Exponential Smoothing for Off-Policy Learning

Imad Aouali [1 2] **Victor-Emmanuel Brunel** [2] **David Rohde** [1] **Anna Korba** [2]

Abstract

Off-policy learning (OPL) aims at finding improved policies from logged bandit data, often by minimizing the inverse propensity scoring (IPS) estimator of the risk. In this work, we investigate a smooth regularization for IPS, for which we derive a two-sided PAC-Bayes generalization bound. The bound is tractable, scalable, interpretable and provides learning certificates. In particular, it is also valid for standard IPS without making the assumption that the importance weights are bounded. We demonstrate the relevance of our approach and its favorable performance through a set of learning tasks. Since our bound holds for standard IPS, we are able to provide insight into when regularizing IPS is useful. Namely, we identify cases where regularization might not be needed. This goes against the belief that, in practice, clipped IPS often enjoys favorable performance than standard IPS in OPL.

1. Introduction

An off-policy contextual bandit (Dudík et al., 2011) is a ubiquitous framework to optimize decision-making using offline data. In practice, logged data reflecting the preferences of the agent in an online setting is available (Bottou et al., 2013). In each round, the agent observes a *context*, takes an *action*, and receives a *reward* that depends on the observed context and the taken action. Off-policy evaluation (OPE) (Dudík et al., 2011) aims at evaluating a policy offline by designing an estimator of its expected reward using logged data. The estimator is often based on the importance sampling trick and it is generally referred to as inverse propensity scoring (IPS) (Horvitz & Thompson, 1952). Off-policy learning (OPL) leverages the latter estimator to learn an improved policy (Swaminathan & Joachims, 2015a).

The literature on OPL has focused so far on using learning principles derived from generalization bounds. First, Swaminathan & Joachims (2015a) used sample variance penalization (SVP) that favors policies with high estimated reward and low empirical variance. Recently, London & Sandler (2019) derived a novel scalable learning principle that favors policies with high estimated reward and whose parameter is not far from that of the logging policy in terms of L_2 distance. While derived from generalization bounds, these learning principles do not give any guarantees on the expected performance of the learned policy. Also, they require additional care to tune their hyper-parameters. Thus, motivated by the results in Sakhi et al. (2022), we derive tractable generalization bounds that we optimize directly.

The paper is organized as follows. In Section 2, we introduce the necessary background. In Section 3, we explain the shortcomings of the widely used *hard clipping* of IPS and present a smoother correction, called exponential smoothing. In Section 4, we focus on OPL and leverage PAC-Bayes theory to derive a *two-sided* generalization bound for our estimator. In contrast with prior works (Swaminathan & Joachims, 2015a; London & Sandler, 2019; Sakhi et al., 2022), our bound is also valid for standard IPS without clipping, and this is without assuming that the importance weights are bounded. We also discuss our results in detail in Section 5. In particular, we give insights into the sample complexity of our learning procedure, an important question not addressed in prior OPL works. Finally, we show in Section 6 that our approach enjoys favorable performance. A detailed comparative review of the literature is provided in Appendix A. The proofs are deferred to Appendices B and C. Refer to Appendix D to reproduce our experiments.

2. Background

Consider an agent interacting with a *contextual bandit* environment over n rounds. In round $t \in [n]$, the agent observes a *context* $x_t \sim \nu$, where ν is a distribution whose support $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$. Then the agent takes an *action* $a_t \in \mathcal{A} = [K]$. Finally, the agent receives a stochastic cost $c_t \in [-1, 0]$ that depends on both x_t and a_t. That is $c_t \sim p(\cdot|x_t, a_t)$ where $p(\cdot|x, a)$ is the *cost distribution* of action a in context x. We let $c(x, a) = \mathbb{E}_{c\sim p(\cdot|x,a)}[c]$ be the *cost function* that outputs the expected cost of action a in context x. Here we use a negative cost since it is seen as the

[1]CREST, ENSAE, IP Paris, France [2]Criteo AI Lab, Paris, France. Correspondence to: Imad Aouali <i.aouali@criteo.com>.

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

negative value of the reward, that is for any $(x, a) \in \mathcal{X} \times \mathcal{A}$, $c(x, a) = -r(x, a)$ where $r : \mathcal{X} \times \mathcal{A} \to [0, 1]$ is the *reward function* that outputs the expected reward of a in context x.

The agent is represented by a stochastic policy π. Given a context $x \in \mathcal{X}$, $\pi(\cdot|x)$ is a probability distribution over $\mathcal{A}$. Our goal is to find a policy $\pi \in \Pi$ among a set of policies Π that minimizes the risk defined as

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

where μ_π is the joint distribution of (x, a, c); $\mu_\pi(x, a, c) = \nu(x)\pi(a|x)p(c|x, a)$. We assume access to logged data $\mathcal{D}_n = (x_i, a_i, c_i)_{i\in[n]}$, where $(x_i, a_i, c_i) \sim \mu_{\pi_0}$ are i.i.d. and π_0 is a *known logging policy*. Given a policy $\pi \in \Pi$, OPE consists in building an estimator for its risk $R(\pi)$ using $\mathcal{D}_n$ such as $\hat{R}_n(\pi) \approx R(\pi)$. After that, OPL is used to find a policy $\hat{\pi}_n \in \Pi$ such that $R(\hat{\pi}_n) \approx \min_{\pi\in\Pi} R(\pi)$.

In this work, we focus on inverse propensity scoring (IPS) (Horvitz & Thompson, 1952; Dudík et al., 2012). Given a policy $\pi \in \Pi$, IPS estimates the risk $R(\pi)$ by re-weighting the samples using the ratio between π and π_0 such as

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

where for any $(x, a) \in \mathcal{X} \times \mathcal{A}$, $w_\pi(a|x) = \pi(a|x)/\pi_0(a|x)$ are the *importance weights*. The variance of $\hat{R}_n^{\text{IPS}}(\pi)$ scales linearly with the importance weights (Swaminathan et al., 2017) which can be large. Thus other OPE methods that do not rely on the importance weights or partially use them were proposed and they can be categorized into two families, direct method (DM) (Jeunen & Goethals, 2021) and doubly robust (DR) (Dudík et al., 2011). The reader may refer to Appendix A.1 for more details about these methods.

Let $\hat{R}_n$ be an estimator of the risk R. For instance, $\hat{R}_n$ can be $\hat{R}_n^{\text{IPS}}$ in (2). The goal in OPL is to minimize the risk R. But since we cannot access it, we only search for $\hat{\pi}_n = \operatorname{argmin}_{\pi\in\Pi} \hat{R}_n(\pi) + \text{pen}(\pi)$ hoping that $R(\hat{\pi}_n) \approx \min_{\pi\in\Pi} R(\pi)$. Here $\text{pen}(\cdot)$ is a penalization term obtained using generalization bounds of the following form. Let $\delta \in (0, 1)$, then we have with probability at least $1 - \delta$ that

$$R(\pi) \leq \hat{R}_n(\pi) + g(\delta, \Pi, \pi, \pi_0, n) \,, \quad \forall \pi \in \Pi \,, \quad (3)$$

for some function g. Improving upon π_0, that is when $R(\pi) - R(\pi_0) < 0$, is guaranteed with high probability when $\hat{R}_n(\pi) + g(\delta, \Pi, \pi, \pi_0, n) - R(\pi_0) < 0$. Thus we minimize $\hat{R}_n(\pi) + g(\delta, \Pi, \pi, \pi_0, n) - R(\pi_0)$ *in the hope* that the minimum is smaller than 0. Since $R(\pi_0)$ is fixed, the final objective reads

$$\hat{\pi}_n = \operatorname*{argmin}_{\pi\in\Pi} \hat{R}_n(\pi) + g(\delta, \Pi, \pi, \pi_0, n) \,. \quad (4)$$

This motivated the concept of *counterfactual risk minimization (CRM)* in Swaminathan & Joachims (2015a); London & Sandler (2019); Sakhi et al. (2022). However, all these works only derived one-sided inequalities similar to (3). In contrast, we derive *two-sided* inequalities of the form

$$|R(\pi) - \hat{R}_n(\pi)| \leq g(\delta, \Pi, \pi, \pi_0, n) \,, \quad \forall \pi \in \Pi \,. \quad (5)$$

This is because (5) can attest to the quality of the estimator $\hat{R}_n$. A one-sided one fails at this. To see why, note that we have with probability 1 that $R(\pi) \leq \hat{R}_n^{\text{POOR}}(\pi)$ with $g(\delta, \Pi, \pi, \pi_0, n) = 0$, considering a poor estimator of the risk, $\hat{R}_n^{\text{POOR}}(\pi) = 0$ for any $\pi \in \Pi$. This holds since by definition $R(\pi) \in [-1, 0]$ while $\hat{R}_n^{\text{POOR}}(\pi) = 0$ for any $\pi \in \Pi$. While this one-sided inequality holds for $\hat{R}_n^{\text{POOR}}$, this estimator is not informative at all about R, so minimizing it is not relevant. This is why we need to control the quality of the upper bound on R, and this is achieved by two-sided inequalities similar to (5). Also, (5) leads to oracle inequalities of the form $R(\hat{\pi}_n) \leq R(\pi_*) + 2g(\delta, \Pi, \pi_*, \pi_0, n)$, where $\hat{\pi}_n$ is the learned policy in (4) and $\pi_* = \operatorname{argmin}_{\pi\in\Pi} R(\pi)$ is the optimal policy. This allows us to quantify the number of samples n needed so that the risk of the learned policy $R(\hat{\pi}_n)$ is close to the optimal one $R(\pi_*)$.

