oracle case where $\pi_0 = \pi_{Q^\star}$, whatever the optimal policy $\pi_{Q^\star}$ is. This issue is not an artifact of the proof technique, but from the performance objective $R(\pi)$, and consequently its estimator. This exhibits a fundamental limitation of this procedure as the slow convergence rate of $1/\sqrt{n_0}$ is inevitable.

We answer those two fundamental issues below. To answer 1, we show in Sec. 3.2 that leveraging online PAC-Bayes techniques allows obtaining a sequential version of Proposition 3.1, permitting both the arrival of new data on-the-fly and naturally re-using former ones. Then, we answer 2, in Sec. 4 by adjusting the performance objective and its respective estimator, providing a novel sequential learning algorithm with accelerated convergence.

3.2 Towards a first sequential algorithm

In this part, we assume, for pedagogical concerns, that all datasets S_k have the same size $n_k = m$. Still exploiting Logarithmic Smoothing, we provide a sequential counterpart of Proposition 3.1 by involving online PAC-Bayes learning techniques [Haddouche and Guedj, 2022, 2023]. Our result is stated as follows:

Theorem 3.2. *For any sequential algorithm* `Alg`*, any* $m > 0$ *and data independent prior* P*. For any* $\lambda > 0, \delta \in (0,1]$*, with probability at least* $1-\delta$ *over* S*, we have, for all* $k \geq 0$*, denoting by* π_{k+1} *the output of* `Alg` *at time* k*, for all* $Q \in \mathcal{P}(\Theta)$*:*

$$R(\pi_Q) \leq \hat{R}_{0:k}^{\lambda-\mathrm{LS}}(\pi_Q) + \frac{\mathrm{KL}(Q,P) + \log\left(\frac{1}{\delta}\right)}{\lambda(k+1)m}.$$

The proof lies in Appendix B. Note that taking $k = 0$ and $m = n_0$ retrieves exactly the batch case of Proposition 3.1. On the contrary taking $k = n$ and $m = 1$ provides a purely online procedure. Our theorem provides an interpolation between batch and online scenarios, allowing sub-batches of data to arrive at each time step and optimize on-the-fly.

While Thm. 3.2 has a similar shape to Proposition 3.1, the involved toolbox differs fundamentally. Indeed, while batch PAC-Bayes only requires controlling an exponential moment via Markov's inequality, Online PAC-Bayes requires a careful design and control of a supermartingale through Ville's inequality [Doob, 1939]. This core difference allows, beyond Thm. 3.2, to reach a more general result, stated in Appendix C, and allowing *any arbitrary evolution of the context distribution*. In particular, we can avoid the *i.i.d.* assumption on contexts made in Sec. 2. This better reflects real-world applications, where user behavior and responses naturally evolve over time. Although analyzing algorithm convergence in this non-stationary setting is of interest, it falls outside the scope of the present work.

Algorithm 1: Sequential Policy Learning via Logarithmic Smoothing

Input: Policy π_0, policy class $\mathcal{C}(\Theta)$, $\lambda > 0$.
Initialize P such that $\pi_P = \pi_0$ and $\mathcal{S}_{-1} = \{\}$.
for $k \geq 0$ **do**
  Collect data S_k of size $|I_k| = m$ with π_k and define global collection $S_{0:k} = S_{0:k-1} \cup S_k$.
  Solve the optimization problem:
  $$\hat{Q}_{k+1} = \mathrm{argmin}_{Q\in\mathcal{C}(\Theta)} \left\{ \hat{R}_{0:k}^{\lambda-\mathrm{LS}}(\pi_Q) + \frac{\mathrm{KL}(Q,P)}{\lambda(k+1)m} \right\}$$
  Set $\pi_{k+1} = \pi_{\hat{Q}_{k+1}}$.

Notice that the optimization problem $\mathrm{argmin}_{Q\in\mathcal{P}(\Theta)} \hat{R}_{0:k}^{\lambda-\mathrm{LS}}(\pi_Q) + \frac{\mathrm{KL}(Q,P)}{\lambda(k+1)m}$ admits a closed form solution: the *Gibbs posterior* Q_λ such that $dQ_\lambda(\theta) \propto \exp\left(-\lambda(k+1)m\hat{R}_{0:k}^{\lambda-\mathrm{LS}}(d_\theta)\right)$ [Catoni, 2007, Theorem 1.2.6]. However, Gibbs posteriors often require costly, advanced Monte Carlo methods [Chopin and Papaspiliopoulos, 2020] to be computed. A common practice is to restrict the optimization problem on a subclass $\mathcal{C}(\Theta) \subset \mathcal{P}(\Theta)$ on which the optimization problem is tractable, *e.g.* the class of Gaussian distributions [Dziugaite and Roy, 2017, Perez-Ortiz et al., 2021]. Restricting the family of distributions is not limiting: as long as the family of parametrized predictors $\{f_\theta \mid \theta \in \Theta\}$ is expressive enough, the induced policy family still contains the optimal policy [Sakhi et al., 2023a].

Finally, a key remark is that, given a fixed batch size $m > 0$ and parameter $\lambda > 0$, Thm. 3.2 holds for a countable number k of sub-batches simultaneously. A fundamental difference with Proposition 3.1 is that we allow the empirical risk $\hat{R}_{0:k}^{\lambda-\mathrm{LS}}$ to evolve with k, while the batch approach only tackles a single S_0. This allows the optimization of the right-hand-side of Thm. 3.2 at each time step and yields a sequential learning algorithm given in Algorithm 1.

In Algorithm 1, at time k, we instantiate $\hat{R}_{0:k}^{\lambda-\mathrm{LS}}$ with

$(\pi_j)_{j=0\cdots k}$ the outputs of Algorithm 1 up to time $k-1$. Similarly to the batch case, Algorithm 1 also enjoys a convergence guarantee to the optimal policy.

Proposition 3.3. *For any fixed subbatch size $m > 0$, any data independent prior P, any $\lambda > 0, \delta \in (0,1]$, with probability at least $1-\delta$, for any $k \geq 0$, denoting by π_{k+1} the output of Algorithm 1 at time k and $\pi_\star = \pi_{Q_\star}$ with $Q_\star \in \operatorname{argmin}_{Q\in\mathcal{C}(\Theta)} R(\pi_Q)$:*

$$R(\pi_{k+1}) - R(\pi_\star) \leq \frac{\lambda}{k+1}\sum_{j=0}^{k}\mathcal{S}_j(\pi_\star) + 2\frac{\mathrm{KL}(Q^\star, P) + \log\left(\frac{2}{\delta}\right)}{\lambda(k+1)m}$$

where $\mathcal{S}_j(\pi) := \mathbb{E}_{(x,a,c)\sim P(\pi_j)}\left[\frac{\pi(a|x)c^2}{\pi_j^2(a|x)}\right]$.

Proposition 3.3 is proven in Appendix B. To obtain explicit convergence rates, we make the assumption that our policy class covers the optimal actions, i.e.:

Assumption 3.4 (Coverage of the optimal actions)**.** *For any policy π in our policy class $\mathcal{C}(\Theta)$, we assume that: $\inf_x \pi(a^\star(x)|x) \geq C^\star$ for some constant $C^\star > 0$.*

Assumption 3.4 only requires coverage of the optimal actions, which is strictly weaker than the uniform coverage condition assumed in prior work [Zenati et al., 2023]. With this assumption, we ensure that $S_j(\pi_\star)$ is finite and uniformly upper bounded by $1/C^\star$ for all j. Taking $\lambda = \frac{1}{\sqrt{(k+1)m}}$ yields an overall convergence rate of $\mathcal{O}\left(\frac{1}{\sqrt{(k+1)m}}\right)$, matching the rate of Proposition 3.1 in the batch case (i.e $k = 0, m = n_0$) and yielding an original rate of $\mathcal{O}(1/\sqrt{n})$ for an online setting ($k = n, m = 1$). Proposition 3.3 also works for any intermediate regime between batch and online settings, showing the strength of the sequential approach. Looking at Proposition 3.3, we observe that the upper bound depends on the average $\frac{1}{k+1}\sum_{j=0}^{k}\mathcal{S}_j(\pi_\star)$. Intuitively, this average is expected to be smaller than $\mathcal{S}_0(\pi_\star)$, as intermediate policies π_j should be progressively closer to $\pi_\star$. However, this algorithm still suffers from the same limitations to the static case here as the pseudo-variance terms $\mathcal{S}_j$ do not decrease to zero, even in the perfect situation where for all j, $\pi_j = \pi_\star$.

In this part, we showed that online PAC-Bayes techniques successfully extend offline algorithms to the sequential setting, revealing the dependence on the data-collection process while enjoying the same convergence rates. A key question is then: is it possible to find a sequential algorithm with a faster convergence rate?

4 SEQUENTIAL LOGARITHMIC SMOOTHING WITH ACCELERATED CONVERGENCE

Algorithm 2: Sequential Policy Learning via Adjusted Logarithmic Smoothing

Input: Policy π_0, a policy class $\mathcal{C}(\Theta)$, $\lambda \in (0,1)$.
Initialize P such that $\pi_P = \pi_0$ and $\mathcal{S}_{-1} = \{\}$.
for $k \geq 0$: **do**
  Collect data S_k of size $|I_k| = n_k$ with π_k.
  Define global collection $S_{0:k} = S_{0:k-1} \cup S_k$, and $N_k := |S_{0:k}|$.
  Solve the following optimization problem:
  $$\hat{Q}_{k+1} = \operatorname{argmin}_{Q\in\mathcal{C}}\left\{\hat{R}_{0:k}^{\lambda-\mathrm{adj}}(\pi_Q) + \frac{\mathrm{KL}(Q,P)}{\lambda N_k}\right\}$$
  Set $\pi_{k+1} = \pi_{\hat{Q}_{k+1}}$.

While Thm. 3.2 provides a natural sequential extension of Sakhi et al. [2024], Algorithm 1 still suffers (Proposition 3.3) from a slow convergence rate of $\mathcal{O}(1/\sqrt{km})$ due to the nature of the performance objective, and consequently the pseudo-variance term. Is it then possible to avoid such a term? The answer is yes, by adjusting $h^{\lambda-\mathrm{Lin}}$ introduced in Sec. 3: we define $h_\lambda^{\mathrm{adj}}(p,q,c) := -\frac{p}{\lambda}\log\left(1 - \frac{\lambda c}{q(1+\lambda c)}\right)$ and denote it as the *adjusted linear LS regularizer*. We then define the adjusted risk for any sequential algorithm `Alg` and $\lambda \in (0,1), k \geq 0$:

$$\hat{R}_{0:k}^{\lambda-\mathrm{adj}}(\pi) := R_{0,k}^{h_\lambda^{\mathrm{adj}}}(\pi) = -\frac{1}{N_k}\sum_{j=0}^{k}\sum_{i\in I_j}\frac{\pi(a_i|x_i)}{\lambda} \cdot \log\left(1 - \frac{\lambda c_i}{\pi_j(a_i|x_i)(1+\lambda c_i)}\right),$$

where again, π_{k+1} is the output of `Alg` at time k. Using similar tools to those in Sec. 3.2, we are able to derive the following theoretical guarantee for our novel adjusted regularizer h_λ^{adj}.

Theorem 4.1. *For any sequential algorithm `Alg`, any sequence $(n_k)_{k\geq 0}$ and data independent prior P. For any $\boldsymbol{\lambda} \in (0,1), \delta \in (0,1]$, with probability at least $1-\delta$ over S, we have, for all $k \geq 0$, denoting by π_{k+1} the output of `Alg` at time k, for all $Q \in \mathcal{P}(\Theta)$:*

$$R(\pi_Q) - \sum_{j=0}^{k}\frac{n_j}{N_k}R(\pi_j) \leq \hat{R}_{0:k}^{\lambda-\mathrm{adj}}(\pi_Q) + \hat{C}_{0,k}(\lambda) + \frac{\mathrm{KL}(Q,P) + \log\left(\frac{1}{\delta}\right)}{\lambda N_k},$$

with $\hat{C}_{0,k}(\lambda) = \frac{1}{N_k}\sum_{j=0}^{k}\sum_{i\in I_j}\frac{1}{\lambda}\log(\frac{1}{1+\lambda c_i})$.

The proof of Thm. 4.1 is established in Appendix D. Thm. 4.1 relies on the same analytical tools as Thm. 3.2, and therefore inherits the same general insights. Notably, the result remains valid for varying batch sizes $(n_k)_{k\geq 0}$, which makes it applicable to real-world scenarios where data may arrive in irregular quantities. This theorem serves as the theoretical foundation of the main contribution of this work: a sequential learning algorithm for off-policy contextual bandits, presented in Algorithm 2.

Note that at time k, we instantiate $\hat{R}_{0:k}^{\lambda-\mathrm{adj}}(\pi_Q)$ with $(\pi_j)_{j=0\cdots k}$ the previous outputs of Algorithm 2.
Note that in Thm. 4.1, contrary to Thm. 3.2, we do not control directly $R(\pi_Q)$ but the difference $R(\pi_Q) - \sum_{j=0}^{k}\frac{n_j}{N_k}R(\pi_j)$. This means that running Algorithm 2 at time k using $\cup_{j=0}^{k}S_j$ tends to ensure a better generalisation ability than what we obtained (on a weighted average) by running Algorithm 2 on $\cup_{j=0}^{k-1}S_j$. This phenomenon gives an intuition about why Algorithm 2 converges faster to the optimum. We prove this intuition in what follows.

Accelerated convergence rate for Algorithm 2. To prove the fast convergence of Algorithm 2 to $\pi_\star$, we state a first proposition.

Proposition 4.2. *For any data independent prior P, any $\boldsymbol{\lambda} \in (\mathbf{0},\mathbf{1})$, $\delta \in (0,1]$, with probability at least $1-\delta$, we have for all $k \geq 0$, denoting by π_{k+1} the output of Algorithm 2 at time k:*

$$0 \leq R(\pi_{k+1}) - R(\pi_\star)$$
$$\leq \frac{\lambda}{1-\lambda}\sum_{j=0}^{k}\frac{n_j}{N_k}L(\pi_\star,\pi_j) + 2\frac{\mathrm{KL}(Q^\star,P)+\log\left(\frac{2}{\delta}\right)}{\lambda N_k},$$

where for all π,
$L(\pi,\pi_j) = \mathbb{E}_{x\sim\nu}\left[\mathbb{E}_{\pi_j}\left[c^2\right] + \mathbb{E}_{\pi}\left[c^2\left(\frac{1}{\pi_j(a|x)}-2\right)\right]\right]$.

Proposition 4.2 is proven in Appendix D and highlights the interest of h_λ^{adj}: we avoid the pseudo variance terms $\mathcal{S}_j(\pi_\star)$ of Proposition 3.3 to get $L(\pi_\star,\pi_j)$. Contrary to $S_j(\pi_\star)$, we have that, for any deterministic policy π_d, $L(\pi_d,\pi_j)\to 0$ if $\pi_j \to \pi_d$. Such desirable behavior leaves room for accelerated convergence rates. To investigate this route, we target the convergence towards the optimal policy $\pi_\star$ and study the behavior of L.

To prove the acceleration of our method, we need to define additional quantities. We define the suboptimality of action a, given a context x by: $\Delta_{a,x} := c(a,x) - c(a_\star(x),x) \geq 0$. Then, for any $u \in [0,1)$, we define Δ_u such that:

$$P_x\left(\min_{a\neq a^\star(x)}\Delta_{a,x} \geq \Delta_u\right) = 1-u.$$

For a fixed u, Δ_u quantifies how separable $a^\star(x)$ is from the rest of actions. In other words, the magnitude of Δ_u quantifies the difficulty of the problem. $\Delta_u = 0$ for all $u \in [0,1)$ means that the optimal action $a^\star(x)$ is not distinguishable from the second best action, almost surely. Oppositely, $\Delta_u = 1$ for a specific u means that with probability $1-u$, the costs are deterministic and the optimal action $a^\star(x)$ is the only one to have nonzero cost (of -1). We want to avoid the first case, as the existence of a strictly positive $\Delta_u > 0$ will be important next. Finally, we introduce the minimum coverage of the optimal actions under π_k:

$$C_k^\star = \inf_x \pi_k(a^\star(x)|x)\,.$$

Lemma 4.3. *Let $\pi_k \in \Pi$, assume there exists $u \in [0,1)$ such that $\Delta_u > 0$ and $C_k^\star > 0$, then:*

$$L(\pi_\star,\pi_k) \leq \gamma_k(R(\pi_k) - R(\pi_\star)),$$

with $\gamma_k = 1 + \left(\frac{1}{4}+\frac{1}{C_k^\star}\right)\frac{1}{\Delta_u(1-u)} = \mathcal{O}\left(\frac{1}{(1-u)\Delta_u C_k^\star}\right)$.

The proof is given in Appendix D. The condition $\Delta_u > 0$ is analogous to margin assumptions commonly used in multi-arm contextual bandits [Hu et al., 2022, Bibaut et al., 2021], while $C_k^\star > 0$ imposes a coverage requirement ensuring that π_k sufficiently explores optimal actions. Importantly, this condition is strictly weaker than the uniform coverage assumption imposed in Bibaut et al. [2021], Zenati et al. [2023]. Instead of controlling L, a similar inequality on the variance of importance weights was used as an assumption to obtain fast rates in Zenati et al. [2023]. Here, we prove such result and characterize when it is possible to obtain. For instance, this inequality cannot hold when the costs are constant. Finally, while we formulate the conditions in terms of Δ_u and $C_k^\star$ for interpretability, a sharper, but less transparent, formulation is given in Appendix D.3.

Lemma 4.3 shows that the expansion of L is controlled by a difference of risks. Smaller γ_k sharpens this inequality, which is the case for problems with distinguishable optimal actions and near optimal logging policies π_k. Plugging this inequality in Proposition 4.2 then yields a recursive control on the sequence $(R(\pi_k))_{k\geq 0}$ (see Cor. D.5). Such a recursive relation can then be transformed into an explicit accelerated convergence rate for the output of Algorithm 2, provided the following assumption holds: for all $k \geq 0$, the algorithm outputs policies π_k such that

$$\forall k \geq 0\,, \quad C_k^\star \geq C^\star > 0.$$

Intuitively, this is plausible since the initial policy π_0 is expected to cover optimal actions, and subsequent policies π_k for $k \geq 1$ should become progressively closer to

$\pi_\star$. Formally, this condition is guaranteed by Assumption 3.4, which requires the policy class to sufficiently cover the optimal actions. Combined with the separability condition ($\Delta_u > 0$ for some $u \in [0,1)$), we obtain the uniform upper bound on γ_k:

$$\forall k, \quad \gamma_k \leq \gamma = \frac{3}{\Delta_u(1-u)C^\star} < +\infty\,.$$

With this condition, we state our principal result below.

Theorem 4.4. *Let $n_j = m > 0$ for all j. Then, for any $\alpha \in [0,1)$, denote by π_{k+1} the output of Algorithm 2 at time k with parameter $\lambda_m := \frac{1-\alpha}{8\gamma\sqrt{m}}$. Then with probability at least $1-\delta$ over S, for all $k \geq 0$:*

$$R(\pi_{k+1}) - R(\pi^\star) \leq \frac{C_\alpha}{\sqrt{m}(k+1)^\alpha},$$

where $C_\alpha := 64\gamma\frac{\mathrm{KL}(\mathrm{Q}^\star,P)+\log(2/\delta)}{1-\alpha}$.

Thm. 4.4 is proven in Appendix D and provides an accelerated rate of $\mathcal{O}\left(\frac{1}{\sqrt{m}k^\alpha}\right)$ compared to the $\mathcal{O}\left(\frac{1}{\sqrt{mk}}\right)$ of Proposition 3.3 as long as $\alpha > 1/2$. Again, we recover the same rate as Sakhi et al. [2024] in the batch setting (i.e. policy after one batch update: $k = 1, m = n$) and improve the rate in the online setting (i.e. updating the behavior policy after each interaction: $k = n, m = 1$) to $\mathcal{O}(1/n^\alpha)$. Thm. 4.4 also works for any intermediate regime between batch and online settings, and suggests that considering several updates is beneficial to learn efficiently. Indeed, as long as we choose a λ that ensures $\alpha > 1/2$, the number of updates k will have a stronger influence than the size of the batch m. It should be noted, however, that the constant C_α deteriorates as α approaches 1. Larger values of λ_m yield smaller constants but slower convergence rates, whereas smaller λ_m yield faster rates at the cost of heavier constants. In practice, this creates a trade-off between accelerating convergence and keeping C_α at reasonable levels.

Finally, to capture a broad range of real-world scenarios where the size of the batches received n_k may vary over time, we extend these results in Cor. 4.5, which provides convergence guarantees for Algorithm 2 in this more general setting.

Corollary 4.5. *Let $\beta_1 := \sup_{k\geq 0}\frac{k+1}{N_k} \leq 1$. Assume there exists $\beta_2 \in \mathbb{R}^+$ such that for all k, $n_k \leq \beta_2$. For any $\alpha \in [0,1)$, denote by π_{k+1} the output of Algorithm 2 at time k with parameter $\lambda = \frac{1}{1+2^{2+\alpha}\gamma\beta_1\beta_2 B(\alpha)}$. Then with probability at least $1-\delta$ over S, for all $k \geq 0$:*

$$R(\pi_{k+1}) - R(\pi^\star) \leq \frac{C_\alpha}{(k+1)^\alpha},$$

where $C_\alpha \leq \max\left(1, \frac{\mathrm{KL}(\mathrm{Q}^\star,P)+\log(2/\delta)}{\lambda^2\gamma\beta_2 B(\alpha)}\right)$, $B(\alpha) = \frac{1}{1-\alpha}$.

Observe that the convergence rate in Cor. 4.5 is weaker than that of Thm. 4.4, since it no longer depends explicitly on the batch sizes n_j. This relaxation is necessary to accommodate adversarial scenarios where some n_j may be as small as 1. For example, let $n \geq 2$ and consider a sequence where $n_{2j-1} = m$ and $n_{2j} = 1$ for all $j \geq 1$. Then, for all $k \geq 0$, we have $N_{2k-1} = k(m+1) - 1$ and $N_{2k} = k(m+1)$. Thus, for all k, it holds that $\frac{k+1}{N_k} \leq \frac{2}{m}$, and the assumptions of Cor. 4.5 are satisfied. Importantly, Cor. 4.5 matches the optimal convergence rate of Thm. 4.4 in the fully online regime (i.e., $k = n$, $n_j = 1$). Having established that Algorithm 2 enjoys provably accelerated convergence rates, we turn to empirical validation to assess whether this theoretical advantage is reflected in practice.

5 EXPERIMENTS

General Setup. Our experimental framework follows the standard multiclass-to-bandit conversion widely used in prior work [Dudik et al., 2014, Swaminathan and Joachims, 2015]. Each multiclass dataset consists of feature-label pairs, which we transform into contextual bandit problems by treating features as contexts and labels as actions. The reward r for taking action a for context x is modeled as Bernoulli with probability $p_x = \epsilon + \mathbb{1}\left[a = a^\star(x)\right](1-2\epsilon)$, where $a^\star(x)$ is the true label of features x, and ϵ is a noise parameter, that we set to $\epsilon = 0.2$ in all our experiments. This setup ensures an average reward of $1-\epsilon$ for the optimal action $a^\star(x)$ and ϵ for all others. In our experiments, we use parametrized Linear Gaussian Policies [Sakhi et al., 2023a, Aouali et al., 2023] as our class of policies. These policies interact with the contextual bandit environment to construct a logged bandit feedback dataset in the form $\{x_i, a_i, c_i\}_{i\in[n]}$, where $c_i = -r_i$ is the associated cost. Appendix E details the experimental design and obtained results.

Effect of Multiple Deployments. Unlike prior batch approaches, our sequential algorithms support multiple updates of the behavior policy π_0, while seamlessly reducing to the batch case when $k = 1$. Using the `MNIST` [LeCun et al., 2010] and `CIFAR100` [Krizhevsky, 2009] datasets, we evaluate how the number of intermediate policy updates k affects the performance of the final learned policy π. Starting from a uniform policy π_0, we run both `SeqLS` (`Alg` 1) and the improved `SeqAdjLS` (`Alg` 2). For each $k \in \{1, 5, 10, 100\}$, we fix the total number of interactions to $N_k = N$, resulting in $m = N/k$ interactions per deployment. The case $k = 1$ corresponds to the static setting studied in Sakhi et al. [2024], where all data is collected using π_0. For $k > 1$, each intermediate policy π_j interacts with m examples, collectively producing N sequentially collected samples. We set $\lambda = 1/\sqrt{m}$, following our theoretical guidance.

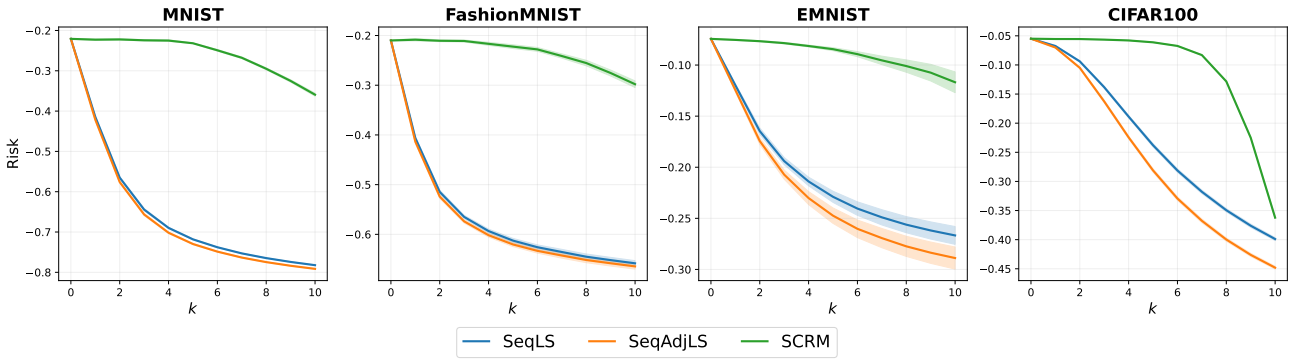

Figure 1: $R(\pi_k)$ at each intermediate step, for $k = 10$.

Table 1: $R(\pi)$ while fixing $N_k = N$, varying k in `MNIST` and `CIFAR100` (lower is better).

Dataset	k	`SeqLS` (1)	`SeqAdjLS` (2)
`MNIST`	1	−0.674	−0.678
	5	−0.758	−0.766
	10	−0.782	**−0.791**
	100	−0.781	**−0.789**
`CIFAR100`	1	−0.197	−0.207
	5	−0.344	−0.380
	10	−0.406	−0.455
	100	−0.508	**−0.558**

Table 1 summarizes the results: increasing k consistently improves final policy performance, with diminishing returns as π nears optimality (e.g., `MNIST` at $k = 10$ and 100). The gains are especially notable on `CIFAR100`, where the larger action space ($|\mathcal{A}| = 100$) makes uniform policies particularly suboptimal, thus highlighting the importance of the sequential updates. Finally, our results demonstrate the empirical advantage of `SeqAdjLS`, validating the need for our refined estimator. Across all settings, and especially the more challenging ones, `SeqAdjLS` consistently outperforms `SeqLS`, providing empirical evidence that the adjustment is essential for achieving accelerated convergence.

Benchmarks. Having demonstrated the benefits of intermediate updates and the effectiveness of our proposed `SeqAdjLS` algorithm, we now compare our methods against the recently introduced `SCRM` approach Zenati et al. [2023], which extends the `CRM` principle of Swaminathan and Joachims [2015] to the sequential setting. We evaluate all methods on `MNIST`, `FashionMNIST` [Xiao et al., 2017], `EMNIST` [Cohen et al., 2017], and `CIFAR100` with varying action space sizes, fixing the total number of interactions to N, setting $k = 10$, and using theoretically recommended values for λ to ensure a fair comparison.

`SeqLS` and `SeqAdjLS` divide the interaction budget uniformly across rounds ($n_j = N/k$), reusing all previously collected data to update policies. In contrast, `SCRM` employs exponentially increasing interaction sizes ($n_j = n_0 \cdot 2^j$ with $n_0 = \lceil N/2^k \rceil$) and updates policies using only the most recent batch, rather than the cumulative data (n_j instead of N_j). Each method is evaluated across six random seeds, and the results are summarized in Fig. 1.

We observe that `SCRM` tends to be overly conservative, particularly in early rounds when the number of interactions is small. Although it improves upon the initial behavior policy π_0 over time, it lags significantly behind both of our sequential methods. While it narrows the gap with `SeqLS` on `CIFAR100`, it is consistently outperformed, especially by `SeqAdjLS`, which achieves superior results across all datasets. These findings further confirm the empirical advantages of our approach. Additional experimental details and results are provided in Appendix E.

6 CONCLUSION

In this work, we extended the Logarithmic Smoothing learning principle to the sequential setting. We proposed two novel sequential algorithms for off-policy learning and established convergence guarantees to their respective optima. One of these algorithms enjoys accelerated convergence under mild assumptions and demonstrated improved empirical performance. These theoretical and practical benefits highlight the potential of combining online PAC-Bayes techniques with sequential contextual bandits. Our theoretical analysis requires our policy class to cover the optimal actions, thus weakening standard uniform coverage assumptions. An interesting open question is whether the structure of the pessimistic update can be exploited to relax this requirement. Moreover, in Appendix C, we showed that online PAC-Bayes techniques yield learning algo-

rithms that relax the *i.i.d.* assumption on contexts, thereby addressing real-world scenarios with distribution shifts. A natural and promising direction is to investigate whether the (accelerated) convergence guarantees established here extend to this more general, non-stationary setting.

Acknowledgements M.H. is partially supported by the European Research Council Starting Grant DYNASTY – 101039676. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

P. Alquier. User-friendly Introduction to PAC-Bayes Bounds. *Foundations and Trends® in Machine Learning*, 2024.

R. Andreeva, B. Dupuis, R. Sarkar, T. Birdal, and U. Şimşekli. Topological Generalization Bounds for Discrete-Time Stochastic Optimization Algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

I. Aouali, V.-E. Brunel, D. Rohde, and A. Korba. Exponential Smoothing for Off-Policy Learning. In *Proceedings of the 40th International Conference on Machine Learning*, pages 984–1017. PMLR, 2023.

I. Aouali, V.-E. Brunel, D. Rohde, and A. Korba. Unified pac-bayesian study of pessimism for offline policy learning with regularized importance sampling. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, UAI '24. JMLR.org, 2024.

A. Bibaut, N. Kallus, M. Dimakopoulou, A. Chambaz, and M. van der Laan. Risk minimization from adaptively collected data: guarantees for supervised and policy learning. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

L. Bottou, J. Peters, J. Quiñonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(11), 2013.

O. Catoni. A PAC-Bayesian approach to adaptive classification. *preprint*, 840, 2003.

O. Catoni. *PAC-Bayesian supervised classification: the thermodynamics of statistical learning.* Institute of Mathematical Statistics, 2007.

N. Chopin and O. Papaspiliopoulos. *An introduction to Sequential Monte Carlo / Nicolas Chopin, Omiros Papaspiliopoulos.* Springer Series in Statistics. Springer, Cham, Switzerland, 1st ed. 2020. edition, 2020. ISBN 3-030-47845-9.

B. Chugg, H. Wang, and A. Ramdas. A Unified Recipe for Deriving (Time-Uniform) PAC-Bayes Bounds. *Journal of Machine Learning Research*, 2023. URL http://jmlr.org/papers/v24/23-0401.html.

G. Cohen, S. Afshar, J. Tapson, and A. Van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 international joint conference on neural networks (IJCNN)*, pages 2921–2926. IEEE, 2017.

I. Csiszár. *I*-Divergence Geometry of Probability Distributions and Minimization Problems. *The Annals of Probability*, 1975.

M. D. Donsker and S. R. S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time—III. *Communications on Pure and Applied Mathematics*, 1976.

J. Doob. Jean Ville, Étude Critique de la Notion de Collectif. *Bulletin of the American mathematical society*, 45(11):824–824, 1939.

M. Dudik, D. Erhan, J. Langford, and L. Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014.

G. K. Dziugaite and D. Roy. Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017.

M. M. Fard and J. Pineau. PAC-Bayesian model selection for reinforcement learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2010.

G. Gabbianelli, G. Neu, and M. Papini. Importance-Weighted Offline Learning Done Right. In *Proceedings of The 35th International Conference on Algorithmic Learning Theory*, volume 237 of *Proceedings of Machine Learning Research*, pages 614–634. PMLR, 25–28 Feb 2024.

B. Guedj. A Primer on PAC-Bayesian Learning. In *Proceedings of the second congress of the French Mathematical Society*, 2019.

M. Haddouche and B. Guedj. Online PAC-Bayes Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

M. Haddouche and B. Guedj. PAC-Bayes Generalisation Bounds for Heavy-Tailed Losses through Supermartingales. *Transactions on Machine Learning Research*, 2023.

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

Y. Hu, N. Kallus, and X. Mao. Fast rates for contextual linear optimization. *Manage. Sci.*, 68(6):4236–4245, June 2022. ISSN 0025-1909. doi: 10.1287/mnsc.2022.4383. URL https://doi.org/10.1287/mnsc.2022.4383.

Y. Jin, Z. Yang, and Z. Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021.

N. Kallus and A. Zhou. Policy evaluation and optimization with continuous treatments. In *International conference on artificial intelligence and statistics*, pages 1243–1251. PMLR, 2018.

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

A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Y. LeCun, C. Cortes, and C. Burges. MNIST Handwritten Digit Database. http://yann.lecun.com/exdb/mnist, 2010.