Moreover, in many prior works, the objective in (4) is not optimized directly. Instead, the function g is used to motivate a heuristic-based learning principle. Here we review these principles briefly. But the reader may refer to Appendix A.2 for more detail. First, Swaminathan & Joachims (2015a) minimized the estimated risk while penalizing its empirical variance. This was inspired by a function g that contains a variance term; discarding more complicated terms like the covering number of the space of policies Π. Similarly, London & Sandler (2019) parameterize policies by a mean parameter and propose to penalize the estimated risk by the L_2 distance between the mean of the logging and the learning policies; discarding all the other terms from their bound. In contrast, we follow the theoretically grounded approach that consists in directly optimizing the objective in (4) as it is. It may also be relevant to note that some works (Metelli et al., 2021) derived *evaluation* bounds and used them in OPL. In evaluation, we *fix a policy* $\pi \in \Pi$, and show that

$$\mathbb{P}(|R(\pi) - \hat{R}_n(\pi)| \leq f(\delta, \pi, \pi_0, n)) \geq 1 - \delta \,,$$

for some function f that does not necessarily depend on the space of policies Π. In contrast, the generalization bound in (5) holds simultaneously for any policy $\pi \in \Pi$, and it is the one that should be used in OPL. That said, in this work, we derive a *two-sided* generalization bound that holds *simultaneously for any policy* $\pi \in \Pi$ as in (5).

3. Exponential Smoothing

The estimator $\hat{R}_n^{\text{IPS}}(\pi)$ in (2) is unbiased when $\pi_0(a|x) = 0$ implies that $\pi(a|x) = 0$ for any $(x, a) \in \mathcal{X} \times \mathcal{A}$. But its variance can be large as it grows linearly with the importance

weights $w_\pi(a|x)$. Thus they are often clipped (Swaminathan & Joachims, 2015a) such as on the following estimators

$$\texttt{IPS-min} \quad \tilde{R}_n^{\mathrm{M}}(\pi) = \frac{1}{n}\sum_{i=1}^{n} c_i \min\left(w_\pi(a_i|x_i), M\right),$$
$$\texttt{IPS-max} \quad \hat{R}_n^{\tau}(\pi) = \frac{1}{n}\sum_{i=1}^{n} c_i \frac{\pi(a_i|x_i)}{\max(\pi_0(a_i|x_i), \tau)}. \quad (6)$$

Here IPS-min clips the weights while IPS-max only clips π_0 in the denominator since π is always smaller than 1. For instance, $M \in \mathbb{R}^+$ in $\tilde{R}_n^{\mathrm{M}}(\pi)$ trades the bias and variance of the estimator. When M is large, the bias of $\tilde{R}_n^{\mathrm{M}}(\pi)$ is small but its variance may be large. On the other hand, the variance goes to 0 when $M \approx 0$ since in that case $\tilde{R}_n^{\mathrm{M}}(\pi) \approx 0$ for any $\pi \in \Pi$. Similarly, $\tau \in [0, 1]$ trades the bias and variance of $\hat{R}_n^{\tau}(\pi)$ and can be seen as $\tau \approx \frac{1}{M}$.

This *hard* clipping has some limitations. First, $\min(\cdot, M)$ leads to non-differentiable objectives that may require additional care in optimization (Papini et al., 2019). Also, $\min(\cdot, M)$ is constant on $[M, \infty)$ leading to objectives with zero gradients for any policy π that satisfies $w_\pi(a_i|x_i) > M$ for any $i \in [n]$. More importantly, hard clipping is sensitive to the choice of the clipping threshold M. In practice, tuning M is challenging and may cause the learned policy to match the logging policy, leading to minimal improvements. To see this, consider the following illustrative example.

For simplicity, suppose that the problem is non-contextual, in which case the reward function r only depends on the actions $a \in \mathcal{A}$. It follows that policies do not depend on $x \in \mathcal{X}$; they are now probability distributions $\pi(\cdot)$ over $\mathcal{A}$. Also, assume that $\mathcal{A} = [100]$ and that the reward received after taking action $a \in [100]$ is binary. That is, $r \sim \mathrm{Bern}(r(a))$ where $r(a) = 0.1 - 10^{-3}(a - 1)$ is the expected reward of action a, and for any $p \in [0, 1]$, $\mathrm{Bern}(p)$ is the Bernoulli distribution with parameter p. This means that the best action is 1 and the worst is 100. Finally, the logging policy $\pi_0(\cdot)$ is ϵ-greedy centered at action 50. That is $\pi_0(50) = 1 - \epsilon$, and for any $a \neq 50$, $\pi_0(a) = \frac{\epsilon}{99}$, with $\epsilon = 0.05$,.

Now consider 100 deterministic policies $\pi_a(\cdot)$ for $a \in [100]$ such that $\pi_a(\cdot)$ is the Dirac distribution centered at a. In Figure 1, we plot the estimated reward of the policies π_a using either IPS in (2) or IPS-min in (6). We generate $n = 50$k samples and set $M = 100 = \mathcal{O}(\sqrt{n})$ as suggested by Ionides (2008). *With this choice of* M, IPS-min underestimates the reward of all policies π_a for $a \neq 100$ since their weights π_a/π_0 are either 0 or $99/\epsilon > M$. The estimated reward of IPS-min is maximized in $\pi_{50} \approx \pi_0$ only. Thus, if we optimize $\tilde{R}_n^{M}(\cdot)$ over Dirac policies, we will converge to the logging policy despite its bad performance.

Although the other variant of hard clipping, IPS-max in (6), is differentiable, it is still sensitive to τ and may induce high bias similar to Figure 1. This is due to some loss of information related to the preferences of the logging policy. Indeed, for two actions a and a' such that $\pi_0(a \mid x_i) \ll \pi_0(a' \mid x_i) < \tau$ for an observed context x_i, the propensity scores $\pi_0(a \mid x_i)$ and $\pi_0(a' \mid x_i)$ will be clipped to the same value τ. Thus the information that, for context x_i, action a' is preferred by the logging policy than action a will be lost.

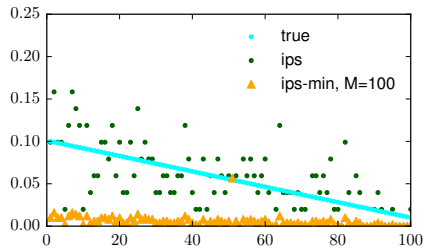

Figure 1. Effect of hard clipping on the estimation quality. The x-axis corresponds to actions $a \in [100]$. The y-axis is the estimated reward of each of the 100 policies π_a using either IPS or IPS-min. The cyan line is the true reward for each policy π_a.

To mitigate this, we propose the following *exponential smoothing* correction for IPS. Our estimators are defined as

$$\texttt{IPS-}\alpha: \ \hat{R}_n^{\alpha}(\pi) = \frac{1}{n}\sum_{i=1}^{n} c_i \hat{w}_\pi^{\alpha}(a_i|x_i), \ \alpha \in [0, 1],$$
$$\texttt{IPS-}\beta: \ \tilde{R}_n^{\beta}(\pi) = \frac{1}{n}\sum_{i=1}^{n} c_i \tilde{w}_\pi^{\beta}(a_i|x_i), \ \beta \in [0, 1], \quad (7)$$

where $\hat{w}_\pi^{\alpha}(a|x) = \frac{\pi(a|x)}{\pi_0(a|x)^\alpha}$ and $\tilde{w}_\pi^{\beta}(a|x) = \frac{\pi(a|x)^\beta}{\pi_0(a|x)^\beta}$. Here standard IPS is recovered for $\alpha = 1$ and $\beta = 1$. These estimators are differentiable in π and do not suffer from stationary points in optimization as they are not constant in π when $\beta \neq 0$ and $\alpha \neq 0$. Also, in contrast with IPS-max in (6), $\hat{R}_n^{\alpha}(\pi)$ preserves the preferences of the logging policy. Precisely, for two actions a and a' such that $\pi_0(a \mid x_i) < \pi_0(a' \mid x_i)$ for an observed context x_i, we still have $\pi_0(a \mid x_i)^\alpha < \pi_0(a' \mid x_i)^\alpha$ and the information that action a' is preferred by the logging policy than action a is preserved.

While a similar correction to IPS-β was proposed in Korba & Portier (2022), its use in off-policy contextual bandits is novel. Also, Su et al. (2020); Metelli et al. (2021) regularized the importance weights w as $\frac{\lambda_1 w}{\lambda_1 + w^2}$, $\lambda_1 > 0$ and $\frac{w}{1 - \lambda_2 + \lambda_2 w}$, $\lambda_2 \in [0, 1]$, respectively. Thus, the expression of both corrections is very different from ours. More importantly, these corrections entail different properties than ours. Roughly speaking, our correction allows us to *simultaneously* **(1)** control a tuning parameter $\alpha \in [0, 1]$ that is in a bounded domain $[0, 1]$, **(2)** without constraining the resulting importance weights to be bounded, **(3)** and to obtain PAC-Bayes generalization guarantees as the correction $\frac{\pi}{\pi_0^\alpha}$ is linear in π; a technical requirement of our analysis. In contrast, Metelli et al. (2021); Su et al. (2020) do not provide generalization guarantees; they focus on OPE and only propose heuristics for OPL. Those heuristics are not

based on theory, in contrast with ours which is directly derived from our generalization bound. Also, our approach has favorable empirical performance (Appendix D.6).

Although Korba & Portier (2022, Lemma 1) show that smoothing the importance weights similarly to IPS-β in (7) reduces the variance, it might still be unclear how α and β trade the bias and variance of our estimators in off-policy contextual bandits. To see this, let $\alpha \in [0,1]$, then we have

$$|\mathbb{B}(\hat{R}_n^\alpha(\pi))| \leq \mathbb{E}_{x\sim\nu, a\sim\pi(\cdot|x)}\left[1-\pi_0(a|x)^{1-\alpha}\right], \quad (8)$$
$$\mathbb{V}\left[\hat{R}_n^\alpha(\pi)\right] \leq \frac{1}{n}\mathbb{E}_{x\sim\nu, a\sim\pi(\cdot|x)}\left[\frac{\pi(a|x)}{\pi_0(a|x)^{2\alpha-1}}\right],$$

with $\mathbb{B}(\hat{R}_n^\alpha(\pi)) = \mathbb{E}[\hat{R}_n^\alpha(\pi)] - R(\pi)$ and $\mathbb{V}[\hat{R}_n^\alpha(\pi)] = \mathbb{E}[(\hat{R}_n^\alpha(\pi) - \mathbb{E}[\hat{R}_n^\alpha(\pi)])^2]$ are respectively the bias and the variance of $\hat{R}_n^\alpha(\pi)$. The bound of the bias in (8) is minimized in $\alpha = 1$ (standard IPS); in which case it is equal to 0 (standard IPS is unbiased). In contrast, the bound of the variance is minimized in $\alpha = 0$. Thus if the variance is small or n is large enough such that $\mathbb{E}[\pi(a|x)/\pi_0(a|x)^{2\alpha-1}]/n \to 0$, then we set $\alpha \to 1$. Otherwise, we set $\alpha \to 0$. This shows that α trades the bias and variance of $\hat{R}_n^\alpha$. More details and a similar discussion for $\tilde{R}_n^\beta(\pi)$ are deferred to Appendix B.

4. PAC-Bayes Analysis for Off-Policy Learning

We now derive generalization bounds for our estimator. We opt for the PAC-Bayes framework for the following reasons. First, it is known to provide some of the tightest generalization bounds in challenging scenarios (Farid & Majumdar, 2021), for aggregated and randomized predictors (Alquier, 2021). Second, the bounds have a Kullback–Leibler (KL) divergence (Van Erven & Harremos, 2014) term $D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P})$ that depends on a *fixed prior* $\mathbb{P}$ and a *learning posterior* $\mathbb{Q}$ (see Section 4.1 for a brief introduction). This quantity can be seen as a complexity measure, similarly to the covering number (Maurer & Pontil, 2009). The difference is that complexity measures are uniform on the space of policies while the KL term in PAC-Bayes depends on the prior $\mathbb{P}$ and the posterior $\mathbb{Q}$. This allows getting sharper bounds when the former is well chosen. Third, the PAC-Bayes perspective fits very well with OPL. In fact, a policy π can be written as an aggregation of predictors under some distribution $\mathbb{Q}$. Thus the prior $\mathbb{P}$ can be associated with the logging policy π_0 that we want to improve upon while the posterior $\mathbb{Q}$ is related to the learning policy π. Fourth, London & Sandler (2019) showed that PAC-Bayes can lead to tractable and scalable objectives, an important consideration in practice.

4.1. Elements of PAC-Bayes

Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ be an instance space: e.g., $\mathcal{X}$ and $\mathcal{Y}$ are the input and output space in supervised learning. Let $\mathcal{H} = \{h : \mathcal{X} \to \mathcal{Y}\}$ denote a hypothesis space of mappings from $\mathcal{X}$ to $\mathcal{Y}$ (predictors). Also, let $L : \mathcal{H}\times\mathcal{Z} \to \mathbb{R}$ be a loss function and assume access to data $\mathcal{D}_n = (z_i)_{i\in[n]}$ drawn from an unknown distribution $\mathbb{D}$. Let $R(h) = \mathbb{E}_{z\sim\mathbb{D}}[L(h,z)]$ be the risk of $h \in \mathcal{H}$ while $\hat{R}_n(h) = \frac{1}{n}\sum_{i=1}^n L(h, z_i)$ is its empirical counterpart. Then the main focus in PAC-Bayes is to study the generalization capabilities of random hypothesis $\mathbb{Q}$ on $\mathcal{H}$ by controlling the gap between the expected risk under $\mathbb{Q}$, $\mathbb{E}_{h\sim\mathbb{Q}}[R(h)]$ and the expected empirical risk under $\mathbb{Q}$, $\mathbb{E}_{h\sim\mathbb{Q}}[\hat{R}_n(h)]$. For example, assume that $L(h,z) \in [0,1]$ for any $(h,z) \in \mathcal{H}\times\mathcal{Z}$, let $\mathbb{P}$ be a *fixed prior* distribution on $\mathcal{H}$ and let $\delta \in (0,1)$. Then with probability at least $1-\delta$ over $\mathcal{D}_n \sim \mathbb{D}^n$, the following inequality holds *simultaneously for any posterior* distribution $\mathbb{Q}$ on $\mathcal{H}$

$$\mathbb{E}_{h\sim\mathbb{Q}}[R(h)] \leq \mathbb{E}_{h\sim\mathbb{Q}}[\hat{R}_n(h)] + \sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{2\sqrt{n}}{\delta}}{2n}}.$$

This was originally proposed by McAllester (1998), and the reader may refer to Alquier (2021); Guedj (2019) for more elaborate introductions of PAC-Bayes theory.

4.2. PAC-Bayes for Off-Policy Contextual Bandits

Let $\mathcal{H} = \{h : \mathcal{X} \to \mathcal{A}\}$ be a hypothesis space of mappings from $\mathcal{X}$ (contexts) to $\mathcal{A}$ (actions). Given a policy π and a context $x \in \mathcal{X}$, the action distribution $\pi(\cdot|x)$ is induced by a distribution $\mathbb{Q}$ over $\mathcal{H}$ (London & Sandler, 2019) such as

$$\pi(a|x) = \pi_{\mathbb{Q}}(a|x) = \mathbb{E}_{h\sim\mathbb{Q}}\left[\mathbb{I}_{\{h(x)=a\}}\right]. \quad (9)$$

This is not an assumption since any policy π has this form when $\mathcal{H}$ is rich enough (Sakhi et al., 2022, Theorem 2). From (9), we observe that policies can be seen as an aggregation $\mathbb{E}_{h\sim\mathbb{Q}}[\cdot]$ (under some distribution $\mathbb{Q}$ on the predefined hypothesis space $\mathcal{H}$) of deterministic decision rules $\mathbb{I}_{\{h(x)=a\}}$. This allows formulating OPL as a PAC-Bayes problem. Before showing how this is achieved, we start by providing two practical policies of such form.

Example 1 (softmax and mixed-logit policies): we define the space $\mathcal{H} = \left\{h_{\theta,\gamma}; \theta \in \mathbb{R}^{dK}, \gamma \in \mathbb{R}^K\right\}$ of mappings $h_{\theta,\gamma}(x) = \mathrm{argmax}_{a\in\mathcal{A}}\, \phi(x)^\top\theta_a + \gamma_a$. Here $\phi(x)$ outputs a d-dimensional representation of x, and γ_a is a standard Gumbel perturbation, $\gamma_a \sim \mathrm{G}(0,1)$ for any $a \in \mathcal{A}$. Then

$$\pi_\theta^{\mathrm{SOF}}(a|x) = \frac{\exp(\phi(x)^\top\theta_a)}{\sum_{a'\in\mathcal{A}}\exp(\phi(x)^\top\theta_{a'})},$$
$$\overset{(i)}{=} \mathbb{E}_{\gamma\sim\mathrm{G}(0,1)^K}\left[\mathbb{I}_{\{h_{\theta,\gamma}(x)=a\}}\right], \quad (10)$$

where (i) follows from the Gumbel-Max trick (GMT) (Luce, 2012; Maddison et al., 2014). Thus a `softmax` policy $\pi_\theta^{\mathrm{SOF}}$ can be written as in (9). Now we also consider random parameters $\theta \sim \mathcal{N}(\mu, \sigma^2 I_{dK})$ with $\mu \in \mathbb{R}^{dK}$ and $\sigma > 0$. Then, let $\mathbb{Q} = \mathcal{N}(\mu, \sigma^2 I_{dK}) \times \mathrm{G}(0,1)^K$, it follows that

$\pi_{\mathbb{Q}} = \pi^{\text{MIXL}}_{\mu,\sigma}$ is a mixed-logit policy and it reads

$$\pi^{\text{MIXL}}_{\mu,\sigma}(a|x) = \mathbb{E}_{\theta\sim\mathcal{N}(\mu,\sigma^2 I_d)}\left[\frac{\exp(\phi(x)^\top\theta_a)}{\sum_{a'\in\mathcal{A}}\exp(\phi(x)^\top\theta_{a'})}\right], \\ = \mathbb{E}_{\theta\sim\mathcal{N}(\mu,\sigma^2 I_d),\gamma\sim\text{G}(0,1)^K}\left[\mathbb{I}_{\{h_{\theta,\gamma}(x)=a\}}\right]. \quad (11)$$

Example 2 (Gaussian policies): Sakhi et al. (2022) removed the Gumbel noise γ in (11) and consequently defined the hypothesis space as $\mathcal{H} = \{h_\theta\,;\theta\in\mathbb{R}^{dK}\}$ of mappings $h_\theta(x) = \text{argmax}_{a\in\mathcal{A}}\,\phi(x)^\top\theta_a$ for any $x\in\mathcal{X}$. Then, let $\mathbb{Q} = \mathcal{N}(\mu,\sigma^2 I_{dK})$, it follows that $\pi_{\mathbb{Q}} = \pi^{\text{GAUS}}_{\mu,\sigma}$ reads

$$\pi^{\text{GAUS}}_{\mu,\sigma}(a|x) = \mathbb{E}_{\theta\sim\mathcal{N}(\mu,\sigma^2 I_d)}\left[\mathbb{I}_{\{h_\theta(x)=a\}}\right]. \quad (12)$$

To see why removing the Gumbel noise can be beneficial, the reader may refer to Appendix D.2. After motivating the definition of policies in (9), we are in a position to relate our estimators to the general PAC-Bayes framework in Section 4.1. One technical requirement of our proof is that the estimator should be linear in π. Thus we focus on $\hat{R}^\alpha_n(\cdot)$ since $\tilde{R}^\beta_n(\pi)$ is non-linear in π. Let $h\in\mathcal{H}$, $x\in\mathcal{X}$, $a\in\mathcal{A}$ and $c\in[-1,0]$, we define the loss L_α as

$$L_\alpha(h,x,a,c) = \frac{\mathbb{I}_{\{h(x)=a\}}}{\pi_0(a|x)^\alpha}c. \quad (13)$$