B. London and T. Sandler. Bayesian counterfactual risk minimization. In *International Conference on Machine Learning*, pages 4125–4133. PMLR, 2019.

D. A. McAllester. Some PAC-Bayesian theorems. In *Proceedings of the eleventh annual conference on Computational Learning Theory*, pages 230–234. ACM, 1998.

D. A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational Learning Theory*, pages 164–170. ACM, 1999.

D. A. McAllester. PAC-Bayesian Stochastic Model Selection. *Machine Learning*, 2003.

A. M. Metelli, A. Russo, and M. Restelli. Subgaussian and differentiable importance sampling for off-policy evaluation and learning. *Advances in Neural Information Processing Systems*, 34:8119–8132, 2021.

W. Mou, L. Wang, X. Zhai, and K. Zheng. Generalization Bounds of SGLD for Non-convex Learning: Two Theoretical Viewpoints. In *Conference On Learning Theory (COLT)*, 2018.

K. Nozawa, P. Germain, and B. Guedj. PAC-Bayesian contrastive unsupervised representation learning. In *Conference on Uncertainty in Artificial Intelligence [UAI]*, 2020.

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

M. Perez-Ortiz, O. Rivasplata, J. Shawe-Taylor, and C. Szepesvari. Tighter Risk Certificates for Neural Networks. *Journal of Machine Learning Research*, 2021.

O. Sakhi, S. Bonner, D. Rohde, and F. Vasile. BLOB: A probabilistic model for recommendation that combines organic and bandit signals. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 783–793, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3403121.

O. Sakhi, P. Alquier, and N. Chopin. PAC-Bayesian Offline Contextual Bandits with Guarantees. In *International Conference on Machine Learning*, pages 29777–29799. PMLR, 2023a.

O. Sakhi, D. Rohde, and A. Gilotte. Fast Offline Policy Optimization for Large Scale Recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(8):9686–9694, Jun. 2023b.

O. Sakhi, I. Aouali, P. Alquier, and N. Chopin. Logarithmic Smoothing for Pessimistic Off-Policy Evaluation, Selection and Learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

J. Shawe-Taylor and R. C. Williamson. A PAC analysis of a Bayes estimator. In *Proceedings of the 10th annual conference on Computational Learning Theory*, pages 2–9. ACM, 1997.

Y. Su, M. Dimakopoulou, A. Krishnamurthy, and M. Dudík. Doubly robust off-policy evaluation with shrinkage. In *International Conference on Machine Learning*, pages 9167–9176. PMLR, 2020.

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

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

H. Zenati, E. Diemert, M. Martin, J. Mairal, and P. Gaillard. Sequential counterfactual risk minimization. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

R. Zhan, Z. Ren, S. Athey, and Z. Zhou. Policy learning with adaptively collected data, 2022. URL https://arxiv.org/abs/2105.02344.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] We provide all proofs and algorithm statement either in the main text or appendices.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes] We study convergence guarantees of our algorithms as well as providing PAC-Bayes bound which justifies their existence.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] All results are stated with respect to our detailed theoretical setup of Sec. 2.
   (b) Complete proofs of all theoretical results. [Yes] All proofs are detailed in the appendix due to space constraints.
   (c) Clear explanations of any assumptions. [Yes] We explain and contextualize our various assumptions in Secs. 2 to 4.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] Code lies in the supplementary material.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] All practical details are gathered in Sec. 5 and Appendix E.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes] Such details are retrievable in Sec. 5 and Appendix E.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes] All experiments were run on a single machine. Such details are retrievable in Appendix E.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes] We mention all the inspirations we exploited as well as references of the datasets we used in Sec. 5.
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable] We only exploited standard datasets such as MNIST and CIFAR100, publicly available and not requiring any specific consent to be used.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

APPENDIX

A EXTENDED RELATED WORK

Off-Policy Learning. Off-Policy Learning in the contextual bandit setting, or counterfactual risk minimization [Swaminathan and Joachims, 2015] is a widely adopted framework for learning from past interactions. In most real world applications, large datasets summarizing past interactions are often available, allowing agents to improve their policies offline. In this setting, the goal is to find the optimal policy using previously collected interactions. Two primary approaches exist for learning from this data. The direct method [Sakhi et al., 2020] focuses on modeling the reward and subsequently derives a straightforward policy. This approach is well-understood, and the quality of the learned policy depends entirely on our ability to accurately model the reward. The other approach directly learns a policy from the data by minimizing an importance weighting estimator [Dudik et al., 2014, Bottou et al., 2013] of the expected cost. The research community has predominantly focused on this second approach, as it presents significant theoretical challenges and offers greater generality by not relying on any reward modeling. At the core of directly learning a policy is the development of tailored off-policy estimators with improved concentration properties [Bottou et al., 2013, Aouali et al., 2023, Metelli et al., 2021, Sakhi et al., 2024], combined with tight generalization bounds allowing efficient learning [Swaminathan and Joachims, 2015, Sakhi et al., 2023a, 2024]. The used methods are motivated by the pessimism principle [Jin et al., 2021] that was proven to be optimal for learning in these scenarios. Particularly, analyzing this learning problem from the PAC-Bayesian learning theory lens London and Sandler [2019] proved efficient, improving the seminal work of [Swaminathan and Joachims, 2015]. This direction was further explored by Sakhi et al. [2023a], Aouali et al. [2023], Gabbianelli et al. [2024], Sakhi et al. [2024], reducing assumptions and improving the pessimistic principle, with latest developments [Sakhi et al., 2024] resulting in state of the art pessimistic approaches. We further develop this research direction and generalize it to the sequential setting, where an online algorithm redeploys the learned policy to the environment and collect more interactions to train the next iteration.

Sequential Off-Policy Learning. Sequential policy learning extends off-policy learning by allowing the decision-maker to deploy a learned policy, collect additional interactions, and then train the next policy on the entire set of previously gathered data. This approach is widely used in practice since decision systems are continuously updated. Nevertheless, its theoretical foundations remain limited and underexplored, and only a few works have addressed this question.

A related direction is off-policy learning from adaptively collected data, studied in Zhan et al. [2022] and Bibaut et al. [2021]. These works focus on purely offline settings, where data is collected under multiple logging policies with decaying exploration. However, they remain one-shot batch approaches: their goal is to learn a single policy from the collected interactions, without analyzing the effects of redeployment. In contrast, we consider sequential updates and redeployments that occur after the arrival of batches of arbitrary sizes. Our analysis characterizes the conditions under which sequential deployment achieves fast convergence rates. This framework more accurately models real-world scenarios, provides theoretical insight into a widely used practical approach, and unifies the static (fully batch) and dynamic (fully online) regimes. Our proposed algorithms match the convergence rates of the batch setting and further accelerate learning as the behavior policy becomes refined.

Recently, Zenati et al. [2023] extended *Counterfactual Risk Minimization* [Swaminathan and Joachims, 2015] to the sequential setting, allowing deployments of learned policies within the same framework as ours. Their algorithm achieves accelerated convergence under a Hölderian assumption, but this assumption is not always justified. Moreover, it requires bounded importance weights, which in discrete action spaces is equivalent to assuming uniform coverage. Their method further relies on unrealistic, exponentially growing batch sizes and only uses the most recent batch of data, discarding all past interactions. These limitations arise from the theoretical tools employed, making the algorithm impractical. By contrast, we develop a PAC-Bayesian approach that avoids these shortcomings. Our algorithms are the first to seamlessly interpolate between full batch and full online regimes, achieving the optimal convergence rates of both while requiring only coverage of the optimal actions.

PAC-Bayes learning. PAC-Bayes learning is a recent branch of learning theory which emerged in the late 90s via the seminal work of [Shawe-Taylor and Williamson, 1997, McAllester, 1998, 1999, 2003] and later pursued by [Catoni, 2003, 2007]. Modern surveys are available to describe the various advances in the field [Guedj, 2019, Alquier, 2024]. PAC-Bayes provide generalization bounds involving a complexity term (often a KL divergence), inspired here from the Bayesian learning paradigm of designing a *posterior* knowledge of the learning problem based on both training data and a *prior* knowledge of the considered situation. On the contrary, PAC-Bayes, while inspired from the Bayesian philosophy, relies historically on tools from information theory. This general approach benefits from additional flexibility as PAC-Bayes can be linked and applied to Bayesian learning (see Guedj, 2019) but also blurs the notion of prior and posterior distributions, now independent of the fundamental Bayes formula. This flexibility allowed PAC-Bayes to reach various sub-fields of learning theory: optimization dynamics of learning algorithms [Mou et al., 2018, Andreeva et al., 2024], reinforcement learning [Fard and Pineau, 2010], online learning [Haddouche and Guedj, 2022] and contrastive learning [Nozawa et al., 2020] to name a few.

B PROOFS OF SECTION 3

In this whole section, we assume that all datasets S_k have the same cardinal $n_k = m$. This section contains the proof of Thm. 3.2 and Proposition 3.3.

B.1 Proof of Theorem 3.2

We start by restating the theorem of interest:

Theorem 3.2. *For any sequential algorithm* `Alg`*, any $m > 0$ and data independent prior P. For any $\lambda > 0, \delta \in (0,1]$, with probability at least $1-\delta$ over S, we have, for all $k \geq 0$, denoting by π_{k+1} the output of* `Alg` *at time k, for all $Q \in \mathcal{P}(\Theta)$:*

$$R(\pi_Q) \leq \hat{R}_{0:k}^{\lambda-\mathrm{LS}}(\pi_\mathrm{Q}) + \frac{\mathrm{KL}(Q,P) + \log\left(\frac{1}{\delta}\right)}{\lambda(k+1)m}.$$

Proof of Thm. 3.2. For any $k \geq 0$, let:

$$f_k(S,\theta) := \sum_{j=0}^{k} \underbrace{\sum_{i \in I_k} -\log\left(1-\lambda R(d_\theta)\right) + \log\left(1 - \lambda \frac{d_\theta(a_i \mid x_i)c_i}{\pi_j(a_i \mid x_i)}\right)}_{:=Y_j(S,\theta)}.$$

Now for a given posterior Q and a prior P, we properly apply the change of measure inequality for any m:

$$\mathbb{E}_{\theta\sim Q}\left[f_k(S,\theta)\right] \leq \mathrm{KL}(Q,P) + \log\left(\mathbb{E}_{\theta\sim P}\exp(f_k(S,\theta))\right). \tag{1}$$

We need to control the exponential random variable. To do so we prove the following lemma:

Lemma B.1. *The sequence $(M_k := \mathbb{E}_{\theta\sim P}\exp(f_k(S,\theta))_{k\geq 0}$ is a non-negative martingale.*

Proof. We fix $k \geq 0$. We show that $\mathbb{E}_{k-1}[M_k] = M_{k-1}$.

$$\begin{aligned}
\mathbb{E}_{k-1}[M_k] &= \mathbb{E}_{k-1}\left[\mathbb{E}_{\theta\sim P}\exp(f_k(S,\theta)\right] \\
&= \mathbb{E}_{\theta\sim P}\left[\mathbb{E}_{k-1}\left[\exp\left(f_{k-1}(S,\theta)\right)\exp\left(\sum_{i\in I_k}\log\left(\frac{1-\lambda\frac{d_\theta(a_i|x_i)c_i}{\pi_k(a_i|x_i)}}{1-\lambda R(d_\theta)}\right)\right)\right]\right] \\
&= \mathbb{E}_{\theta\sim P}\left[\exp\left(f_{k-1}(S,\theta)\right)\mathbb{E}_{k-1}\left[\exp\left(\sum_{i\in I_k}\log\left(\frac{1-\lambda\frac{d_\theta(a_i|x_i)c_i}{\pi_k(a_i|x_i)}}{1-\lambda R(d_\theta)}\right)\right)\right]\right].
\end{aligned}$$

The second line comes from the application of Fubini as P is data-independent and last line already demonstrated. Now we have for any $\theta \in \Theta$:

$$\mathbb{E}_{k-1}\left[\exp\left(\sum_{i\in I_k}\log\left(\frac{1-\lambda\frac{d_\theta(a_i|x_i)c_i}{\pi_k(a_i|x_i)}}{1-\lambda R(d_\theta)}\right)\right)\right] = \mathbb{E}_{k-1}\left[\Pi_{i\in I_k}\frac{1-\lambda\frac{d_\theta(a_i|x_i)c_i}{\pi_k(a_i|x_i)}}{1-\lambda R(d_\theta)}\right]$$
$$= \Pi_{i\in I_k}\frac{1-\lambda\mathbb{E}_{k-1}\left[\frac{d_\theta(a_i|x_i)c_i}{\pi_k(a_i|x_i)}\right]}{1-\lambda R(d_\theta)},$$

these computations holding as we exploited the assumption that, conditionally to $\mathcal{F}_{k-1}$, $S_k = ((x_i, a_i, c_i))_{i\in I_k}$ is constituted of mutually independent random variables (thus an expectation of product equals a product of expectations). Now, recall that π_k is $\mathcal{F}_{k-1}$-measurable and that for all $i \in I_k, a_i \sim \pi_k(\cdot \mid x_i)$. We then remark that, for any θ, $i \in I_k$:

$$\begin{aligned}\mathbb{E}_{k-1}\left[\frac{d_\theta(a_i \mid x_i)c_i}{\pi_k(a_i \mid x_i)}\right] &= \mathbb{E}_{x_i'\sim\nu}\mathbb{E}_{a_i'\sim\pi_k(\cdot|x_i')}\mathbb{E}_{c_i'\sim p(\cdot|x_i',a_i')}\left[\frac{d_\theta(a_i' \mid x_i')c_i'}{\pi_k(a_i' \mid x_i')}\right]\\ &= \mathbb{E}_{x_i'\sim\nu}\mathbb{E}_{a_i'\sim\pi_k(\cdot|x_i')}\left[\frac{d_\theta(a_i' \mid x_i')}{\pi_k(a_i' \mid x_i')}\mathbb{E}_{c_i'\sim p(\cdot|x_i',a_i')}\left(c_i'\right)\right]\\ &= \mathbb{E}_{x_i'\sim\nu}\mathbb{E}_{a_i'\sim\pi_k(\cdot|x_i')}\left[c\left(x_i', d_\theta(x_i')\right)\right]\\ &= R(d_\theta)\end{aligned}$$

Thus $\mathbb{E}_{k-1}[M_k] = M_{k-1}$, this concludes the lemma's proof. □

Now; given a martingale is also a supermartingale, we can apply Ville's inequality which implies that with probability at least $1-\delta$, for any $k \geq 0$:

$$\mathbb{E}_{\theta\sim P}\exp(f_k(S,\theta)) \leq \frac{1}{\delta}.$$

Plugging this back into Eq. (1) yields, with probability at least $1-\delta$ over S, for any $k \geq 0$, any posterior Q:

$$\sum_{j=0}^{k}\mathbb{E}_{\theta\sim Q}\left[\sum_{i\in I_k} -\log\left(1-\lambda R(d_\theta)\right) + \log\left(1-\lambda\frac{d_\theta(a_i \mid x_i)c_i}{\pi_j(a_i \mid x_i)}\right)\right] \leq \mathrm{KL}(Q,P) + \log\left(\frac{1}{\delta}\right).$$

Now remark that for any $j \geq 0, i \in I_j, \theta$ because $d_\theta(a,x) = \mathbb{1}\{f_\theta(x) = a\} \in \{0,1\}$, we have for any (a,x):

$$\log\left(1-\lambda\frac{d_\theta(a \mid x)c_i}{\pi_j(a \mid x)}\right) = d_\theta(a,x)\log\left(1-\lambda\frac{c_i}{\pi_j(a \mid x)}\right).$$

We then have:

$$\begin{aligned}\sum_{j=1}^{k}\mathbb{E}_{\theta\sim Q}\left[\sum_{i\in I_j} -\log\left(1-\lambda\frac{d_\theta(a_i \mid x_i)c_i}{\pi_j(a_i \mid x_i)}\right)\right] &= -\sum_{j=0}^{k}\sum_{i\in I_j}\pi_Q(a_i \mid x_i)\log\left(1-\lambda\frac{c_i}{\pi_j(a_i \mid x_i)}\right)\\ &= (k+1)m\lambda\hat{R}_{0:k}^{\lambda-\mathrm{LS}}(\pi_Q).\end{aligned}$$

We also use the convexity of $x \to -\log(1-\lambda x)$ to obtain:

$$\begin{aligned}\sum_{j=0}^{k}\mathbb{E}_{\theta\sim Q}\left[\sum_{i\in I_j} -\log\left(1-\lambda R(d_\theta)\right)\right] &= m(k+1)\mathbb{E}_{\theta\sim Q}\left[-\log\left(1-\lambda R(d_\theta)\right)\right]\\ &\geq -m(k+1)\log\left(1-\lambda R(\pi_Q)\right).\end{aligned}$$

Finally, mixing all those calculations together, dividing by $\lambda(k+1)m$ and applying the increasing function $\psi_\lambda(x) := \frac{1}{\lambda}(1-\exp(-\lambda x)$ gives:

$$R(\pi_Q) \leq \psi_\lambda \left(\hat{R}_{0:k}^{\lambda-\mathrm{LS}}(\pi_\mathrm{Q}) + \frac{\mathrm{KL}(Q,P) + \log\left(\frac{1}{\delta}\right)}{\lambda(k+1)m} \right).$$

Finally, using that $\psi_\lambda(x) \leq x$ concludes the proof. □

B.2 Proof of Proposition 3.3

We prove the convergence guarantee of Algorithm 1.