Using the definition in (9) and the linearity of the expectation, we have that $\hat{R}^\alpha_n(\cdot)$ in (7) can be written as

$$\hat{R}^\alpha_n(\pi_{\mathbb{Q}}) = \mathbb{E}_{h\sim\mathbb{Q}}\left[\frac{1}{n}\sum_{i=1}^n L_\alpha(h,x_i,a_i,c_i)\right].$$

Moreover, the expectation of $\hat{R}_n(\pi_{\mathbb{Q}})$ reads

$$R^\alpha(\pi_{\mathbb{Q}}) = \mathbb{E}_{h\sim\mathbb{Q}}\mathbb{E}_{(x,a,c)\sim\mu_{\pi_0}}\left[L_\alpha(h,x,a,c)\right].$$

Finally, the main quantity of interest, the risk $R(\pi_{\mathbb{Q}})$, can be expressed in terms of the loss with $\alpha = 1$, L_1, as

$$R(\pi_{\mathbb{Q}}) = \mathbb{E}_{h\sim\mathbb{Q}}\mathbb{E}_{(x,a,c)\sim\mu_{\pi_0}}\left[L_1(h,x,a,c)\right].$$

Since $\hat{R}^\alpha_n(\pi_{\mathbb{Q}})$ is an unbiased estimator of $R^\alpha(\pi_{\mathbb{Q}})$, PAC-Bayes can be used to bound $R^\alpha(\pi_{\mathbb{Q}}) - \hat{R}^\alpha_n(\pi_{\mathbb{Q}})$. This will allow bounding our quantity of interest $R(\pi_{\mathbb{Q}}) - \hat{R}^\alpha_n(\pi_{\mathbb{Q}})$.

4.3. Main Result

To ease the exposition, we assume that the costs are deterministic. Then, in logged data $\mathcal{D}_n$, $c_i = c(x_i,a_i)$ for any $i\in[n]$. Note that the same result holds for stochastic costs. We discuss our result and sketch its proof in Section 5. The complete proof can be found in Appendix C.1.

Theorem 4.1. *Let $\lambda > 0$, $n\geq 1$, $\delta\in(0,1)$, $\alpha\in[0,1]$, and let $\mathbb{P}$ be a fixed prior on $\mathcal{H}$, then with probability at least $1-\delta$ over draws $\mathcal{D}_n\sim\mu^n_{\pi_0}$, the following holds simultaneously for any posterior $\mathbb{Q}$ on $\mathcal{H}$*

$$|R(\pi_{\mathbb{Q}}) - \hat{R}^\alpha_n(\pi_{\mathbb{Q}})| \leq \sqrt{\frac{\text{KL}_1(\pi_{\mathbb{Q}})}{2n}} + B^\alpha_n(\pi_{\mathbb{Q}}) + \frac{\text{KL}_2(\pi_{\mathbb{Q}})}{n\lambda} + \frac{\lambda}{2}\bar{V}^\alpha_n(\pi_{\mathbb{Q}}).$$

where $\text{KL}_1(\pi_{\mathbb{Q}}) = D_{\text{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{4\sqrt{n}}{\delta}$, *and*

$$\text{KL}_2(\pi_{\mathbb{Q}}) = D_{\text{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{4}{\delta},$$

$$B^\alpha_n(\pi_{\mathbb{Q}}) = 1 - \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{a\sim\pi_{\mathbb{Q}}(\cdot|x_i)}\left[\pi_0^{1-\alpha}(a|x_i)\right],$$

$$\bar{V}^\alpha_n(\pi_{\mathbb{Q}}) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{\pi_{\mathbb{Q}}(a|x_i)}{\pi_0(a|x_i)^{2\alpha}}\right] + \frac{\pi_{\mathbb{Q}}(a_i|x_i)c_i^2}{\pi_0(a_i|x_i)^{2\alpha}}.$$

We start by clarifying that the prior $\mathbb{P}$ can be any fixed distribution on $\mathcal{H}$. If we have access to $\mathbb{P}_0$ on $\mathcal{H}$ such that $\pi_0 = \pi_{\mathbb{P}_0}$, then it is natural to set $\mathbb{P} = \mathbb{P}_0$. But this is just a choice and one may use priors that do not depend on π_0. Now we explain the main terms in our bound. First, the terms $\text{KL}_1(\pi_{\mathbb{Q}})$ and $\text{KL}_2(\pi_{\mathbb{Q}})$ contain the divergence $D_{\text{KL}}(\mathbb{Q}\|\mathbb{P})$ which penalizes posteriors $\mathbb{Q}$ that differ a lot from the prior $\mathbb{P}$. Moreover, $B^\alpha_n(\pi_{\mathbb{Q}})$ is the bias conditioned on the contexts $(x_i)_{i\in[n]}$; $B^\alpha_n(\pi_{\mathbb{Q}}) = 0$ when $\alpha = 1$ and $B^\alpha_n(\pi_{\mathbb{Q}}) > 0$ otherwise. Also, the first term in $\bar{V}^\alpha_n(\pi_{\mathbb{Q}})$ *resembles* the theoretical second moment of the regularized importance weights $\frac{\pi}{\pi_0^\alpha}$ (without the cost) when they are seen as random variables. Similarly, the second term in $\bar{V}^\alpha_n(\pi_{\mathbb{Q}})$ *resembles* the empirical second moment of $\frac{\pi}{\pi_0^\alpha}c$ (with the cost). Finally, if $\bar{V}^\alpha_n(\pi_{\mathbb{Q}})$ is bounded, then we can set $\lambda = 1/\sqrt{n}$, in which case our bound scales as $\mathcal{O}(1/\sqrt{n} + B^\alpha_n(\pi_{\mathbb{Q}}))$. In practice, we set $\alpha\approx 1$ leading to $B^\alpha_n(\pi_{\mathbb{Q}})) \approx 0$ and the bound would scale as $\mathcal{O}(1/\sqrt{n})$.

This bound motivates the idea that we only need to control the second moments $\bar{V}^\alpha_n(\pi_{\mathbb{Q}})$ to get generalization guarantees for $\hat{R}^\alpha_n(\cdot)$. In particular, one of the main strengths of our result is that it holds for standard IPS with $\alpha = 1$ under the assumption that $\bar{V}^1_n(\pi_{\mathbb{Q}})$ is bounded. This assumption is less restrictive than assuming that the importance weight as a random variable, $\pi_{\mathbb{Q}}(a|x)/\pi_0(a|x)$, is bounded, a required assumption for traditional concentration bounds. In contrast, $\bar{V}^\alpha_n(\pi_{\mathbb{Q}})$ only involves the *expectations* of the random variables $\pi_{\mathbb{Q}}(a|x_i)/\pi_0(a|x_i)^{2\alpha}$, and ratios of π_0 evaluated at observed contexts and actions and $(x_i,a_i)_{i\in[n]}$, that have non-zero probabilities under π_0 by definition.

Our result holds for fixed $\lambda > 0$ and $\alpha\in[0,1]$. In Appendix C.2, we extend this to any potentially data-dependent $\lambda\in(0,1)$ and $\alpha\in(0,1]$. The assumption that $c\in[-1,0]$ can be relaxed to $c\in[-B,0]$ up to additional factors B^2 and B in $\bar{V}^\alpha_n(\pi_{\mathbb{Q}})$ and $\text{KL}_1(\pi_{\mathbb{Q}})$, respectively. Finally, our

bound is suitable for stochastic first-order optimization (Robbins & Monro, 1951) since data-dependent quantities are not inside a square root. This is important for scalability.

4.4. Adaptive and Data-Driven Tuning of α

Theorem 4.1 assumes that α is fixed (although we extend it for data-dependent α in Appendix C.2). However, providing a procedure to tune α in an adaptive and data-dependent fashion is important in practice. Thus we propose to set

$$\alpha_* = \operatorname*{argmin}_{\alpha\in[0,1]} B_n^\alpha(\pi_\mathbb{Q}) + \sqrt{\frac{2\mathrm{KL}_2(\pi_\mathbb{Q})\bar{V}_n^\alpha(\pi_\mathbb{Q})}{n}}\,, \quad (14)$$

where all the terms are defined in Theorem 4.1. Roughly speaking, α_* establishes a bias-variance trade-off; it minimizes the sum of the bias term $B_n^\alpha(\pi_\mathbb{Q})$ and the square root of the second moment term $\bar{V}_n^\alpha(\pi_\mathbb{Q})$, weighted by $\sqrt{\frac{2\mathrm{KL}_2(\pi_\mathbb{Q})}{n}}$. Here (14) is obtained by minimizing the bound in Theorem 4.1 with respect to both α and λ as follows. First, we minimize the bound in Theorem 4.1 with respect to λ; the minimizer is $\lambda_* = \sqrt{\frac{2\mathrm{KL}_2(\pi_\mathbb{Q})}{n\bar{V}_n^\alpha(\pi_\mathbb{Q})}}$. Then, the bound in Theorem 4.1 evaluated at $\lambda = \lambda_*$ becomes

$$\sqrt{\frac{\mathrm{KL}_1(\pi_\mathbb{Q})}{2n}} + B_n^\alpha(\pi_\mathbb{Q}) + \sqrt{\frac{2\mathrm{KL}_2(\pi_\mathbb{Q})\bar{V}_n^\alpha(\pi_\mathbb{Q})}{n}}\,. \quad (15)$$

Finally, α_* is defined as the minimizer of (15) with respect to $\alpha \in [0,1]$, and $\sqrt{\frac{\mathrm{KL}_1(\pi_\mathbb{Q})}{2n}}$ does not appear in (14) as it does not depend on α. Note that α_* depends on both logged data $\mathcal{D}_n$ and the learning policy $\pi_\mathbb{Q}$. Thus it is adaptive; its value changes in each iteration during optimization.

5. Discussion

We start by interpreting and comparing our results to related work. Then, we present the technical challenges in Section 5.2. After that, we sketch our proof in Section 5.3.

5.1. Interpretation and Comparison to Related Work

Theorem 4.1 gives insight into the number of samples needed so that the performance of $\hat{\pi}_n$ is close to that of the optimal policy π_*. To simplify the problem, we consider the Gaussian policies in (12) and assume that there exists $\mathbb{Q}_* = \mathcal{N}(\mu_*, I_{dK})$ with $\mu_* \in \mathbb{R}^{dK}$ such that the optimal policy is $\pi_* = \pi_{\mathbb{Q}_*}$. Also, we let the prior $\mathbb{P} = \mathcal{N}(\mu_0, I_{dK})$ and assume that π_0 is uniform. This is possible since as we said before, the prior $\mathbb{P}$ does not have to depend on the logging policy π_0. Then we have that $D_{\mathrm{KL}}(\mathbb{Q}_*\|\mathbb{P}) = \|\mu_* - \mu_0\|^2/2$, $B_n^\alpha(\pi_{\mathbb{Q}_*}) = 1 - 1/K^{1-\alpha}$ and $\bar{V}_n^\alpha(\pi_{\mathbb{Q}_*}) \le 2K^{2\alpha}$. The last inequality is not tight but it allows getting an easy-to-interpret term that does not depend on n. Now let $\epsilon > 2(1 - K^{\alpha-1})$ for $\alpha \in [1 - \log 2/\log K, 1]$. This condition on α ensures that $\epsilon \in [0,1]$ and it is mild as α is often close to 1. Then, it holds with high probability that

$$n \,\tilde{>}\, \Big(\frac{\|\mu_* - \mu_0\|^2 + K^{2\alpha}}{\epsilon - 2(1 - K^{\alpha-1})}\Big)^2 \implies R(\hat{\pi}_n) \le R(\pi_{\mathbb{Q}_*}) + \epsilon\,,$$

where we omit constant and logarithmic terms in $\tilde{>}$. This gives an intuition on the sample complexity for our procedure. In particular, fewer samples are needed in four cases. The first is when ϵ is large, which means that we afford to learn a policy whose performance is far from the optimal one. The second is when the prior $\mathbb{P}$ is close to $\mathbb{Q}_*$, that is when $\|\mu_* - \mu_0\|$ is small. This highlights that the choice of the prior $\mathbb{P}$ is important. The third is when the second-moment term $K^{2\alpha}$ is small. The fourth is when the bias $B_n^\alpha(\pi_{\mathbb{Q}_*})$ is small. In particular, when $\alpha = 1$, the bias is 0. In contrast, the second-moment term is minimized in $\alpha = 0$. This is where the choice of α matters. The proofs of these claims and more detail can be found in Appendix C.4.

Prior works (Swaminathan & Joachims, 2015a; London & Sandler, 2019; Sakhi et al., 2022) do not provide such insight for two reasons. They only derived one-sided inequalities and thus they cannot relate the risk of the learned policy with the optimal one as we discussed in the last three paragraphs of Section 2. Also, their bounds do not contain a bias term and as a result, they are minimized in $\tau = 1$. In contrast, ours have a bias term and this allows seeing the effect of α.

Our paper derives a *tractable generalization bound* for an estimator other than clipped IPS in (6), which also holds for the standard IPS in (2). The bounds in Swaminathan & Joachims (2015a); London & Sandler (2019); Sakhi et al. (2022) have a multiplicative dependency on the clipping threshold (M or $1/\tau$ in (6)). Standard IPS is recovered when $M \to \infty$ (or $\tau = 0$) in which case their bounds explode. We successfully avoid any similar dependency on α. Moreover, Swaminathan & Joachims (2015a); London & Sandler (2019) only used their generalization bounds to inspire learning principles. Although we directly optimize our theoretical bound (Theorem 4.1) in our experiments, our analysis also inspires a learning principle where we simultaneously penalize the L_2 distance, the variance and the bias. That is, we find $\mu \in \mathbb{R}^{dK}$ that minimizes

$$\hat{R}_n^\alpha(\pi_\mu) + \lambda_1\|\mu - \mu_0\|^2 + \lambda_2\bar{V}_n^\alpha(\pi_\mu) + \lambda_3 B_n^\alpha(\pi_\mu)\,. \quad (16)$$

Here λ_1, λ_2 and λ_3 are tunable hyper-parameters, π_μ can be the Gaussian policy in (12), $\pi_\mu = \pi_{\mu,1}^{\mathrm{GAUS}}$, with a fixed $\sigma = 1$, and μ_0 is the mean of the prior $\mathbb{P} = \mathcal{N}(\mu_0, I_{dK})$. Existing works either penalize the L_2 distance or the variance. For completeness, we also show that this learning principle should be preferred over existing ones in Appendix D.5.

5.2. Technical Challenges

London & Sandler (2019); Sakhi et al. (2022) derived PAC-

Bayes generalization bounds for the estimator IPS-max in (6). Extending their analyses to our case is not straightforward. First, their estimator IPS-max is lower bounded by $-1/\tau$, and thus they relied on traditional techniques for $[0,1]$-losses (Alquier, 2021). In contrast, our loss in (13) is not lower bounded, and controlling it without assuming that the importance weights are bounded is challenging.

Moreover, their bounds have a multiplicative dependency on $1/\tau$, hence they explode as $\tau \to 0$. This makes them vacuous for small values of τ and inapplicable to the standard IPS estimator in (2) recovered for $\tau = 0$. In contrast, our bound does not have a similar dependency on α and it is also valid for standard IPS recovered for $\alpha = 1$. Moreover, we derive two-sided inequalities rather than one-sided ones for the important reasons that we priorly discussed. This requires carefully controlling in *closed-form* the absolute value of the bias. Prior works only used that the bias is negative which was enough to obtain one-sided inequalities.

Explaining other challenges requires stating a result that inspired our analysis: Kuzborskij & Szepesvári (2019) derived PAC-Bayes generalization bounds for unbounded losses by only controlling their second moments. Recently, Haddouche & Guedj (2022) proposed a similar result using Ville's inequality (Bercu & Touati, 2008). Adapting their theorem to our problem is given Proposition 5.1. We slightly adapt their proof to get a *two-sided* inequality for a *negative* loss. The proof is deferred to Appendix C.3.

Proposition 5.1. *Let $\lambda > 0$, $n \geq 1$, $\delta \in (0,1)$, $\alpha \in [0,1]$ and let $\mathbb{P}$ be a fixed prior on $\mathcal{H}$, then with probability at least $1-\delta$ over draws $\mathcal{D}_n \sim \mu_{\pi_0}^n$, the following holds simultaneously for all posteriors, $\mathbb{Q}$, on $\mathcal{H}$*

$$|R^\alpha(\pi_\mathbb{Q}) - \hat{R}_n^\alpha(\pi_\mathbb{Q})| \leq \frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{2}{\delta}}{\lambda n} \tag{17}$$
$$+ \frac{\lambda}{2n}\sum_{i=1}^n \frac{\pi_\mathbb{Q}(a_i|x_i)}{\pi_0^{2\alpha}(a_i|x_i)}c_i^2 + \frac{\lambda}{2}\mathbb{E}_{(x,a,c)\sim\mu_{\pi_0}}\left[\frac{\pi_\mathbb{Q}(a|x)}{\pi_0^{2\alpha}(a|x)}c^2\right],$$

There are two main issues with Proposition 5.1. First, the term $\mathbb{E}_{(x,a,c)\sim\mu_{\pi_0}}\left[\frac{\pi_\mathbb{Q}(a|x)}{\pi_0^{2\alpha}(a|x)}c^2\right]$ in (17) is intractable. One could bound c^2 by 1, but the resulting term will still be intractable due to the expectation over the unknown distribution of contexts ν. Second, we need an upper bound of $|R(\pi_\mathbb{Q}) - \hat{R}_n^\alpha(\pi_\mathbb{Q})|$ while Proposition 5.1 only provides one for $|R^\alpha(\pi_\mathbb{Q}) - \hat{R}_n^\alpha(\pi_\mathbb{Q})|$. Thus it remains to quantify the approximation error $|R(\pi_\mathbb{Q}) - R^\alpha(\pi_\mathbb{Q})|$. This will also require computing an expectation over $x \sim \nu$, which is intractable.