Proposition 3.3. *For any fixed subbatch size $m > 0$, any data independent prior P, any $\lambda > 0, \delta \in (0,1]$, with probability at least $1-\delta$, for any $k \geq 0$, denoting by π_{k+1} the output of Algorithm 1 at time k and $\pi_\star = \pi_{Q_\star}$ with $Q_\star \in \mathrm{argmin}_{Q\in\mathcal{C}(\Theta)} R(\pi_Q)$:*

$$R\left(\pi_{k+1}\right) - R\left(\pi_\star\right) \leq \frac{\lambda}{k+1}\sum_{j=0}^{k} \mathcal{S}_j\left(\pi_\star\right) + 2\frac{\mathrm{KL}(Q^\star,P) + \log\left(\frac{2}{\delta}\right)}{\lambda(k+1)m}$$

where $\mathcal{S}_j(\pi) := \mathbb{E}_{(x,a,c)\sim P(\pi_j)}\left[\frac{\pi(a|x)c^2}{\pi_j^2(a|x)}\right]$.

Proof. We fix $k \geq 0$, a sequential algorithm `Alg` and P a data-free prior. To highlight the dependency on λ, we denote in this proof $\mathcal{S}_j = \mathcal{S}_{\lambda,j}$ Our goal is to apply the change of measure inequality on $\mathcal{H}$ to a specific function f_k. We define this function below, for any sample S and any predictor θ

$$f_k(S,\theta) := \sum_{j=0}^{k} \underbrace{\sum_{i\in I_k} -\lambda R(d_\theta) - \lambda^2 \mathcal{S}_{\lambda,j}\left(d_\theta\right) - \log\left(1 - \lambda\frac{d_\theta(a_i \mid x_i)c_i}{\pi_j(a_i \mid x_i)}\right)}_{:=Y_j(S,\theta)},$$

where $\mathcal{S}_{\lambda,j}(\pi) := \mathbb{E}_{(x,a,c)\sim P(\pi_j)}\left[\frac{\pi(a|x)c^2}{\pi_j^2(a|x) - \lambda\pi_j(a|x)c}\right]$ Now for a given posterior Q we apply the change of measure inequality for any m:

$$\sum_{j=0}^{k} \mathbb{E}_{\theta\sim Q}\left[Y_j(S,\theta)\right] = \mathbb{E}_{\theta\sim Q}\left[f_k(S,\theta)\right] \leq \mathrm{KL}(Q,P) + \log\left(\mathbb{E}_{\theta\sim P}\exp(f_k(S,\theta))\right). \tag{2}$$

We need to control the exponential random variable. To do so we prove the following lemma:

Lemma B.2. *The sequence $(M_k := \mathbb{E}_{\theta\sim P}\exp(f_k(S,\theta))_{k\geq 0}$ is a non-negative supermartingale.*

Proof. We fix $k \geq 0$ and we recall that for any k, P is data free. We show that $\mathbb{E}_{k-1}[M_k] \leq M_{k-1}$.

$$\begin{aligned}
\mathbb{E}_{k-1}[M_k] &= \mathbb{E}_{k-1}\left[\mathbb{E}_{\theta\sim P}\exp(f_k(S,\theta)\right] \\
&= \mathbb{E}_{\theta\sim P}\left[\mathbb{E}_{k-1}\exp(f_k(S,\theta)\right] \\
&= \mathbb{E}_{\theta\sim P}\left[\mathbb{E}_{k-1}\left[\exp\left(f_{k-1}(S,\theta)\right)\exp\left(Y_j(S,\theta)\right)\right]\right] \\
&= \mathbb{E}_{\theta\sim P}\left[\exp\left(f_{k-1}(S,\theta)\right)\mathbb{E}_{k-1}\left[\exp\left(Y_j(S,\theta)\right)\right]\right].
\end{aligned}$$

The second line holding, as P is data free, by Fubini.

Now we note $c = c(a, x)$ for the sake of conciseness. For any $K, \theta \in \mathbb{R}^d$, using the fact that conditionally to $\mathcal{F}_{k-1}$, all the data in $\mathcal{S}_k$ are independent we have:

$$\mathbb{E}_{k-1}\left[\exp\left(Y_j(S,\theta)\right]\right. = \prod_{i \in I_k} \exp\left(-\lambda\left(R\left(d_\theta\right) - \lambda \mathcal{S}_{\lambda,k}\left(d_\theta\right)\right)\right) \mathbb{E}_{k-1}\left[\frac{1}{1-\lambda\frac{d_\theta(a|x)c}{\pi_k(a|x)}}\right]$$

Using $\log(x) \leq x - 1$ yields

$$\begin{aligned}
&\leq \prod_{i \in I_k} \exp\left(-\lambda\left(R\left(d_\theta\right) - \lambda \mathcal{S}_{\lambda,k}\left(d_\theta\right)\right) + \mathbb{E}_{k-1}\left[\frac{1}{1-\lambda\frac{d_\theta(a|x)c}{\pi_k(a|x)}}\right] - 1\right) \\
&= \prod_{i \in I_k} \exp\left(-\lambda\left(R\left(d_\theta\right) - \lambda \mathcal{S}_{\lambda,k}\left(d_\theta\right)\right) + \mathbb{E}_{k-1}\left[\frac{\lambda d_\theta(a \mid x)c}{\pi_k(a \mid x) - \lambda d_\theta(a \mid x)c}\right]\right)
\end{aligned}$$

Using that d_θ is binary gives the equality $\frac{\lambda d_\theta(a|x)c}{\pi_k(a|x)-\lambda d_\theta(a|x)c} = \frac{\lambda d_\theta(a|x)c}{\pi_k(a|x)-\lambda c}$, thus

$$\begin{aligned}
&= \prod_{i \in I_k} \left(\exp\left(-\lambda\left(R\left(d_\theta\right) - \lambda \mathcal{S}_{\lambda,k}\left(d_\theta\right)\right) + \mathbb{E}_{k-1}\left[\frac{\lambda d_\theta(a \mid x)c}{\pi_k(a \mid x) - \lambda c}\right]\right)\right) \\
&= \prod_{i \in I_K} \left(\exp\left(-\lambda^2 \mathcal{S}_{\lambda,k}\left(d_\theta\right) + \mathbb{E}_{k-1}\left[\frac{\lambda d_\theta(a \mid x)c}{\pi_k(a \mid x) - \lambda c} - \lambda\frac{d_\theta(a \mid x)c}{\pi_k(a \mid x)}\right]\right)\right) \\
&\leq \prod_{i \in I_k} \left(\exp\left(-\lambda^2 \mathcal{S}_{\lambda,k}\left(d_\theta\right) + \lambda^2 \mathcal{S}_{\lambda,k}\left(d_\theta\right)\right)\right) \\
&\leq 1.
\end{aligned}$$

This allows concluding that $\mathbb{E}_{k-1}[M_k] \leq M_{k-1}$, thus the proof. □

Now, we can apply Ville's inequality which implies that with probability at least $1 - \delta$, for any $K \geq 1$:

$$\mathbb{E}_{\theta \sim P} \exp(f_k(S, \theta)) \leq \frac{1}{\delta}.$$

Plugging this back into Eq. (2) yields, with probability at least $1 - \delta$ over S, for any $k \geq 0$, any Q:

$$\mathbb{E}_{\theta \sim Q}\left[\sum_{j=0}^{k} \sum_{i \in I_k} -\lambda R(d_\theta) - \lambda \mathcal{S}_{\lambda,j}\left(d_\theta\right) - \log\left(1 - \lambda\frac{d_\theta(a_i \mid x_i)c_i}{\pi_j(a_i \mid x_i)}\right)\right] \leq \mathrm{KL}(Q, P) + \log\left(\frac{1}{\delta}\right).$$

Then recall that

$$\hat{R}_{0:k}^{\lambda-\mathrm{LS}}(\pi_Q) := -\frac{1}{(k+1)m\lambda} \sum_{j=1}^{k} \sum_{i \in I_j} \pi_Q(a_i \mid x_i) \log\left(1 - \lambda\frac{c_i}{\pi_j(a_i \mid x_i)}\right).$$

We have, using again that d_θ is binary for any θ:

$$\begin{aligned}
\mathbb{E}_{\theta \sim Q}\left[\sum_{j=0}^{k} \sum_{i \in I_k} -\log\left(1 - \lambda\frac{d_\theta(a_i \mid x_i)c_i}{\pi_i^0(a_i \mid x_i)}\right)\right] &= -\sum_{j=0}^{k} \sum_{i \in I_k} \pi_Q(a_i \mid x_i) \log\left(1 - \lambda\frac{c_i}{\pi_j(a_i \mid x_i)}\right) \\
&= (k+1)m\lambda \hat{R}_{0:k}^{\lambda-\mathrm{LS}}(\pi_Q).
\end{aligned}$$

Mixing all those calculations together and dividing by $\lambda(k+1)m$ gives: with probability at least $1-\delta$, for any k, Q:

$$\hat{R}_{0:k}^{\lambda-\mathrm{LS}}(\pi_Q) \leq R(\pi_Q) + \frac{\lambda}{k+1}\sum_{j=0}^{k}\mathcal{S}_{\lambda,j}\left(\pi_Q\right) + \frac{\mathrm{KL}(Q,P)+\log\left(\frac{1}{\delta}\right)}{\lambda(k+1)m} \tag{3}$$

Now, we consider the optimal sequence, defined such that, $Q^\star := \mathrm{argmin}_Q\, R(\mathrm{Q})$.

We also consider π_{k+1} the output of Algorithm 1 time k satisfying:

$$\pi_{k+1} = \pi_{\hat{Q}_k}, \quad \text{with } \hat{Q}_k := \mathrm{argmin}_Q\, \hat{R}_{0:k}^{\lambda-\mathrm{LS}}(\pi_Q) + \frac{\mathrm{KL}(Q,P)}{\lambda(k+1)m}$$

Then using Thm. 3.2, and the definition of $\hat{Q}$ yields for any λ, data-free P with probability at least $1-\delta/2$, for any k:

$$\begin{aligned} R\left(\pi_{k+1}\right) &\leq \hat{R}_{0:k}^{\lambda-\mathrm{LS}}(\pi_{k+1}) + \frac{\mathrm{KL}(\hat{Q}_k,P)+\log\left(\frac{2}{\delta}\right)}{\lambda(k+1)m} \\ &\leq \hat{R}_{0:k}^{\lambda-\mathrm{LS}}(\pi_\star) + \frac{\mathrm{KL}(Q^\star,P)+\log\left(\frac{2}{\delta}\right)}{\lambda(k+1)m} \end{aligned}$$

Then, exploiting Eq. (3) with probability $1-\delta/2$ and taking a union bound yields, with probability at least $1-\delta$, for any k:

$$R\left(\pi_{k+1}\right) - R\left(\pi_\star\right) \leq \frac{\lambda}{k+1}\sum_{j=0}^{k}\mathcal{S}_{\lambda,j}\left(\pi_\star\right) + 2\frac{\mathrm{KL}(Q^\star,P)+\log\left(\frac{2}{\delta}\right)}{\lambda(k+1)m}. \tag{4}$$

Finally, using that $\mathcal{S}_{\lambda,j} \leq \mathcal{S}_j$ for all j concludes the proof. □

C SEQUENTIAL LEARNING ALGORITHMS WITH NON-IID CONTEXTS.

In the main text, we always considered an *i.i.d.* assumption on the contexts, which may be restrictive when considering real-life situations such as the behavior of an individual, susceptible to evolve through time. In what follows, *we make no assumption on the context distribution and only assume that at time $k \in \{1\cdots,K\}$, conditionally to $\mathcal{F}_{k-1}$, $S_k = ((x_i,a_i,c_i))_{i\in I_k}$ is constituted of mutually independent random variable.* We show that we are still able to obtain a generalization bound, thanks to the supermartingale toolbox of Haddouche and Guedj [2022, 2023], Chugg et al. [2023]. We first start with some additional framework.

Framework Assume that batches are of size m. We define, for all k, the risk at time k of π_k as

$$R_k(\pi_k) = \mathbb{E}_{x'_k\sim\nu_k, a'_k\sim\pi_k(\cdot|x_k)}[c(x'_k,a'_k)],$$

where ν_k denotes, at time k, the conditional distribution of contexts with respect to the past (*i.e.* $\mathcal{F}_{k-1}$) and x'_k, a'_k are independent of x_k, a_k. We also define, for any π_k, π_{k-1}, the empirical risk at time k as

$$\hat{R}_k^{\lambda-LS}(\pi_k,\pi_{k-1}) := \frac{1}{m}\sum_{i\in I_k} h_{LS}^{\lambda-Lin}(\pi_k(a_i \mid x_i), \pi_{k-1}(a_i \mid x_i), c_i),$$

with $h_{LS}^{\lambda-Lin}$ defined in Sec. 3. Given context distributions are evolving, we aim to learn a sequence of posteriors Q_k which is $\mathcal{F}_k$- measurable from an initial prior $P = Q_0$ to make the cumulative risk $\frac{1}{K}\sum_{k=0}^{K} R_k(\pi_{Q_k})$ small.

Adapting our proof technique yields the following result.

Theorem C.1. *For any subbatch size $m > 0$, any $\lambda > 0, \delta \in (0, 1]$, any learning algorithm $\mathcal{A}$ such that for all k, $Q_k = \mathcal{A}(S_1, \cdots, S_k)$ with probability at least $1 - \delta$ over S, for any $K \geq 1$:*

$$\frac{1}{K+1}\sum_{k=0}^{K} R_k(\pi_{Q_k}) \leq \frac{1}{K+1}\sum_{k=0}^{K} \hat{R}_k^{\lambda-LS}(\pi_{Q_k}, \pi_{Q_{k-1}}) + \frac{\sum_{k=0}^{K} \mathrm{KL}(Q_k, Q_{k-1}) + \log\left(\frac{1}{\delta}\right)}{\lambda(K+1)m}.$$

This is dynamic counterpart of Thm. 3.2, allowing shifting context distribution. Note that, similarly to Thm. 3.2, this bound suggests a new learning algorithm detailed in Algorithm 3. This algorithm only exploits S_k and $\pi_{Q_{k-1}}$ to create π_{Q_k}, contrary to Algorithm 1 which exploits all previous data $S_0 \cup \cdots, \cup S_k$. The reason behind this discrepancy is that past data gathers outdated information as the context distribution may arbitrarily evolve. Then our theoretical result suggests to only consider current data. Note that the toolbox we use also allows an alternative result to Thm. 4.1 for non-iid contexts, exploiting our novel regularized IPS.

Algorithm 3: Sequential Policy Learning via Logarithmic Smoothing for shifting context distributions

Input: Behavior policy $\pi_{Q_{-1}}$, a policy class $\mathcal{C}(\Theta)$, $\lambda > 0$.
Initialize P such that $\pi_P = \pi_{Q_{-1}}$ and $\mathcal{S}_{-1} = \{\}$.
for $k \geq 0$ **do**
  Collect data S_k of size $|I_k| = m$ with π_k.
  Solve the optimization problem:

$$\hat{Q}_{k+1} = \mathrm{argmin}_{Q \in \mathcal{C}(\Theta)} \left\{ \hat{R}_k^{\lambda-LS}(\pi_{Q_k}, \pi_{Q_{k-1}}) + \frac{\mathrm{KL}(Q_k, Q_{k-1})}{\lambda m} \right\}$$

  Set $\pi_{k+1} = \pi_{\hat{Q}_{k+1}}$.

We then detail the proof of Thm. C.1.

Proof. First, we follow the route described in Haddouche and Guedj [2023], Section 3.

For any $K > 0$, we look for a K-tuple of probabilities. Thus, our predictor set of interest is $\mathcal{H}_K := \mathcal{H}^{\otimes K}$ and then, our predictor θ^K is a tuple $(\theta_1, .., \theta_K) \in \mathcal{H}_K$.