5.3. Sketch of Proof for Theorem 4.1

We conclude by showing how the technical challenges above were solved. First, We decompose $R(\pi_\mathbb{Q}) - \hat{R}_n^\alpha(\pi_\mathbb{Q})$ as

$$R(\pi_\mathbb{Q}) - \hat{R}_n^\alpha(\pi_\mathbb{Q}) = I_1 + I_2 + I_3\,, \qquad \text{where}$$

$$I_1 = R(\pi_\mathbb{Q}) - \frac{1}{n}\sum_{i=1}^n R(\pi_\mathbb{Q}|x_i)\,,$$
$$I_2 = \frac{1}{n}\sum_{i=1}^n R(\pi_\mathbb{Q}|x_i) - \frac{1}{n}\sum_{i=1}^n R^\alpha(\pi_\mathbb{Q}|x_i)\,,$$
$$I_3 = \frac{1}{n}\sum_{i=1}^n R^\alpha(\pi_\mathbb{Q}|x_i) - \hat{R}_n^\alpha(\pi_\mathbb{Q})\,,$$
$$R(\pi_\mathbb{Q}|x_i) = \mathbb{E}_{a\sim\pi_\mathbb{Q}(\cdot|x_i)}\left[c(x_i,a)\right]\,,$$
$$R^\alpha(\pi_\mathbb{Q}|x_i) = \mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{\pi_\mathbb{Q}(a|x_i)}{\pi_0(a|x_i)^\alpha}c(x_i,a)\right].$$

I_1 is the estimation error of the empirical mean of the risk using n i.i.d. contexts $(x_i)_{i\in[n]}$. This term is introduced to avoid the intractable expectation over $x \sim \nu$. Moreover, I_2 is the bias term conditioned on the contexts $(x_i)_{i\in[n]}$ and we bound it in closed-form. Finally, I_3 is the estimation error of the risk conditioned on the contexts $(x_i)_{i\in[n]}$. Again, this conditioning allows us to avoid the intractable expectation over $x \sim \nu$ and to consequently bound $|I_3|$ by tractable terms. First, Alquier (2021, Theorem 3.3) yields that with probability at least $1-\frac{\delta}{2}$, it holds for any $\mathbb{Q}$ on $\mathcal{H}$ that

$$|I_1| \leq \sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{4\sqrt{n}}{\delta}}{2n}}\,.$$

Also, $|I_2|$ is bounded similarly to (8) as

$$|I_2| \leq \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{a\sim\pi_\mathbb{Q}(\cdot|x_i)}\left[1-\pi_0^{1-\alpha}(a|x_i)\right]\,.$$

Bounding $|I_3|$ is achieved by expressing it using martingale difference sequences $(f_i(a_i,h))_{i\in[n]}$ that we construct as follows. Let $(\mathcal{F}_i)_{i\in\{0\}\cup[n]}$ be a filtration adapted to $(S_i)_{i\in[n]}$ where $S_i = (a_\ell)_{\ell\in[i]}$ for any $i \in [n]$, we define

$$f_i(a_i,h) = \mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{\mathbb{I}_{\{h(x_i)=a\}}c(x_i,a)}{\pi_0(a|x_i)^\alpha}\right] - \frac{\mathbb{I}_{\{h(x_i)=a_i\}}c_i}{\pi_0(a_i|x_i)^\alpha}\,.$$

Then we show that for any $h \in \mathcal{H}$, $(f_i(a_i,h))_{i\in[n]}$ is a martingale difference sequence. After that, we apply Haddouche & Guedj (2022, Theorem 5) and obtain that with probability at least $1-\delta/2$, it holds for any $\mathbb{Q}$ on $\mathcal{H}$ that

$$|\mathbb{E}_{h\sim\mathbb{Q}}\left[M_n(h)\right]| \leq \frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{4}{\delta}}{\lambda} + \frac{\lambda}{2}\mathbb{E}_{h\sim\mathbb{Q}}\left[\bar{V}_n(h)\right],$$

where $M_n(h) = \sum_{i=1}^n f_i(a_i,h)$ and $\bar{V}_n(h) = \sum_{i=1}^n f_i(a_i,h)^2 + \mathbb{E}\left[f_i(a_i,h)^2|\mathcal{F}_{i-1}\right]$. Then notice that $\mathbb{E}_{h\sim\mathbb{Q}}\left[M_n(h)\right]$ can be expressed in terms of I_3 as

$$\mathbb{E}_{h\sim\mathbb{Q}}\left[M_n(h)\right] = \sum_{i=1}^n R^\alpha(\pi_\mathbb{Q}|x_i) - n\hat{R}_n^\alpha(\pi_\mathbb{Q}) = nI_3\,,$$

Moreover, $\mathbb{E}_{h\sim\mathbb{Q}}\left[\bar{V}_n(h)\right]$ is bounded by

$$\sum_{i=1}^{n}\mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{\pi_{\mathbb{Q}}(a|x_i)}{\pi_0(a|x_i)^{2\alpha}}\right]+\frac{\pi_{\mathbb{Q}}(a_i|x_i)}{\pi_0(a_i|x_i)^{2\alpha}}c_i^2\,.$$

Thus with probability at least $1-\frac{\delta}{2}$, it holds for any $\mathbb{Q}$ that

$$|I_3|\leq\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P})+\log\frac{4}{\delta}}{n\lambda}+\frac{\lambda}{2n}\sum_{i=1}^{n}\frac{\pi_{\mathbb{Q}}(a_i|x_i)}{\pi_0(a_i|x_i)^{2\alpha}}c_i^2$$
$$+\frac{\lambda}{2n}\sum_{i=1}^{n}\mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{\pi_{\mathbb{Q}}(a|x_i)}{\pi_0(a|x_i)^{2\alpha}}\right]\,.$$

Our result is obtained by bounding $|I_1|+|I_2|+|I_3|$. One shortcoming of our analysis is that $\bar{V}_n^{\alpha}(\pi_{\mathbb{Q}})$ is not exactly and only resembles the sum of the theoretical and empirical second moments of our estimator. Precisely, the terms $\pi_{\mathbb{Q}}/\pi_0^{2\alpha}$ should be $\pi_{\mathbb{Q}}^2/\pi_0^{2\alpha}$. This problem arises due to our definition of the martingale difference sequences $(f_i(a_i,h))_{i\in[n]}$ in (13). Precisely, in our proof, we compute the square $f_i(a_i,h)^2$. However, the square of an indicator function is the indicator function itself. Thus applying the expectation afterwards, $\mathbb{E}_{h\sim\mathbb{Q}}\left[f_i(a_i,h)^2\right]$, leads to $\pi_{\mathbb{Q}}$ appearing instead of $\pi_{\mathbb{Q}}^2$. This issue is inherent in the PAC-Bayes formulation and seminal works (London & Sandler, 2019; Sakhi et al., 2022) would suffer the same issue. Solving this would be beneficial and we leave it to future work.

6. Experiments

We briefly present our experiments. More details and discussions can be found in Appendix D. We consider the standard supervised-to-bandit conversion (Agarwal et al., 2014) where we transform a supervised training set $\mathcal{S}_n^{\mathrm{TR}}$ to a logged bandit data $\mathcal{D}_n$ as described in Algorithm 1 in Appendix D.1. Here the action space $\mathcal{A}$ is the label set and the context space $\mathcal{X}$ is the input space. Then, $\mathcal{D}_n$ is used to train our policies. After that, we evaluate the reward of the learned policies on the supervised test set $\mathcal{S}_{n_{\mathrm{TS}}}^{\mathrm{TS}}$ as described in Algorithm 2 in Appendix D.1. Roughly speaking, the resulting reward quantifies the ability of the learned policy to predict the true labels of the inputs in the test set. This is our performance metric; the higher the better. We use 4 image classification datasets `MNIST` (LeCun et al., 1998), `FashionMNIST` (Xiao et al., 2017), `EMNIST` (Cohen et al., 2017) and `CIFAR100` (Krizhevsky et al., 2009).

The logging policy is defined as $\pi_0=\pi_{\eta_0\cdot\mu_0}^{\mathrm{SOF}}$ in (10), where $\mu_0=(\mu_{0,a})_{a\in\mathcal{A}}\in\mathbb{R}^{dK}$ and $\eta_0\in[0,1]$ is the inverse-temperature parameter. The higher η_0, the better the performance of π_0. When $\eta_0=0$, π_0 is uniform. The parameters μ_0 are learned using 5% of the training set $\mathcal{S}_n^{\mathrm{TR}}$. In our experiments, we consider both, Gaussian and mixed-logit policies, in (11) and (12), for which we set the prior as $\mathbb{P}=\mathcal{N}(\eta_0\mu_0,I_{dK})$ and $\mathbb{P}=\mathcal{N}(\eta_0\mu_0,I_{dK})\times\mathrm{G}(0,1)^K$, respectively. Given that μ_0 are learnt on 5% of $\mathcal{S}_n^{\mathrm{TR}}$, we train our policies on the remaining 95% portion of $\mathcal{S}_n^{\mathrm{TR}}$ to match our theory that requires the prior to not depend on training data. The policies are trained using Adam (Kingma & Ba, 2014) with a learning rate of 0.1 for 20 epochs.

We compare our bound to those in London & Sandler (2019); Sakhi et al. (2022); discarding the intractable bound in Swaminathan & Joachims (2015a) as it requires computing a covering number. Here we do not include the learning principles in Swaminathan & Joachims (2015a); London & Sandler (2019) since we directly optimize our bounds. But we make such a comparison in Appendix D.5 for completeness, showing the favorable performance of our bound and the newly proposed learning principle in (16). Also, we do not compare to Su et al. (2020); Metelli et al. (2021) since they do not provide generalization guarantees; they focus on OPE and only propose a heuristic for OPL. However, we still show the favorable performance of our approach in OPL compared to Su et al. (2020); Metelli et al. (2021) in Appendix D.6 for completeness.

Prior methods are not named. Thus we refer to them as **(Author, Policy)** where **Author** $\in$ {**Ours, London et al., Sakhi et al. 1, Sakhi et al. 2**} **and Policy** $\in$ {**Gaussian, Mixed-Logit**}. Here **Ours**, **London et al.**, **Sakhi et al. 1** and **Sakhi et al. 2** correspond to Theorem 4.1, London & Sandler (2019, Theorem 1), Sakhi et al. (2022, Proposition 1), and Sakhi et al. (2022, Proposition 3), respectively. Since we have two classes of policies, each bound leads to two baselines. For example, London & Sandler (2019, Theorem 1) leads to **(London et al., Gaussian)** and **(London et al., Mixed-Logit)**. More details are provided in Appendix D.3.