Our goal is to apply the change of measure inequality on $\mathcal{H}_K$ to a specific function f_K. We define this function below, for any sample S and any predictor $\theta^K = (\theta_1, ..., \theta_K)$

$$f_K(S, \theta^K) := \sum_{k=0}^{K} \underbrace{\sum_{i \in I_k} -\log\left(1 - \lambda R_k(d_{\theta_k})\right) + \log\left(1 - \lambda \frac{d_{\theta_k}(a_i \mid x_i) c_i}{\pi_{Q_{k-1}}(a_i \mid x_i)}\right)}_{:= Y_k(S, \theta_k)}.$$

Now for a given posterior tuple $Q_1, ...Q_K$ we define $Q^K = Q_1 \otimes ... \otimes Q_K$ and also $P^K = Q_0 \otimes ... \otimes Q_{K-1}$. We can now properly apply the change of measure inequality for any m:

$$\begin{aligned}\sum_{k=0}^{K} \mathbb{E}_{\theta_k \sim Q_k}\left[Y_k(S, \theta_k)\right] &= \mathbb{E}_{\theta^k \sim Q}\left[f_K(S, \theta^K)\right] \leq \mathrm{KL}(Q^K, P^K) + \log\left(\mathbb{E}_{\theta^K \sim P^K} \exp(f_K(S, \theta^K))\right) \\ &= \sum_{k=0}^{K} \mathrm{KL}(Q_k, Q_{k-1}) + \log\left(\mathbb{E}_{\theta^K \sim P^K} \exp(f_K(S, \theta^K))\right), \end{aligned} \tag{5}$$

where we define $f_K(S, \theta^K) := \sum_{k=0}^{K} Y_k(S, \theta_k)$.

We need to control the exponential random variable. To do so we prove the following lemma:

Lemma C.2. *The sequence* $(M_k := \mathbb{E}_{\theta^K \sim P^K} \exp(f_K(S, \theta^K))_{K \geq 0}$ *is a non-negative martingale.*

Proof. We fix $K \geq 1$ and we recall that for any k, Q_{k-1} is $\mathcal{F}_{k-1}$-measurable. We show that $\mathbb{E}_{K-1}[M_K] = M_{K-1}$.

$$\begin{aligned}
\mathbb{E}_{K-1}[M_K] &= \mathbb{E}_{K-1}\left[\mathbb{E}_{\theta^K \sim P^K} \exp(f_K(S, \theta^K)\right] \\
&= \mathbb{E}_{K-1}\left[\mathbb{E}_{\theta_1, .., \theta_K \sim Q_0 \otimes ... \otimes Q_{K-1}} \exp(f_K(S, \theta^K)\right] \\
&= \mathbb{E}_{K-1}\left[\mathbb{E}_{\theta_1, .., \theta_K \sim Q_0 \otimes ... \otimes Q_{K-1}}\left[\exp\left(f_{K-1}(S, \theta^{K-1})\right) \exp\left(\sum_{i \in I_K} \log\left(\frac{1 - \lambda \frac{d_{\theta_K}(a_i|x_i)c_i}{\pi_{Q_{K-1}}(a_i|x_i)}}{1 - \lambda R_K(d_{\theta_K})}\right)\right)\right]\right] \\
&= M_{K-1}\mathbb{E}_{K-1}\left[\mathbb{E}_{\theta_K \sim Q_{K-1}}\left[\Pi_{i \in I_K} \frac{1 - \lambda \frac{d_{\theta_K}(a_i|x_i)c_i}{\pi_{Q_{K-1}}(a_i|x_i)}}{1 - \lambda R_K(d_{\theta_K})}\right]\right].
\end{aligned}$$

The last line holding because $P^{K-1} = Q_0 \otimes ... \otimes Q_{K-2}$ is $\mathcal{F}_{K-1}$ measurable.

Now we exploit the fact that Q_{K-1} is $\mathcal{F}_{K-1}$ measurable to apply a conditional Fubini lemma stated in Haddouche and Guedj [2022, Lemma D.3]. We have:

$$\begin{aligned}
\mathbb{E}_{K-1}\left[\mathbb{E}_{\theta_K \sim Q_{K-1}}\left[\Pi_{i \in I_K} \frac{1 - \lambda \frac{d_{\theta_K}(a_i|x_i)c_i}{\pi_{Q_{K-1}}(a_i|x_i)}}{1 - \lambda R_K(d_{\theta_K})}\right]\right] &= \mathbb{E}_{\theta_K \sim Q_{K-1}}\left[\mathbb{E}_{K-1}\left[\Pi_{i \in I_K} \frac{1 - \lambda \frac{d_{\theta_K}(a_i|x_i)c_i}{\pi_{Q_{K-1}}(a_i|x_i)}}{1 - \lambda R_K(d_{\theta_K})}\right]\right] \\
&= \mathbb{E}_{\theta_K \sim Q_{K-1}}\left[\Pi_{i \in I_K} \frac{1 - \lambda \mathbb{E}_{K-1}\left[\frac{d_{\theta_K}(a_i|x_i)c_i}{\pi_{Q_{K-1}}(a_i|x_i)}\right]}{1 - \lambda R_K(d_{\theta_K})}\right].
\end{aligned}$$

The last line holding as $R_K(d_{\theta_K})$ is $\mathcal{F}_{K-1}$-measurable for all K, θ_K. This also holds as we exploited the assumption that, conditionally to $\mathcal{F}_{K-1}$, $S_K = ((x_i, a_i, c_i))_{i \in I_K}$ is constituted of mutually independent random variable (thus an expectation of product equals a product of expectations). Now we remark that, for any θ_K, $i \in I_K$:

$$\begin{aligned}
\mathbb{E}_{K-1}\left[\frac{d_{\theta_K}(a_i \mid x_i)c_i}{\pi_{Q_{K-1}}(a_i \mid x_i)}\right] &= \mathbb{E}_{x_i' \sim \nu_K}\mathbb{E}_{a_i' \sim \pi_{Q_{K-1}}(\cdot|x_i')}\mathbb{E}_{c_i' \sim p(\cdot|x_i', a_i')}\left[\frac{d_{\theta_K}(a_i' \mid x_i')c_i'}{\pi_{Q_{K-1}}(a_i' \mid x_i')}\right] \\
&= \mathbb{E}_{x_i' \sim \nu_K}\mathbb{E}_{a_i' \sim \pi_{Q_{K-1}}(\cdot|x_i')}\left[\frac{d_{\theta_K}(a_i' \mid x_i')}{\pi_{Q_{K-1}}(a_i' \mid x_i')}\mathbb{E}_{c_i' \sim p(\cdot|x_i', a_i')}\left(c_i'\right)\right] \\
&= \mathbb{E}_{x_i' \sim \nu_K}\mathbb{E}_{a_i' \sim \pi_{Q_{K-1}}(\cdot|x_i')}\left[c\left(x_i', d_{\theta_K}(x_i')\right)\right] \\
&= R_K(d_{\theta_K})
\end{aligned}$$

Thus $\mathbb{E}_{K-1}[M_K] = M_{K-1}$, this concludes the lemma's proof. □

Now; given a martingale is also a supermartingale, we can apply Ville's inequality which implies that with probability at least $1 - \delta$, for any $k \geq 0$:

$$\mathbb{E}_{\theta^k \sim P^k} \exp(f_k(S, \theta)) \leq \frac{1}{\delta}.$$

Plugging this back into Eq. (5) yields, with probability at least $1 - \delta$ over S, for any $K \geq 0$,

$$\sum_{k=0}^{K} \mathbb{E}_{\theta_k \sim Q_k}\left[\sum_{i \in I_k} -\log\left(1 - \lambda R_k(d_{\theta_k})\right) + \log\left(1 - \lambda \frac{d_{\theta_k}(a_i \mid x_i)c_i}{\pi_{Q_{k-1}}(a_i \mid x_i)}\right)\right] \leq \sum_{k=0}^{K} \mathrm{KL}(Q_k, Q_{k-1}) + \log\left(\frac{1}{\delta}\right).$$

Now remark that for any $k \geq 1, i \in I_k, \theta_k$, because $d_{\theta_k}(a,x) = \mathbb{1}\{f_{\theta_k}(x) = a\} \in \{0,1\}$, we have for any (a,x):

$$\log\left(1 - \lambda \frac{d_{\theta_k}(a \mid x) c_i}{\pi_{Q_{k-1}}(a \mid x)}\right) = d_{\theta_k}(a,x) \log\left(1 - \lambda \frac{c_i}{\pi_{Q_{k-1}}(a \mid x)}\right).$$

We then have:

$$\begin{aligned}\sum_{k=0}^{K} \mathbb{E}_{\theta_k \sim Q_k}\left[\sum_{i \in I_k} -\log\left(1 - \lambda \frac{d_{\theta_i}(a_i \mid x_i) c_i}{\pi_i^0(a_i \mid x_i)}\right)\right] &= -\sum_{k=0}^{K} \sum_{i \in I_k} \pi_{Q_k}(a_i \mid x_i) \log\left(1 - \lambda \frac{c_i}{\pi_{Q_{k-1}}(a_i \mid x_i)}\right) \\ &= m\lambda \sum_{k=0}^{K} \hat{R}_k^{\lambda - LS}(\pi_{Q_k}).\end{aligned}$$

We also use twice the convexity of $x \to -\log(1 - \lambda x)$ to obtain:

$$\begin{aligned}\sum_{k=0}^{K} \mathbb{E}_{\theta_k \sim Q_k}\left[\sum_{i \in I_k} -\log\left(1 - \lambda R_k(d_{\theta_k})\right)\right] &= m \sum_{k=0}^{K} \mathbb{E}_{\theta_k \sim Q_k}\left[-\log\left(1 - \lambda R_k(d_{\theta_k})\right)\right] \\ &\geq m \sum_{k=0}^{K} -\log\left(1 - \lambda R_k(\pi_{Q_k})\right), \\ &\geq -(K+1)m \log\left(1 - \frac{\lambda}{K+1} \sum_{k=0}^{K} R_k(\pi_{Q_k})\right).\end{aligned}$$

Finally, mixing all those calculations together and dividing by $\lambda K m$ gives: with probability at least $1 - \delta$, for any K:

$$-\frac{1}{\lambda} \log\left(1 - \frac{\lambda}{K+1} \sum_{k=0}^{K} R_k(\pi_{Q_k})\right) \leq \frac{1}{K+1} \sum_{k=0}^{K} \hat{R}_k^{\lambda - LS}(\pi_{Q_k}, \pi_{Q_{k-1}}) + \frac{\sum_{k=0}^{K} \mathrm{KL}(Q_k, Q_{k-1}) + \log\left(\frac{1}{\delta}\right)}{\lambda(K+1)m}$$

Applying the function ψ_λ Finally, mixing all those calculations together, dividing by $\lambda(k+1)m$ and applying the increasing function $\psi_\lambda(x) := \frac{1}{\lambda}(1 - \exp(-\lambda x)$ gives:

$$\frac{1}{K+1} \sum_{k=0}^{K} R_k(\pi_{Q_k}) \leq \psi_\lambda\left(\frac{1}{K+1} \sum_{k=0}^{K} \hat{R}_k^{\lambda - LS}(\pi_{Q_k}, \pi_{Q_{k-1}}) + \frac{\sum_{k=0}^{K} \mathrm{KL}(Q_k, Q_{k-1}) + \log\left(\frac{1}{\delta}\right)}{\lambda(K+1)m}\right).$$

Finally, using that $\psi_\lambda(x) \leq x$ concludes the proof. □

D PROOFS OF SECTION 4

D.1 Proof of Theorem 4.1

Proof. For any $k \geq 0$, let:

$$\begin{aligned}Y_{i,j} &:= \log\left(1 - \lambda\left(R\left(d_\theta\right) - R\left(\pi_j\right)\right)\right) - \log\left(1 - \lambda\left(\frac{d_\theta(a_i|x_i)}{\pi_j(a_i|x_i)} - 1\right) c_i\right) \\ f_k(S,\theta) &:= -\sum_{j=1}^{k} \sum_{i \in I_j} Y_{i,j}.\end{aligned}$$

Now for a given posterior Q and a prior P, we apply the change of measure inequality [Csiszár, 1975, Donsker and Varadhan, 1976] for any k:

$$\mathbb{E}_{\theta \sim Q}\left[f_k(S,\theta)\right] \leq \mathrm{KL}(Q,P) + \log\left(\mathbb{E}_{\theta \sim P} \exp(f_k(S,\theta))\right). \tag{6}$$

We need to control the exponential random variable. To do so we prove the following lemma:

Lemma D.1. *The sequence* $(M_k := \mathbb{E}_{\theta\sim P}\exp(f_k(S,\theta))_{k\geq 0}$ *is a non-negative martingale.*

Proof. We fix $k \geq 0$. We show that $\mathbb{E}_{k-1}[M_k] = M_{k-1}$.

$$\begin{aligned}
\mathbb{E}_{k-1}[M_k] &= \mathbb{E}_{k-1}\left[\mathbb{E}_{\theta\sim P}\exp(f_k(S,\theta)\right] \\
&= \mathbb{E}_{\theta\sim P}\left[\mathbb{E}_{k-1}\left[\exp\left(f_{k-1}(S,\theta)\right)\exp\left(\sum_{i\in I_k}\log\left(\frac{1-\lambda\left(\frac{d_\theta(a_i|x_i)}{\pi_k(a_i|x_i)}-1\right)c_i}{1-\lambda(R(d_\theta)-R(\pi_k))}\right)\right)\right]\right] \\
&= \mathbb{E}_{\theta\sim P}\left[\exp\left(f_{k-1}(S,\theta)\right)\mathbb{E}_{k-1}\left[\exp\left(\sum_{i\in I_k}\log\left(\frac{1-\lambda\left(\frac{d_\theta(a_i|x_i)}{\pi_k(a_i|x_i)}-1\right)c_i}{1-\lambda(R(d_\theta)-R(\pi_k))}\right)\right)\right]\right].
\end{aligned}$$

The second line comes from the application of Fubini as P is data-independent. Now we have for any $\theta \in \Theta$:

$$\begin{aligned}
&\mathbb{E}_{k-1}\left[\exp\left(\sum_{i\in I_k}\log\left(\frac{1-\lambda\left(\frac{d_\theta(a_i|x_i)}{\pi_k(a_i|x_i)}-1\right)c_i}{1-\lambda(R(d_\theta)-R(\pi_k))}\right)\right)\right] \\
&\qquad = \mathbb{E}_{k-1}\left[\Pi_{i\in I_k}\frac{1-\lambda\left(\frac{d_\theta(a_i|x_i)}{\pi_k(a_i|x_i)}-1\right)c_i}{1-\lambda(R(d_\theta)-R(\pi_k))}\right] \\
&\qquad = \Pi_{i\in I_k}\frac{1-\lambda\mathbb{E}_{k-1}\left[\left(\frac{d_\theta(a_i|x_i)}{\pi_k(a_i|x_i)}-1\right)c_i\right]}{1-\lambda(R(d_\theta)-R(\pi_k))},
\end{aligned}$$

these computations holding as we exploited the assumption that, conditionally to $\mathcal{F}_{k-1}$, $S_k = ((x_i, a_i, c_i))_{i\in I_k}$ is constituted of mutually independent random variables (thus an expectation of product equals a product of expectations). Now, recall that π_k is $\mathcal{F}_{k-1}$-measurable and that for all $i \in I_k, a_i \sim \pi_k(\cdot \mid x_i)$. We then remark that, for any θ, $i \in I_k$:

$$\begin{aligned}
\mathbb{E}_{k-1}\left[\frac{d_\theta(a_i \mid x_i)c_i}{\pi_k(a_i \mid x_i)}\right] &= \mathbb{E}_{x_i'\sim\nu}\mathbb{E}_{a_i'\sim\pi_k(\cdot|x_i')}\mathbb{E}_{c_i'\sim p(\cdot|x_i',a_i')}\left[\frac{d_\theta(a_i' \mid x_i')c_i'}{\pi_k(a_i' \mid x_i')}\right] \\
&= \mathbb{E}_{x_i'\sim\nu}\mathbb{E}_{a_i'\sim\pi_k(\cdot|x_i')}\left[\frac{d_\theta(a_i' \mid x_i')}{\pi_k(a_i' \mid x_i')}\mathbb{E}_{c_i'\sim p(\cdot|x_i',a_i')}\left(c_i'\right)\right] \\
&= \mathbb{E}_{x_i'\sim\nu}\mathbb{E}_{a_i'\sim\pi_k(\cdot|x_i')}\left[c\left(x_i', d_\theta(x_i')\right)\right] \\
&= R(d_\theta)
\end{aligned}$$

Similarly, notice that $\mathbb{E}_{k-1}[c_i] = R(\pi_k)$.

Thus $\mathbb{E}_{k-1}[M_k] = M_{k-1}$, this concludes the lemma's proof. □

Now; given a martingale is also a supermartingale, we can apply Ville's inequality [Doob, 1939], which implies that with probability at least $1-\delta$, for any $k \geq 0$:

$$\mathbb{E}_{\theta\sim P}\exp(f_k(S,\theta)) \leq \frac{1}{\delta}.$$

Plugging this into Eq. (6) yields, with probability at least $1-\delta$, for any $k \geq 0$, any Q:

$$\sum_{j=1}^{k}\mathbb{E}_{\theta\sim Q}\left[\sum_{i\in I_j} -\log\left(1-\lambda\left(R(d_\theta)-R(\pi_j)\right)\right)+\log\left(1-\lambda\left(\frac{d_\theta(a_i \mid x_i)}{\pi_j(a_i \mid x_i)}-1\right)c_i\right)\right] \leq \mathrm{KL}(Q,P)+\log\left(\frac{1}{\delta}\right).$$

Now remark that for any $j \geq 1, i \in I_j, \theta$ because $d_\theta(a,x) = \mathbb{1}\{f_\theta(x) = a\} \in \{0,1\}$, we have for any (a,x):

$$\log\left(1 - \lambda\left(\frac{d_\theta(a_i \mid x_i)}{\pi_j(a_i \mid x_i)} - 1\right)c_i\right) = d_\theta(a_i|x_i)\log\left(1 - \lambda\frac{c_i}{\pi_j(a_i \mid x_i)(1+\lambda c_i)}\right) - \log(1+\lambda c_i).$$

We then have:

$$\sum_{j=1}^{k}\mathbb{E}_{\theta\sim Q}\left[\sum_{i\in I_j} -\log\left(1 - \lambda\left(\frac{d_\theta(a_i \mid x_i)}{\pi_j(a_i \mid x_i)} - 1\right)c_i\right)\right] = N_k\lambda\left(\hat{R}_{0:k}^{\lambda-\mathrm{adj}}(\pi_Q) + \hat{C}_{0,k}(\lambda)\right).$$

We also use the convexity of $x \to -\log(1-\lambda x)$ to obtain:

$$\begin{aligned}
\sum_{j=1}^{k}\mathbb{E}_{\theta\sim Q}\left[\sum_{i\in I_j} -\log\left(1 - \lambda\left(R(d_\theta) - R(\pi_j)\right)\right)\right] &= -\sum_{j=1}^{k}\mathbb{E}_{\theta\sim Q}\left[n_j\log\left(1 - \lambda\left(R(d_\theta) - R(\pi_j)\right)\right)\right] \\
&= -N_k\sum_{j=1}^{k}\mathbb{E}_{\theta\sim Q}\left[\frac{n_j}{N_k}\log\left(1 - \lambda\left(R(d_\theta) - R(\pi_j)\right)\right)\right] \\
&\geq -N_k\log\left(1 - \lambda\left(R(\pi_Q) - \sum_{j=1}^{k}\frac{n_j}{N_k}R(\pi_j)\right)\right).
\end{aligned}$$

Finally, mixing all those calculations together, dividing by λN_k and applying the increasing function $\psi_\lambda(x) := \frac{1}{\lambda}(1-\exp(-\lambda x)$ gives:

$$R(\pi_Q) \leq \psi_\lambda\left(\hat{R}_{0,k}^{\lambda}(\pi_Q) + \hat{C}_{0,k}(\lambda) + \frac{\mathrm{KL}(Q,P) + \log\left(\frac{1}{\delta}\right)}{\lambda N_k}\right).$$

Finally, using that $\psi_\lambda(x) \leq x$ concludes the proof. □

D.2 Proof of Proposition 4.2

We need a preliminary proposition to prove Proposition 4.2.

After we proved our upper bound, we need another theorem to analyze the algorithm.

Proposition D.2. *For any sequential algorithm* `Alg`*, data independent prior P, any $\boldsymbol{\lambda} \in (\mathbf{0},\mathbf{1})$, $\delta \in (0,1]$, with probability at least $1-\delta$ over S, we have, for all $k \geq 0$, denoting by π_{k+1} the output of* `Alg` *at time k, for all $Q \in \mathcal{P}(\Theta)$:*

$$\hat{R}_{0:k}^{\lambda-\mathrm{adj}}(\pi_Q) + \hat{C}_{0,k}(\lambda) \leq R(\pi_Q) - \sum_{j=1}^{k}\frac{n_j}{N_k}R(\pi_j) + \frac{\lambda}{1-\lambda}\sum_{j=1}^{k}\frac{n_j}{N_k}L(\pi_Q,\pi_j) + \frac{\mathrm{KL}(Q,P) + \log\left(\frac{1}{\delta}\right)}{\lambda N_k}$$

with $L(\pi_Q,\pi_j) = \mathbb{E}_{x\sim\nu}\left[\mathbb{E}_{\pi_j}\left[c^2\right] + \mathbb{E}_{\pi_Q}\left[c^2\left(\frac{1}{\pi_j(a|x)} - 2\right)\right]\right]$.

Proof. For any $k \geq 0$, and $\lambda \in (0,1)$, let:

$$\begin{aligned}
Y_{i,j} &:= \log\left(1 - \lambda\left(\frac{d_\theta(a_i|x_i)}{\pi_j(a_i|x_i)} - 1\right)c_i\right) + \lambda\left(R(d_\theta) - R(\pi_j) + \lambda M_\lambda(d_\theta,\pi_j)\right) \\
f_k(S,\theta) &:= -\sum_{j=1}^{k}\sum_{i\in I_j}Y_{i,j}.
\end{aligned}$$

with $M_\lambda(d_\theta, \pi_j) = \mathbb{E}_{(x,a,c)\sim P(\pi_j)}\left[\frac{\left(\frac{d_\theta(a|x)}{\pi_j(a|x)}-1\right)^2 c^2}{1-\lambda\left(\frac{d_\theta(a|x)}{\pi_j(a|x)}-1\right)c}\right]$. Now for a given posterior Q and a prior P, we apply the change of measure inequality Csiszár [1975], Donsker and Varadhan [1976] for any k:

$$\mathbb{E}_{\theta\sim Q}\left[f_k(S,\theta)\right] \leq \mathrm{KL}(Q,P) + \log\left(\mathbb{E}_{\theta\sim P}\exp(f_k(S,\theta))\right). \tag{7}$$

We also need to control the exponential random variable. We prove the following lemma:

Lemma D.3. *The sequence* $(M_k := \mathbb{E}_{\theta\sim P}\exp(f_k(S,\theta)))_{k\geq 0}$ *is a non-negative supermartingale.*

Proof. We fix $k \geq 0$. We show that $\mathbb{E}_{k-1}[M_k] \leq M_{k-1}$.

Let $L_k = \mathbb{E}_{k-1}\left[\exp\left(\sum_{i\in I_k}\log\left(\frac{1}{1-\lambda\left(\frac{d_\theta(a_i|x_i)}{\pi_k(a_i|x_i)}-1\right)c_i}\right) - \lambda\left(R\left(d_\theta\right) - R\left(\pi_k\right) + \lambda M_\lambda(d_\theta,\pi_k)\right)\right)\right]$. We have:

$$\begin{aligned}\mathbb{E}_{k-1}[M_k] &= \mathbb{E}_{k-1}\left[\mathbb{E}_{\theta\sim P}\exp(f_k(S,\theta)\right] \\ &= \mathbb{E}_{\theta\sim P}\left[\exp\left(f_{k-1}(S,\theta)\right)L_k\right],\end{aligned}$$

where the last line comes from Fubini's theorem as P is data-independent. Let us prove that L_k is smaller than 1. We have:

$$L_k = \prod_{i\in I_k}\exp\left(-\lambda\left(R\left(d_\theta\right) - R\left(\pi_k\right) + \lambda M_\lambda\left(d_\theta,\pi_k\right)\right)\right)\mathbb{E}_{k-1}\left[\frac{1}{1-\lambda(\frac{d_\theta(a|x)}{\pi_k(a|x)}-1)c}\right]$$

Using $\log(x) \leq x-1$ yields

$$\begin{aligned}L_k &\leq \prod_{i\in I_k}\exp\left(-\lambda\left(R\left(d_\theta\right) - R\left(\pi_k\right) + \lambda M_\lambda\left(d_\theta,\pi_k\right)\right) + \mathbb{E}_{k-1}\left[\frac{1}{1-\lambda(\frac{d_\theta(a|x)}{\pi_k(a|x)}-1)c}\right] - 1\right) \\ &= \prod_{i\in I_k}\exp\left(-\lambda\left(R\left(d_\theta\right) - R\left(\pi_k\right) + \lambda M_\lambda\left(d_\theta,\pi_k\right)\right) + \mathbb{E}_{k-1}\left[\frac{\lambda(\frac{d_\theta(a|x)}{\pi_k(a|x)}-1)c}{1-\lambda(\frac{d_\theta(a|x)}{\pi_k(a|x)}-1)c}\right]\right) \\ &= \prod_{i\in I_k}\exp\left(-\lambda^2 M_\lambda\left(d_\theta,\pi_k\right) + \mathbb{E}_{k-1}\left[\frac{\lambda(\frac{d_\theta(a|x)}{\pi_k(a|x)}-1)c}{1-\lambda(\frac{d_\theta(a|x)}{\pi_k(a|x)}-1)c} - \lambda(\frac{d_\theta(a\mid x)}{\pi_k(a\mid x)}-1)c\right]\right) \\ &= \prod_{i\in I_k}\exp\left(-\lambda^2 M_\lambda\left(d_\theta,\pi_k\right) + \mathbb{E}_{k-1}\left[\frac{\lambda^2(\frac{d_\theta(a|x)}{\pi_k(a|x)}-1)^2c^2}{1-\lambda(\frac{d_\theta(a|x)}{\pi_k(a|x)}-1)c}\right]\right) \\ &= \prod_{i\in I_k}\exp\left(-\lambda^2 M_\lambda\left(d_\theta,\pi_k\right) + \lambda^2 M_\lambda\left(d_\theta,\pi_k\right)\right) \\ &= 1.\end{aligned}$$

Thus $\mathbb{E}_{k-1}[M_k] \leq M_{k-1}$, this concludes the lemma's proof. □

We can apply Ville's inequality Doob [1939] which implies that with probability at least $1-\delta$, for any $k \geq 0$:

$$\mathbb{E}_{\theta\sim P}\exp(f_k(S,\theta)) \leq \frac{1}{\delta}.$$

Plugging this into Eq. (7) yields, with probability at least $1-\delta$, for any $k \geq 0$, any Q:

$$-\sum_{j=0}^{k}\sum_{i\in I_j}\mathbb{E}_{\theta\sim Q}\left[Y_{i,j}\right] \leq \mathrm{KL}(Q,P) + \log\left(\frac{1}{\delta}\right).$$

with $Y_{i,j} = \log\left(1 - \lambda\left(\frac{d_\theta(a_i|x_i)}{\pi_j(a_i|x_i)} - 1\right)c_i\right) + \lambda\left(R\left(d_\theta\right) - R\left(\pi_j\right) + \lambda M_\lambda(d_\theta, \pi_j)\right)$.

Now remark that for any $j \geq 1, i \in I_j, \theta$ because $d_\theta(a,x) = \mathbb{1}\{f_\theta(x) = a\} \in \{0,1\}$, we have for any (a,x):

$$\log\left(1 - \lambda\left(\frac{d_\theta(a_i \mid x_i)}{\pi_j(a_i \mid x_i)} - 1\right)c_i\right) = d_\theta(a_i|x_i)\log\left(1 - \lambda\frac{c_i}{\pi_j(a_i \mid x_i)(1+\lambda c_i)}\right) - \log(1+\lambda c_i).$$

We thus obtain by re-organising the terms and dividing by λN_k:

$$\hat{R}_{0:k}^{\lambda-\mathrm{adj}}(\pi_Q) + \hat{C}_{0,k}(\lambda) \leq R\left(\pi_Q\right) - \sum_{j=1}^{k}\frac{n_j}{N_k}R\left(\pi_j\right) + \lambda\sum_{j=1}^{k}\frac{n_j}{N_k}\mathbb{E}_{\theta\sim Q}\left[M_\lambda\left(d_\theta, \pi_j\right)\right] + \frac{\mathrm{KL}(Q,P) + \log\left(\frac{1}{\delta}\right)}{\lambda N_k}.$$

We proceed then by bounding $\mathbb{E}_{\theta\sim Q}\left[M_\lambda\left(d_\theta, \pi_j\right)\right]$:

$$\begin{aligned}
\mathbb{E}_{\theta\sim Q}\left[M_\lambda\left(d_\theta, \pi_j\right)\right] &= \mathbb{E}_{\theta\sim Q}\left[M_\lambda\left(d_\theta, \pi_j\right)\right] \\
&= \mathbb{E}_{\theta\sim Q}\left[\mathbb{E}_{\pi_j}\left[\frac{\left(\frac{d_\theta(a|x)}{\pi_j(a|x)} - 1\right)^2 c^2}{1 - \lambda\left(\frac{d_\theta(a|x)}{\pi_j(a|x)} - 1\right)c}\right]\right] \\
&\leq \frac{1}{1-\lambda}\mathbb{E}_{\theta\sim Q}\left[\mathbb{E}_{\pi_j}\left[\left(\frac{d_\theta(a|x)}{\pi_j(a|x)} - 1\right)^2 c^2\right]\right] \\
&= \frac{1}{1-\lambda}\mathbb{E}_{\pi_j}\left[\mathbb{E}_{\theta\sim Q}\left[\left(\frac{d_\theta(a|x)}{\pi_j(a|x)} - 1\right)^2 c^2\right]\right] \\
&= \frac{1}{1-\lambda}\mathbb{E}_{\pi_j}\left[\mathbb{E}_{\theta\sim Q}\left[d_\theta(a|x)\left(\frac{1}{\pi_j(a|x)} - 1\right)^2 c^2 + (1 - d_\theta(a|x))c^2\right]\right] \\
&= \frac{1}{1-\lambda}\mathbb{E}_{\pi_j}\left[\pi_Q(a|x)\left(\frac{1}{\pi_j(a|x)} - 1\right)^2 c^2 + (1 - \pi_Q(a|x))c^2\right] \\
&= \frac{1}{1-\lambda}\mathbb{E}_{\pi_j}\left[\pi_Q(a|x)\left(\frac{1}{\pi_j(a|x)} - 1\right)^2 c^2 + (1 - \pi_Q(a|x))c^2\right] \\
&= \mathbb{E}_{\pi_j}\left[c^2\right] + \mathbb{E}_{\pi_Q}\left[c^2\left(\frac{1}{\pi_j(a|x)} - 2\right)\right] = L(\pi_Q, \pi_j).
\end{aligned}$$

We used several times above the property $d_\theta \in \{0,1\}$ to transform the analytical expression of the random variable with $\mathbb{E}_{\pi_j}$. This gives our result and ends the proof. □

Now we are able to prove our main result. Let us recall it below for conciseness.

Theorem D.4. *For any data independent prior P, any $\boldsymbol{\lambda} \in (\mathbf{0}, \mathbf{1})$, $\delta \in (0,1]$, with probability at least $1-\delta$, we have, denoting by π_j the output of Algorithm 2 at time j, for all $k \geq 0$:*

$$0 \leq R\left(\pi_{k+1}\right) - R\left(\pi_\star\right) \leq \frac{\lambda}{1-\lambda}\sum_{j=1}^{k}\frac{n_j}{N_k}L\left(\pi_\star, \pi_j\right) + 2\frac{\mathrm{KL}(Q^\star, P) + \log\left(\frac{2}{\delta}\right)}{\lambda N_k}$$

with π_{k+1} defined as the output of the sequential algorithm Algorithm 2 at time k.