In Figure 2, we report the reward of the learned policies. Here we fix $\tau=1/\sqrt[4]{n}\approx 0.06$ and $\alpha=1-1/\sqrt[4]{n}\approx 0.94$ so that when n is large enough, both $\hat{R}_n^{\tau}(\pi)$ and $\hat{R}_n^{\alpha}(\pi)$ approach $\hat{R}_n^{\mathrm{IPS}}(\pi)$ (Ionides, 2008). This is because standard IPS should be preferred when $n\to\infty$. To have a fair comparison, we fixed α instead of tuning it in an adaptive fashion as described in Section 4.4. However, we also provide the results with an adaptive α in Figure 3. Let us start with interpreting Figure 2 (with fixed α and τ). Overall, our method outperforms all the baselines. We also observe that Gaussian policies behave better than mixed-logit policies. However, this is less significant for our method where the performances of both Gaussian and mixed-logit policies are comparable. Moreover, our method reaches the maximum reward even when the logging policy has an average performance. In contrast, the baselines only reach their best reward when the logging policy is well-performing ($\eta_0\approx 1$), in which case minor to no improvements are made. Finally, the baselines induce a better reward when the logging policy is uniform ($\eta_0=0$). But our method has a better reward when $\eta_0>0$, which is more common in practice.

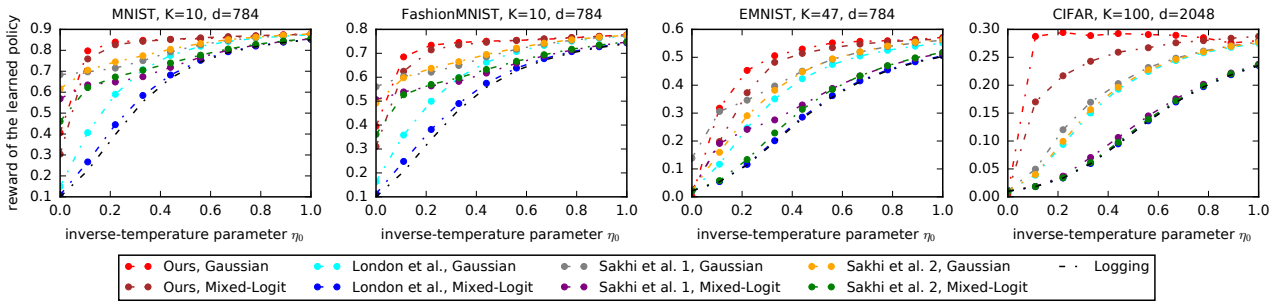

Figure 2. The reward of the learned policy using one of the baselines with varying quality of the logging policy $\eta_0 \in [0, 1]$.

Our choice of τ and α does not affect the above conclusions. In Figure 3 (left-hand side), we compare our method with the best baseline, **(Sakhi et al. 2)** with Gaussian policies, for 20 evenly spaced values of $\tau \in (0, 1)$ and $\alpha \in (0, 1)$. We also include the results using the adaptive tuning procedure of α described in Section 4.4 (green curve). This procedure is reliable since the performance with an adaptive α (green curve) is comparable with the best possible choice of α. Also, our method consistently outperforms the best baseline **(Sakhi et al. 2)** with the best value of τ when the logging policy is not uniform ($\eta_0 > 0$). Also, there is no very bad choice of α, in contrast with $\tau = 10^{-5}$ (dark blue plot) which led to minimal improvement upon all logging policies. This might be due to the $1/\tau$ dependency in existing bounds.

To see the effect of α, we consider the following experiment. We split the logging policies into two groups. The first is called *modest logging* which corresponds to logging policies π_0 whose η_0 is between 0 and 0.5. This group includes the uniform policy and other average-performing policies. The second is called *good logging* and it includes the logging policies whose η_0 is between 0.5 and 1. Then, for each α, we compute the average reward of the learned policy, with that value of α, across these two groups. This leads to the two red and green curves in Figure 3 (right-hand side). Overall, we observe that $\alpha \approx 0.7$ leads to the best performance across the modest logging group. Thus when the performance of the logging policy is bad or average, which is common in practice, regularization can be critical. In contrast, when the performance of the logging policy is already good and n is large enough, regularization might not be needed and $\alpha \approx 1$ would also lead to good performance. This is one of the main strengths of our bound; it holds for the standard IPS recovered with $\alpha = 1$. This result goes against the belief that clipped IPS should always be preferred to standard IPS. Here, our bound applied to standard IPS outperformed clipping by a large margin when the logging policy is relatively well-performing. Similar results for the other datasets are deferred to Appendix D.4.

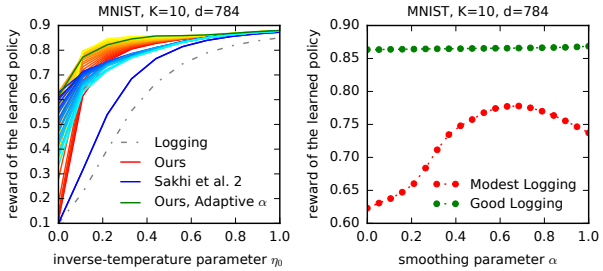

Figure 3. On the left-hand side is the reward of the learned policy with varying $\tau \in (0, 1)$, $\alpha \in (0, 1)$ and $\eta_0 \in [0, 1]$, and for an adaptive α using the procedure in Section 4.4 (green curve). The blue-to-cyan and red-to-yellow colors correspond to varying values of τ and α, respectively. The lighter the color, the higher the value of τ or α. The green curve corresponds to the reward of the learned policy with an adaptive and data-dependent α (Section 4.4). On the right-hand side is the *average* reward of the learned policies using our method across the modest and good logging groups, $\eta_0 \in [0, 0.5]$ (red) and $\eta_0 \in [0.5, 1]$ (green), respectively.

7. Conclusion

In this paper, we investigated a smooth regularization of IPS in the context of OPL. First, we highlighted the pitfalls of hard clipping and advocated for a soft regularization alternative, called exponential smoothing. Moreover, we addressed some fundamental theoretical limitations of existing OPL approaches. Those limitations include the use of one-sided inequalities instead of two-sided ones, the use of learning principles and the use of evaluation bounds in OPL. Building upon this, we successfully derived a *tractable two-sided* PAC-Bayes *generalization bound* for our estimator, which *we directly optimize*. We demonstrated, both theoretically through our bias-variance trade-off analysis in (8) and our bound in Theorem 4.1, and empirically, that this smooth regularization may be critical in some situations. In contrast with all prior works, our bound also applies to the standard IPS. This allowed us to also show that in some other cases, slight to no correction of IPS is needed in OPL.

References

Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L., and Schapire, R. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pp. 1638–1646. PMLR, 2014.

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

Aouali, I., Ivanov, S., Gartrell, M., Rohde, D., Vasile, F., Zaytsev, V., and Legrand, D. Combining reward and rank signals for slate recommendation. *arXiv preprint arXiv:2107.12455*, 2021.

Aouali, I., Hammou, A. A. S., Ivanov, S., Sakhi, O., Rohde, D., and Vasile, F. Probabilistic rank and reward: A scalable model for slate recommendation, 2022.

Aouali, I., Kveton, B., and Katariya, S. Mixed-effect thompson sampling. In *International Conference on Artificial Intelligence and Statistics*, pp. 2087–2115. PMLR, 2023.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.

Bang, H. and Robins, J. M. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61 (4):962–973, 2005.

Bercu, B. and Touati, A. Exponential inequalities for self-normalized martingales with applications. 2008.

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

Chu, W., Li, L., Reyzin, L., and Schapire, R. Contextual bandits with linear payoff functions. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pp. 208–214, 2011.

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

Dudík, M., Langford, J., and Li, L. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pp. 1097–1104, 2011.

Dudík, M., Erhan, D., Langford, J., and Li, L. Sample-efficient nonstationary policy evaluation for contextual bandits. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, UAI'12, pp. 247–254, Arlington, Virginia, USA, 2012. AUAI Press.

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

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

Farid, A. and Majumdar, A. Generalization bounds for meta-learning via pac-bayes and uniform stability. *Advances in Neural Information Processing Systems*, 34:2173–2186, 2021.

Faury, L., Tanielian, U., Dohmatob, E., Smirnova, E., and Vasile, F. Distributionally robust counterfactual risk minimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 3850–3857, 2020.

Gilotte, A., Calauzènes, C., Nedelec, T., Abraham, A., and Dollé, S. Offline a/b testing for recommender systems. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pp. 198–206, 2018.

Guedj, B. A primer on pac-bayesian learning. *arXiv preprint arXiv:1901.05353*, 2019.

Haddouche, M. and Guedj, B. Pac-bayes with unbounded losses through supermartingales. *arXiv preprint arXiv:2210.00928*, 2022.

Hong, J., Kveton, B., Katariya, S., Zaheer, M., and Ghavamzadeh, M. Deep hierarchy in bandits. *arXiv preprint arXiv:2202.01454*, 2022.

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

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

Jeunen, O. and Goethals, B. Pessimistic reward models for off-policy learning in recommendation. In *Fifteenth ACM Conference on Recommender Systems*, pp. 63–74, 2021.

Kallus, N., Saito, Y., and Uehara, M. Optimal off-policy evaluation from multiple logging policies. In *International Conference on Machine Learning*, pp. 5247–5256. PMLR, 2021.

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

Kingma, D. P., Salimans, T., and Welling, M. Variational dropout and the local reparameterization trick. *Advances in neural information processing systems*, 28, 2015.

Korba, A. and Portier, F. Adaptive importance sampling meets mirror descent: a bias-variance tradeoff. In *International Conference on Artificial Intelligence and Statistics*, pp. 11503–11527. PMLR, 2022.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Kuzborskij, I. and Szepesvári, C. Efron-stein pac-bayesian inequalities. *arXiv preprint arXiv:1909.01931*, 2019.

Kuzborskij, I., Vernade, C., Gyorgy, A., and Szepesvári, C. Confident off-policy evaluation and selection through self-normalized importance weighting. In *International Conference on Artificial Intelligence and Statistics*, pp. 640–648. PMLR, 2021.

Lattimore, T. and Szepesvari, C. *Bandit Algorithms*. Cambridge University Press, 2019.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Li, L., Chu, W., Langford, J., and Schapire, R. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, 2010.