Proof. We know that by Thm. 4.1, for any data independent prior P, any $\boldsymbol{\lambda} \in (\mathbf{0}, \mathbf{1})$, $\delta \in (0,1]$, with probability at least $1 - \delta/2$ over S, we have for all $k \geq 0$:

$$\begin{aligned}
R\left(\pi_{k+1}\right) - \sum_{j=1}^{k}\frac{n_j}{N_k}R\left(\pi_j\right) &\leq \psi_\lambda\left(\hat{R}_{0:k}^{\lambda-\mathrm{adj}}(\pi_{k+1}) + \hat{C}_{0,k}(\lambda) + \frac{\mathrm{KL}(Q_{k+1}, P) + \log\left(\frac{2}{\delta}\right)}{\lambda N_k}\right) \\
&\leq \hat{R}_{0:k}^{\lambda-\mathrm{adj}}(\pi_{k+1}) + \hat{C}_{0,k}(\lambda) + \frac{\mathrm{KL}(Q_{k+1}, P) + \log\left(\frac{1}{\delta}\right)}{\lambda N_k},
\end{aligned}$$

as $\psi_\lambda(x) \leq x$. Then, by definition, $\hat{Q}_{k+1} = \text{argmin}_Q \left\{ \hat{R}_{0:k}^{\lambda-\text{adj}}(\pi_\text{Q}) + \frac{\text{KL}(Q,P)}{\lambda N_k} \right\}$, thus:

$$\forall k \geq 0, \quad R\left(\pi_{k+1}\right) - \sum_{j=1}^{k} \frac{n_j}{N_k} R\left(\pi_j\right) \leq \hat{R}_{0:k}^{\lambda-\text{adj}}(\pi_\star) + \hat{C}_{0,k}(\lambda) + \frac{\text{KL}(Q^\star, P) + \log\left(\frac{2}{\delta}\right)}{\lambda N_k}.$$

Now, we exploit Proposition D.2, with probability at least $1-\delta/2$ over S, we have $\forall k \geq 0$:

$$\hat{R}_{0:k}^{\lambda-\text{adj}}(\pi_\star) + \hat{C}_{0,k}(\lambda) \leq R\left(\pi_\star\right) - \sum_{j=1}^{k} \frac{n_j}{N_k} R\left(\pi_j\right) + \frac{\lambda}{1-\lambda} \sum_{j=1}^{k} \frac{n_j}{N_k} L\left(\pi_\star, \pi_j\right) + \frac{\text{KL}(Q^\star, P) + \log\left(\frac{2}{\delta}\right)}{\lambda N_k}.$$

Now, taking an union bound and combining the two previous results gives, with probability at least $1-\delta$ over S, $\forall k \geq 0$:

$$\begin{aligned} R\left(\pi_{k+1}\right) - \sum_{j=1}^{k} \frac{n_j}{N_k} R\left(\pi_j\right) \leq R\left(\pi_\star\right) - \sum_{j=1}^{k} \frac{n_j}{N_k} R\left(\pi_j\right) + \frac{\lambda}{1-\lambda} \sum_{j=1}^{k} \frac{n_j}{N_k} L\left(\pi_\star, \pi_j\right) \\ + 2\frac{\text{KL}(Q^\star, P) + \log\left(\frac{2}{\delta}\right)}{\lambda N_k}. \end{aligned}$$

Re-organising the terms conclude the proof. □

D.3 Proof of Lemma 4.3

Proof. We are looking for $\gamma_k > 0$ such that:

$$\frac{L(\pi_\star, \pi_k)}{R(\pi_k) - R(\pi_\star)} \leq \gamma_k.$$

Let us start rewriting the suboptimality $R(\pi_k) - R(\pi_\star)$. We have:

$$\begin{aligned} R(\pi_k) - R(\pi_\star) &= \mathbb{E}_x \left[\mathbb{E}_{\pi_k}\left[c(a,x)\right] - c(a_\star(x), x)\right] \\ &= \mathbb{E}_x \left[\mathbb{E}_{\pi_k}\left[c(a,x) - c(a_\star(x), x)\right]\right] \\ &= \mathbb{E}_x \left[\sum_{a \neq a_\star(x)} \pi_k(a|x) \Delta_{a,x}\right]. \end{aligned}$$

On the other hand, we have by definition of $L(\pi_\star, \pi_k)$:

$$\begin{aligned} L(\pi_\star, \pi_k) &= \mathbb{E}_x \left[\mathbb{E}_{\pi_k}\left[\mathbb{E}_{p(\cdot|x,a)}\left[c^2\right]\right] + \left(\frac{1}{\pi_k(a^\star(x)|x)} - 2\right) \mathbb{E}_{p(\cdot|x,a_\star(x))}\left[c^2\right]\right] \\ &= \mathbb{E}_x \left[\mathbb{E}_{\pi_k}\left[\mathbb{E}_{p(\cdot|x,a)}\left[c^2\right] - \mathbb{E}_{p(\cdot|x,a_\star(x))}\left[c^2\right]\right] + \left(\frac{1}{\pi_k(a^\star(x)|x)} - 1\right) \mathbb{E}_{p(\cdot|x,a_\star(x))}\left[c^2\right]\right] \\ &= \mathbb{E}_x \left[\sum_{a \neq a_\star(x)} \pi_k(a|x) \left[\mathbb{E}_{p(\cdot|x,a)}\left[c^2\right] - \mathbb{E}_{p(\cdot|x,a_\star(x))}\left[c^2\right]\right] + \left(\frac{1}{\pi_k(a^\star(x)|x)} - 1\right) \mathbb{E}_{p(\cdot|x,a_\star(x))}\left[c^2\right]\right] \\ &= \mathbb{E}_x \left[\sum_{a \neq a_\star(x)} \pi_k(a|x) \left[\sigma^2(a,x) - \sigma^2(a_\star(x), x) + c^2(a,x) - c^2(a_\star(x), x)\right]\right] \\ &+ \mathbb{E}_x \left[\left(\frac{1}{\pi_k(a^\star(x)|x)} - 1\right) \mathbb{E}_{p(\cdot|x,a_\star(x))}\left[c^2\right]\right]. \end{aligned}$$

As the costs are in $[0,1]$, we can bound the variance term by:

$$\mathbb{E}_x\left[\sum_{a\neq a_\star(x)}\pi_k(a|x)\left[\sigma^2(a,x)-\sigma^2(a_\star(x),x)\right]\right]\leq\frac{1}{4}\mathbb{E}_x\left[\sum_{a\neq a_\star(x)}\pi_k(a|x)\right]=\frac{1}{4}\mathbb{E}_x\left[1-\pi_k(a_\star(x)|x)\right].$$

We then use the definition of $\Delta_{x,a}$ to develop $c(a,x)^2$ and obtain:

$$\begin{aligned}c(a,x)^2&=\left(\Delta_{a,x}+c(a_\star(x),x)\right)^2\\&=\Delta_{a,x}^2+2c(a_\star(x),x)\Delta_{x,a}+c(a_\star(x),x)^2\leq\Delta_{a,x}^2+c(a_\star(x),x)^2.\end{aligned}$$

Plugging these two results into our previous equation, we get:

$$\begin{aligned}L(\pi_\star,\pi_k)&\leq\frac{1}{4}\mathbb{E}_x\left[1-\pi_k(a_\star(x)|x)\right]+\mathbb{E}_x\left[\sum_{a\neq a_\star(x)}\pi_k(a|x)\Delta_{x,a}^2\right]+\mathbb{E}_x\left[\left(\frac{1}{\pi_k(a^\star(x)|x)}-1\right)\mathbb{E}_{p(\cdot|x,a_\star(x))}\left[c^2\right]\right]\\&\leq\frac{1}{4}\mathbb{E}_x\left[1-\pi_k(a_\star(x)|x)\right]+\mathbb{E}_x\left[\sum_{a\neq a_\star(x)}\pi_k(a|x)\Delta_{x,a}^2\right]+\mathbb{E}_x\left[\left(\frac{1}{\pi_k(a^\star(x)|x)}-1\right)\right]\\&\leq\frac{1}{4}\mathbb{E}_x\left[1-\pi_k(a_\star(x)|x)\right]+\mathbb{E}_x\left[\sum_{a\neq a_\star(x)}\pi_k(a|x)\Delta_{x,a}^2\right]+\frac{1}{C_k^\star}\mathbb{E}_x\left[1-\pi_k(a^\star(x)|x)\right].\end{aligned}$$

We then proceed at upper bound the quantity:

$$\begin{aligned}\frac{L(\pi_\star,\pi_k)}{R(\pi_k)-R(\pi_\star)}&\leq\frac{\frac{1}{4}\mathbb{E}_x\left[1-\pi_k(a_\star(x)|x)\right]+\mathbb{E}_x\left[\sum_{a\neq a_\star(x)}\pi_k(a|x)\Delta_{x,a}^2\right]+\frac{1}{C_k^\star}\mathbb{E}_x\left[1-\pi_k(a^\star(x)|x)\right]}{\mathbb{E}_x\left[\sum_{a\neq a_\star(x)}\pi_k(a|x)\Delta_{a,x}\right]}\\&\leq1+\left(\frac{1}{4}+\frac{1}{C_k^\star}\right)\frac{\mathbb{E}_x\left[1-\pi_k(a_\star(x)|x)\right]}{\mathbb{E}_x\left[\sum_{a\neq a_\star(x)}\pi_k(a|x)\Delta_{a,x}\right]},\end{aligned}$$

because $\Delta_{x,a}^2\leq\Delta_{x,a}$ as $\Delta_{x,a}\leq1$.

Now, we proceed to define:

$$\bar{\Delta}_k=\frac{\mathbb{E}_x\left[\sum_{a\neq a_\star(x)}\pi_k(a|x)\Delta_{a,x}\right]}{\mathbb{E}_x\left[1-\pi_k(a_\star(x)|x)\right]}.$$

$\bar{\Delta}_k$ represents the expected sub-optimality incurred when policy π_k selects a non-optimal action. A large $\bar{\Delta}_k$ means that the environment penalizes us more for not choosing an optimal action, while $\bar{\Delta}_k\approx0$ means that the environment has multiple near-optimal actions. We can characterize γ_k by this quantity, or loosen our bound and have a more interpretable condition. We choose the second approach. Let us suppose that there exists a $\Delta_u>0$ for a $u\in[0,1)$ such that:

$$P_x\left(\min_{a\neq a^\star(x)}\Delta_{x,a}\geq\Delta_u\right)=1-u.$$

Let us focus on the numerator of $\bar{\Delta}_k$. We have:

$$P\left(\Delta_{x,a}\geq\Delta_u|a\neq a^\star(x)\right)\geq P_x\left(\min_{a\neq a^\star(x)}\Delta_{x,a}\geq\Delta_u\right)=1-u.$$

Indeed, the event $E_1=\{\Delta_{x,a}\geq\Delta_u|a\neq a^\star(x)\}$ represents choosing a random action from the suboptimal set, and finding that $\Delta_{x,a}\geq\Delta_u$, while the event $E_0=\{\min_{a\neq a^\star(x)}\Delta_{x,a}\geq\Delta_u\}$ means that all non-optimal actions have a sub-optimality bigger than Δ_u. We have $E_0\subset E_1$, and thus:

$$P\left(\Delta_{x,a}\geq\Delta_u|a\neq a^\star(x)\right)\geq1-u.$$

Additionally, we have by the definition of conditional expectations:

$$P\left(\Delta_{x,a} \geq \Delta_u | a \neq a^\star(x)\right) = \frac{\mathbb{E}_x\left[\sum_{a\neq a_\star(x)} \pi_k(a|x)\mathbb{1}\left[\Delta_{a,x} > \Delta_u\right]\right]}{\mathbb{E}_x\left[\sum_{a\neq a_\star(x)} \pi_k(a|x)\right]}$$
$$= \frac{\mathbb{E}_x\left[\sum_{a\neq a_\star(x)} \pi_k(a|x)\mathbb{1}\left[\Delta_{a,x} > \Delta_u\right]\right]}{\mathbb{E}_x\left[1-\pi_k(a_\star(x)|x)\right]} \geq 1-u.$$

Now, we write:

$$\mathbb{E}_x\left[\sum_{a\neq a_\star(x)} \pi_k(a|x)\Delta_{a,x}\right] \geq \mathbb{E}_x\left[\sum_{a\neq a_\star(x)} \pi_k(a|x)\Delta_{a,x}\mathbb{1}\left[\Delta_{a,x} > \Delta_u\right]\right]$$
$$\geq \Delta_u \mathbb{E}_x\left[\sum_{a\neq a_\star(x)} \pi_k(a|x)\mathbb{1}\left[\Delta_{a,x} > \Delta_u\right]\right]$$
$$\geq \Delta_u(1-u)\mathbb{E}_x\left[1-\pi_k(a_\star(x)|x)\right].$$

Finally, we have:

$$\bar{\Delta}_k = \frac{\mathbb{E}_x\left[\sum_{a\neq a_\star(x)} \pi_k(a|x)\Delta_{a,x}\right]}{\mathbb{E}_x\left[1-\pi_k(a_\star(x)|x)\right]} \geq \Delta_u(1-u).$$

and we finally obtain our result:

$$\frac{L(\pi_\star,\pi_k)}{R(\pi_k)-R(\pi_\star)} \leq 1 + \left(\frac{1}{4}+\frac{1}{C_k^\star}\right)\frac{1}{\Delta_u(1-u)} = \gamma_k$$

This ends the proof. □

D.4 Proof of Theorem 4.4

We first start with a key intermediary result, derived from Lemma 4.3 coupled with Proposition 4.2.

Corollary D.5. *Let π_{k+1} be the output of Algorithm 2 at time k with parameter $\lambda \in (0,1)$. For any data independent prior P, any $\boldsymbol{\lambda} \in \mathbf{(0,1)}, \delta \in (0,1]$, with probability at least $1-\delta$, we have for all $k \geq 0$:*

$$0 \leq D_{k+1} \leq \frac{\lambda}{1-\lambda}\sum_{j=0}^{k}\frac{n_j}{N_k}\gamma_j D_j + 2\frac{\mathrm{KL}(Q^\star,P)+\log\left(\frac{2}{\delta}\right)}{\lambda N_k},$$

where $D_k = R(\pi_k) - R(\pi^\star)$.

We now proceed to prove Theorem 4.4.

Proof. Assume that for all j, $n_j = m$, then $N_k = (k+1)m$ for all k. We re-use the notation $D_k = R(\pi_k) - R(\pi^\star)$ for all $k \geq 0$. We define

$$\lambda_m := \frac{1-\alpha}{8\gamma\sqrt{m}}$$
$$C_\alpha := 64\gamma\frac{\mathrm{KL}(\mathrm{Q}^\star,P)+\log(2/\delta)}{(1-\alpha)}.$$

Now let us prove the result by strong recursion.

Initialisation Take $k = 0$. We have, by Cor. D.5:

$$D_1 \leq \frac{\lambda}{1-\lambda}\gamma_1 D_0 + 2\frac{\mathrm{KL}(Q^\star, P) + \log\left(\frac{2}{\delta}\right)}{\lambda m},$$
$$\leq \frac{\lambda}{1-\lambda}\gamma D_0 + 2\frac{\mathrm{KL}(Q^\star, P) + \log\left(\frac{2}{\delta}\right)}{\lambda m}$$

Now we take $\lambda = \lambda_m$ and notice that $\lambda_m \leq 1/2$, thus $\frac{\lambda_m}{1-\lambda_m} \leq 2\lambda_m$:

$$\leq \frac{1}{\sqrt{m}}\left(\frac{(1-\alpha)}{4}D_0 + 16\gamma\frac{\mathrm{KL}(Q^\star, P) + \log\left(\frac{2}{\delta}\right)}{1-\alpha}\right),$$
$$\leq \frac{1}{\sqrt{m}}\left(\frac{(1-\alpha)}{4}D_0 + 16\gamma\frac{\mathrm{KL}(Q^\star, P) + \log\left(\frac{2}{\delta}\right)}{1-\alpha}\right),$$

Now as we have $\frac{1-\alpha}{4}D_0 \leq \frac{1}{4} \leq \frac{C_\alpha}{2}$ we write:

$$\leq \frac{1}{\sqrt{m}}\left(\frac{C_\alpha}{2} + \frac{C_\alpha}{4}\right),$$

This then ensures that $D_1 \leq \frac{C_\alpha}{\sqrt{m}}$ and prove the result for $k = 1$

General case. Assume the result true for all $1 \leq j \leq k$. By Cor. D.5, we have for any $\lambda \leq 1/2$:

$$D_{k+1} \leq \frac{\lambda}{1-\lambda}\frac{1}{k+1}\sum_{j=0}^{k}\gamma_j D_j + 2\frac{\mathrm{KL}(Q^\star, P) + \log\left(\frac{2}{\delta}\right)}{\lambda(k+1)m},$$
$$\leq 2\lambda\gamma\frac{1}{k+1}\sum_{j=0}^{k}D_j + 2\frac{\mathrm{KL}(Q^\star, P) + \log\left(\frac{2}{\delta}\right)}{\lambda k m},$$

Now, we use the recursion assumption to get

$$D_{k+1} \leq 2\lambda\gamma\left(\frac{D_0}{k+1} + \frac{1}{(k+1)\sqrt{m}}\sum_{j=1}^{k}\frac{C_\alpha}{k^\alpha}\right) + 2\frac{\mathrm{KL}(Q^\star, P) + \log\left(\frac{2}{\delta}\right)}{\lambda(k+1)m},$$
$$\leq 2\lambda\gamma\left(\frac{D_0}{k+1} + \frac{C_\alpha B(\alpha)}{(k+1)\sqrt{m}}k^{1-\alpha}\right) + 2\frac{\mathrm{KL}(Q^\star, P) + \log\left(\frac{2}{\delta}\right)}{\lambda(k+1)m}$$
$$\leq 2\lambda\gamma\left(\frac{D_0}{k+1} + \frac{C_\alpha B(\alpha)}{(k+1)^\alpha\sqrt{m}}\right) + 2\frac{\mathrm{KL}(Q^\star, P) + \log\left(\frac{2}{\delta}\right)}{\lambda(k+1)m},$$

Now as $\lambda_m \leq 1/2$, take $\lambda = \lambda_m$, we have:

$$D_{k+1} \leq \frac{1}{\sqrt{m}}\left(\frac{D_0(1-\alpha)}{4(k+1)} + \frac{C_\alpha}{4(k+1)^\alpha\sqrt{m}}\right) + 16\gamma\frac{\mathrm{KL}(Q^\star, P) + \log\left(\frac{2}{\delta}\right)}{(1-\alpha)(k+1)\sqrt{m}},$$
$$\leq \frac{1}{\sqrt{m}(k+1)^\alpha}\left(\frac{D_0(1-\alpha)}{4} + \frac{C_\alpha}{4\sqrt{m}} + 16\gamma\frac{\mathrm{KL}(Q^\star, P) + \log\left(\frac{2}{\delta}\right)}{(1-\alpha)}\right),$$

We use $C_\alpha \geq D_0$ and its expression to get:

$$D_{k+1} \leq \frac{1}{\sqrt{m}(k+1)^\alpha}\left(\frac{C_\alpha}{4} + \frac{C_\alpha}{4\sqrt{m}} + \frac{C_\alpha}{4}\right),$$

We then have: $D_{k+1} \leq \frac{C_\alpha}{\sqrt{m}(k+1)^\alpha}$, this concludes the proof. □

D.5 Proof of Corollary 4.5

Proof of Cor. 4.5. We re-use the notation $D_k = R(\pi_k) - R(\pi^\star)$ for all $k \geq 0$. We define, for any $\alpha \in [0,1)$,

$$C_\alpha := \max\left(D_0, (1-\lambda)\frac{\mathrm{KL}(\mathrm{Q}^\star, P) + \log(2/\delta)}{\lambda^2 \gamma^* \beta_2 B(\alpha)}\right) \leq \max\left(1, \frac{\mathrm{KL}(\mathrm{Q}^\star, P) + \log(2/\delta)}{\lambda^2 \gamma^* \beta_2 B(\alpha)}\right),$$

where $\lambda = \frac{1}{1+2^{2+\alpha}\gamma^\star \beta_1 \beta_2 B(\alpha)}$ Note that thanks to this definition, we have

$$C_\alpha \frac{\lambda}{1-\lambda}\gamma^* \beta_2 B(\alpha) \geq \frac{\mathrm{KL}(\mathrm{Q}^\star, P) + \log(2/\delta)}{\lambda}. \tag{8}$$

Now let us prove the result by strong recursion.

Initialisation. By definition of C_α, we have $D_0 \leq C_\alpha$.

General case. For $k \geq 0$, Assume the result true for all $j \leq k$. By Cor. D.5, we have:

$$\begin{aligned} D_{k+1} &\leq \frac{\lambda}{1-\lambda}\sum_{j=0}^{k}\frac{n_j}{N_k}\gamma_j D_j + 2\frac{\mathrm{KL}(Q^\star, P) + \log\left(\frac{2}{\delta}\right)}{\lambda N_k}, \\ &\leq \frac{\lambda}{1-\lambda}\gamma^\star \beta_1 \beta_2 \frac{1}{k+1}\sum_{j=0}^{k} D_j + 2\frac{\mathrm{KL}(Q^\star, P) + \log\left(\frac{2}{\delta}\right)}{\lambda(k+1)}\beta_1. \end{aligned}$$

In the last line, we used $\frac{k+1}{N_k} \leq \beta_1$, $\frac{n_j}{N_k} = \frac{n_j}{k+1}\frac{k+1}{N_k} \leq \frac{n_j}{k+1}\beta_1 \leq \beta_1 \frac{\beta_2}{k+1}$ and $\gamma_j \leq \gamma^\star$ for all $k \geq 0, j \leq k$. Now we use the recursion assumption.

$$D_{k+1} \leq C_\alpha \frac{\lambda}{1-\lambda}\gamma^\star \beta_1 \beta_2 \frac{1}{k+1}\sum_{j=0}^{k}\frac{1}{(j+1)^\alpha} + 2\frac{\mathrm{KL}(Q^\star, P) + \log\left(\frac{2}{\delta}\right)}{\lambda(k+1)}\beta_1.$$

Yet, $\sum_{j=0}^{k}\frac{1}{(j+1)^\alpha} \leq 1 + B(\alpha)(k+1)^{1-\alpha} \leq 2B(\alpha)(k+1)^{1-\alpha}$, thus

$$\begin{aligned} D_{k+1} &\leq 2C_\alpha \frac{\lambda}{1-\lambda}\gamma^\star \beta_1 \beta_2 B(\alpha)\frac{1}{(k+1)^\alpha} + 2\frac{\mathrm{KL}(Q^\star, P) + \log\left(\frac{2}{\delta}\right)}{\lambda(k+1)}\beta_1. \\ &\leq 2C_\alpha \frac{\lambda}{1-\lambda}\gamma^\star \beta_1 \beta_2 B(\alpha)\frac{1}{(k+1)^\alpha} + 2\frac{\mathrm{KL}(Q^\star, P) + \log\left(\frac{2}{\delta}\right)}{\lambda(k+1)^\alpha}\beta_1. \end{aligned}$$

Now, use that for all $k, (\frac{k+2}{k+1})^\alpha \leq 2^\alpha$:

$$D_{k+1} \leq 2^{1+\alpha}\frac{\beta_1}{(k+2)^\alpha}\left(C_\alpha \frac{\lambda}{1-\lambda}\gamma^\star \beta_2 B(\alpha) + \frac{\mathrm{KL}(Q^\star, P) + \log\left(\frac{2}{\delta}\right)}{\lambda}\right).$$

Now using Eq. (8) gives:

$$D_{k+1} \leq \frac{\lambda}{1-\lambda}2^{2+\alpha}\gamma^\star \beta_1 \beta_2 B(\alpha)\frac{C_\alpha}{(k+2)^\alpha}.$$

Now notice that λ is such that

$$\frac{\lambda}{1-\lambda}2^{2+\alpha}\gamma^\star \beta_1 \beta_2 B(\alpha) = 1.$$

This ensures that $D_{k+1} \leq \frac{C_\alpha}{(k+2)^\alpha}$ and concludes the recursion. □

E EXPERIMENTAL DESIGN AND ADDITIONAL RESULTS

E.1 Detailed Setup

General Setup. Our experimental framework follows the standard multiclass-to-bandit conversion widely used in prior work [Dudik et al., 2014, Swaminathan and Joachims, 2015]. Each multiclass dataset consists of feature-label pairs, which we transform into contextual bandit problems by treating features as contexts and labels as actions. The reward r for taking action a for context x is modeled as Bernoulli with probability $p_x = \epsilon + \mathbb{1}\left[a = a^\star(x)\right](1 - 2\epsilon)$, where $a^\star(x)$ is the true label of features x, and ϵ is a noise parameter, that we set to $\epsilon = 0.2$ in all our experiments. This setup ensures an average reward of $1 - \epsilon$ for the optimal action $a^\star(x)$ and ϵ for all others. In our experiments, we use parametrized Linear Gaussian Policies [Sakhi et al., 2023a, Aouali et al., 2023], these policies interact with the contextual bandit environment to construct a logged bandit feedback dataset in the form $\{x_i, a_i, c_i\}_{i\in[n]}$, where $c_i = -r_i$ is the associated cost.

In this procedure, we need three splits: D_l (of size n_l) to train the logging policy π_0, another split D_c (of size n_c) to generate the logging feedback with π_0, and finally a test split D_{test} (of size n_{test}) to compute the true risk $R(\pi)$ of any policy π. In our experiments, we split the training split D_{train} (of size N) of the four datasets considered into D_l ($n_l = 0.05N$) and D_c ($n_c = 0.95N$) and use their test split D_{test}. The detailed statistics of the different splits can be found in Table 2. Recall that K is the number of actions and p the number of features.

All datasets are used without processing aside from `CIFAR100` that we embed using a pretrained Resnet, obtaining a low dimensional vector of 100 for each image. All experiments do not require heavy computations, they were conducted on a single 16 CPU/ 1 GPU V100 machine.

Datasets	N	n_l	n_c	n_{test}	K	p
MNIST	60 000	3000	57 000	10 000	10	784
FashionMNIST	60 000	3000	57 000	10 000	10	784
EMNIST	112 800	5640	107 160	18 800	47	784
CIFAR100	60 000	3000	57 000	10 000	100	500

Table 2: Detailed statistics of the datasets used.

Linear Gaussian Policies We use the Linear Gaussian Policy of [Sakhi et al., 2023a]. To obtain these policies, we restrict f_θ to:

$$\forall x \in \mathcal{X}, \quad f_\theta(x) = \arg\max_{a' \in \mathcal{A}} \left\{x^t \theta_{a'}\right\} \tag{9}$$

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

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

with Φ the cumulative distribution function of the standard normal.

The starting behavior policy π_0. π_0 is trained on D_l (supervised manner) with the following parameters: We use L_2 regularization of 10^{-4}. This is used to prevent the logging policy π_0 from being close to deterministic, allowing efficient learning with importance sampling. We use Adam [Kingma and Ba, 2014] with a learning rate of 10^{-1} for 10 epochs. Once it is trained, we use an inverse temperature parameter α on its score to interpolate between a uniform policy $\alpha = 0$ and a trained policy $\alpha = 1$.

Optimizing our learning objectives. In each optimization subroutine, we use Adam with a learning rate of 10^{-3} for 10 epochs. The gradient of **LIG** policies is a one dimensional integral, and is approximated using $S = 32$

α	k	SeqLS (Alg 1)	SeqAdjLS (Alg 2)
$\alpha = 0.$	0	-0.010	-0.010
	1	-0.011	-0.011
	5	-0.011	-0.011
	10	-0.020	-0.026
	100	-0.044	$\mathbf{-0.077}$
$\alpha = 0.5$	0	-0.268	-0.268
	1	-0.430	-0.440
	5	-0.500	-0.520
	10	-0.527	-0.555
	100	-0.583	$\mathbf{-0.621}$
$\alpha = 1.$	0	-0.531	-0.531
	1	-0.569	-0.572
	5	-0.590	-0.600
	10	-0.600	-0.614
	100	-0.630	$\mathbf{-0.649}$

Table 3: $R(\pi)$ varying k and α in CIFAR100 (lower is better).

samples.

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

E.2 More Experiments

Effect of the quality of π_0. In the main experiments, we used $\alpha = 0.2$, which makes the starting behavior policy keep some information about the good actions while being diffuse enough to put mass on all actions. We give here a more comprehensive view of how the algorithms behave depending on α. The main take away, which reflects our first intuition, is that if we start far from $\pi_\star$, the improvement brought by sequential off-policy approaches is substantial compared to using the static behavior policy π_0 to collect data.

Effect of Multiple Deployments while varying α. We focus on CIFAR100 of large action space ($|\mathcal{A}| = 100$). $\alpha = 0.$ is the purely uniform behavior policy π_0, $\alpha = 0.5$ is a very good policy but is still diffuse and $\alpha = 1.$ is a peaked policy. Results are shown in Appendix E.2. SeqAdjLS outperforms SeqLS in all scenarios, with a bigger gap when π_0 is far from a good policy.

Effect of α on SCRM. We extend our evaluation of SCRM by varying the parameter $\alpha \in 0, 0.5, 1$, with k fixed at 10 across all datasets. When $\alpha = 0$, the behavior policy is entirely uniform, making the learning task considerably more difficult—particularly in environments with large action spaces—since the policy provides minimal guidance and infrequently selects low-cost actions. Despite these challenges, both SeqLS and SeqAdjLS consistently outperform SCRM. As α increases, the behavior policy becomes more informative, improving SCRM's performance. Nonetheless, our sequential methods, especially SeqAdjLS, continue to achieve superior results. The accompanying plots highlight the consistent advantage of SeqAdjLS across all settings.

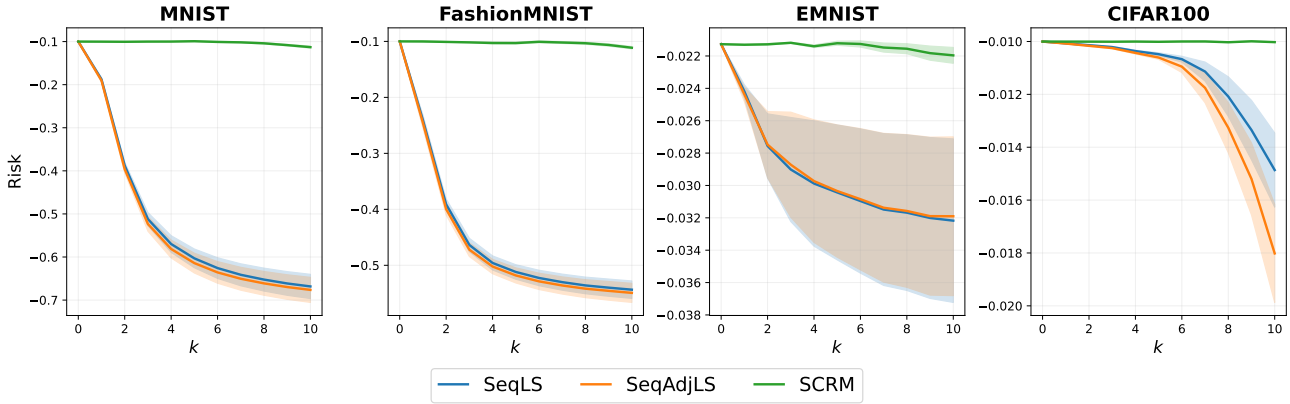

Figure 2: $R(\pi_k)$ at each intermediate step, for $k = 10$ and $\alpha = 0.$.

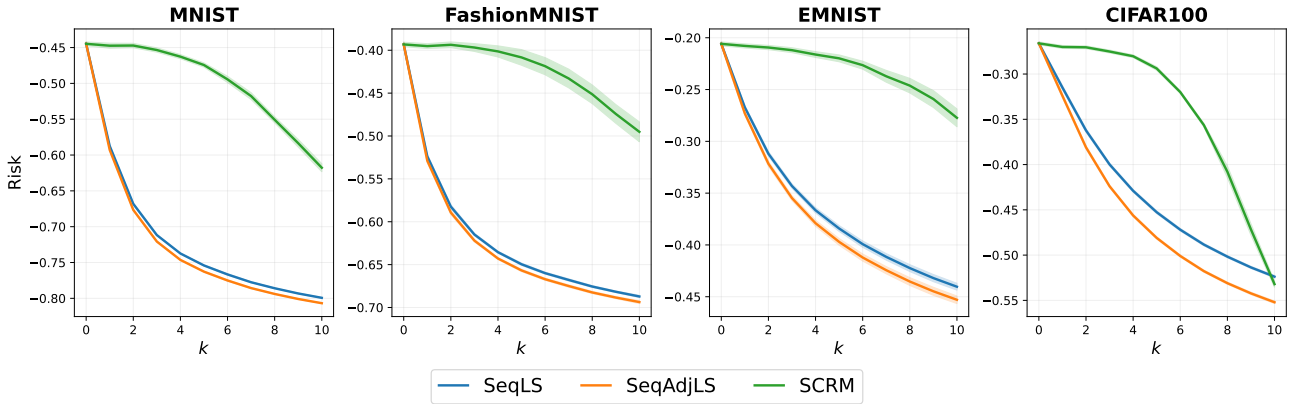

Figure 3: $R(\pi_k)$ at each intermediate step, for $k = 10$ and $\alpha = 0.5$.

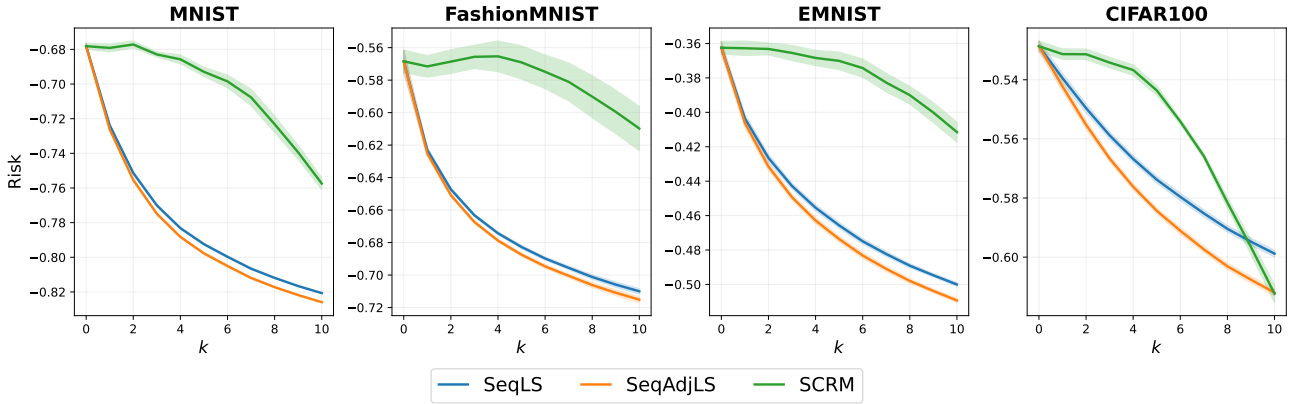

Figure 4: $R(\pi_k)$ at each intermediate step, for $k = 10$ and $\alpha = 1$.