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

Luce, R. D. *Individual choice behavior: A theoretical analysis*. Courier Corporation, 2012.

Maddison, C. J., Tarlow, D., and Minka, T. A* sampling. *Advances in neural information processing systems*, 27, 2014.

Maurer, A. and Pontil, M. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

McAllester, D. A. Some pac-bayesian theorems. In *Proceedings of the eleventh annual conference on Computational learning theory*, pp. 230–234, 1998.

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

Papini, M., Metelli, A. M., Lupo, L., and Restelli, M. Optimistic policy optimization via multiple importance sampling. In *International Conference on Machine Learning*, pp. 4989–4999. PMLR, 2019.

Robbins, H. and Monro, S. A stochastic approximation method. *The annals of mathematical statistics*, pp. 400–407, 1951.

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

Russo, D., Van Roy, B., Kazerouni, A., Osband, I., and Wen, Z. A tutorial on Thompson sampling. *Foundations and Trends in Machine Learning*, 11(1):1–96, 2018.

Sachdeva, N., Su, Y., and Joachims, T. Off-policy bandits with deficient support. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 965–975, 2020.

Saito, Y. and Joachims, T. Off-policy evaluation for large action spaces via embeddings. *arXiv preprint arXiv:2202.06317*, 2022.

Sakhi, O., Bonner, S., Rohde, D., and Vasile, F. Blob: A probabilistic model for recommendation that combines organic and bandit signals. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 783–793, 2020.

Sakhi, O., Chopin, N., and Alquier, P. Pac-bayesian offline contextual bandits with guarantees. *arXiv preprint arXiv:2210.13132*, 2022.

Su, Y., Wang, L., Santacatterina, M., and Joachims, T. Cab: Continuous adaptive blending for policy evaluation and learning. In *International Conference on Machine Learning*, pp. 6005–6014. PMLR, 2019.

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

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

Swaminathan, A. and Joachims, T. The self-normalized estimator for counterfactual learning. *advances in neural information processing systems*, 28, 2015b.

Swaminathan, A., Krishnamurthy, A., Agarwal, A., Dudik, M., Langford, J., Jose, D., and Zitouni, I. Off-policy

evaluation for slate recommendation. *Advances in Neural Information Processing Systems*, 30, 2017.

Thompson, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Van Erven, T. and Harremos, P. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

Wang, Y.-X., Agarwal, A., and Dudık, M. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, pp. 3589–3597. PMLR, 2017.

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

Zenati, H., Bietti, A., Martin, M., Diemert, E., and Mairal, J. Counterfactual learning of continuous stochastic policies. 2020.

Zhu, Y., Foster, D. J., Langford, J., and Mineiro, P. Contextual bandits with large action spaces: Made practical. In *International Conference on Machine Learning*, pp. 27428–27453. PMLR, 2022.

Zong, S., Ni, H., Sung, K., Ke, N. R., Wen, Z., and Kveton, B. Cascading bandits for large-scale recommendation problems. *arXiv preprint arXiv:1603.05359*, 2016.

Organization of the Supplementary Material

The supplementary material is organized as follows.

- In **Appendix A**, we give a detailed comparative review of the literature of OPE and OPL.
- In **Appendix B**, we provide some results and proofs for the bias and variance trade-off for our estimators.
- In **Appendix C**, we prove Theorem 4.1. We also provide the proofs for all the claims made in Section 4.
- In **Appendix D**, we present in detail our experimental setup for reproducibility. This appendix also includes additional experiments.

A. Related Work

A contextual bandit (Lattimore & Szepesvari, 2019) is a popular and practical framework for online learning to act under uncertainty (Li et al., 2010; Chu et al., 2011). In practice, the action space is large and short-term gains are important. Thus the agent should be *risk-averse* which goes against the core principle of online algorithms that seek to explore the action space for the sake of long-term gains (Auer et al., 2002; Thompson, 1933; Russo et al., 2018). Although some practical algorithms have been proposed to efficiently explore the action space of a contextual bandit (Zong et al., 2016; Hong et al., 2022; Zhu et al., 2022; Aouali et al., 2023). A clear need remains for an offline procedure that allows optimizing decision-making using offline data. Fortunately, we have access to logged data about previous interactions. The agent can leverage such data to learn an improved policy *offline* (Swaminathan & Joachims, 2015a; London & Sandler, 2019; Sakhi et al., 2022) and consequently enhance the performance of the current system. In this work, we are concerned with this offline, or *off-policy*, formulation of contextual bandits (Dudík et al., 2011; 2012; Dudik et al., 2014; Wang et al., 2017; Farajtabar et al., 2018). Before learning an improved policy, an important intermediary step is to estimate the performance of policies using logged data, as if they were evaluated online. This task is referred to as *off-policy evaluation (OPE)* (Dudík et al., 2011). After that, the resulting estimator is optimized to approximate the optimal policy, and this is referred to as *off-policy learning (OPL)* (Swaminathan & Joachims, 2015a). Next, we review both OPE and OPL approaches.

A.1. Off-Policy Evaluation

Off-policy evaluation in contextual bandits has seen a lot of interest these recent years (Dudík et al., 2011; 2012; Dudik et al., 2014; Wang et al., 2017; Farajtabar et al., 2018; Su et al., 2019; 2020; Kallus et al., 2021; Metelli et al., 2021; Kuzborskij et al., 2021; Saito & Joachims, 2022; Sakhi et al., 2020; Jeunen & Goethals, 2021). We can distinguish between three main families of approaches in the literature. First, *direct method (DM)* (Jeunen & Goethals, 2021) learns a model that approximates the expected cost and then uses it to estimate the performance of evaluated policies. Unfortunately, DM can suffer from modeling bias and misspecification. Thus DM is often designed for specific use cases, in particular large-scale recommender systems (Sakhi et al., 2020; Jeunen & Goethals, 2021; Aouali et al., 2021; 2022). Second, *inverse propensity scoring (IPS)* (Horvitz & Thompson, 1952; Dudík et al., 2012) estimates the cost of the evaluated policies by removing the preference bias of the logging policy in logged data. Under the assumption that the evaluation policy is absolutely continuous with respect to the logging policy, IPS is unbiased, but it can suffer high variance. Note that it can also be highly biased when such an assumption is violated (Sachdeva et al., 2020). The variance issue is acknowledged and some fixes were proposed. For instance, clipping the importance weights (Ionides, 2008; Swaminathan & Joachims, 2015a), self normalizing them (Swaminathan & Joachims, 2015b), etc. (see Gilotte et al. (2018) for a survey). Third, *doubly robust (DR)* (Robins & Rotnitzky, 1995; Bang & Robins, 2005; Dudík et al., 2011; Dudik et al., 2014; Farajtabar et al., 2018) is a combination of DM and IPS. Here a model of the expected cost is used as a control variate for IPS to reduce the variance. Finally, the accuracy of an estimator $\hat{R}_n(\pi)$ in OPE is assessed using the mean squared error (MSE) defined as

$$\begin{aligned}\mathrm{MSE}(\hat{R}_n(\pi)) &= \mathbb{E}\big[\big(\hat{R}_n(\pi) - R(\pi)\big)^2\big] \\ &= \mathbb{B}(\hat{R}_n(\pi))^2 + \mathbb{V}\big[\hat{R}_n(\pi)\big]\,,\end{aligned}$$

where $\mathbb{B}(\hat{R}_n(\pi)) = \mathbb{E}_{\mathcal{D}_n}[\hat{R}_n(\pi)] - R(\pi)$ and $\mathbb{V}[\hat{R}_n(\pi)] = \mathbb{E}_{\mathcal{D}_n}[(\hat{R}_n(\pi) - \mathbb{E}_{\mathcal{D}_n}[\hat{R}_n(\pi)])^2]$ are respectively the bias and the variance of the estimator. It may be relevant to note that Metelli et al. (2021) argued that high-probability concentration rates should be preferred over the MSE to evaluate OPE estimators as they provide non-asymptotic guarantees. In this work, we highlighted the effect of α and β in OPE following the common methodology of using the MSE as a performance metric. However, we also derived two-sided high-probability generalization bounds that attest to the quality of our estimator.

A.2. Off-Policy Learning

Previous works focused on deriving learning principles inspired by generalization bounds. First, Swaminathan & Joachims (2015a) derived a generalization bound for the `IPS-min` estimator in (6) of the form

$$R(\pi) \leq \tilde{R}_n^M(\pi) + \mathcal{O}\left(\sqrt{\frac{\hat{V}_n(\pi)\mathcal{C}_n(\Pi,\delta)}{n}} + M\frac{\mathcal{C}_n(\Pi,\delta)}{n}\right), \tag{18}$$

where $\mathcal{C}_n(\Pi,\delta)$ is the complexity measure of the class of learning policies Π while $\hat{V}_n(\pi)$ is the empirical variance of the estimator on the logged data $\mathcal{D}_n$. The term $\mathcal{C}_n(\Pi,\delta)$ is not necessarily tractable. Thus the generalization bound above was only used to inspire the following learning principle

$$\min_{\mu} \tilde{R}_n^M(\pi_\mu) + \lambda\sqrt{\frac{\hat{V}_n(\pi_\mu)}{n}}, \tag{19}$$

where λ is a tunable hyper-parameter. This learning principle favors policies that simultaneously enjoy low estimated cost and empirical variance. Faury et al. (2020) generalized their work using distributional robustness optimization while Zenati et al. (2020) generalized it to continuous action spaces. The latter also proposed a soft clipping scheme but they derived a generalization bound similar to the one in Swaminathan & Joachims (2015a). Hence they also used the learning principle in (19). Our paper improves upon these work in different ways. First, (18) has a multiplicative dependency on M. Therefore, it is not applicable to standard IPS recovered for $M \to \infty$. In contrast, our bound in Theorem 4.1 does not have a similar dependency on α and thus it also provides generalization guarantees for standard IPS without assuming that the importance weights are bounded. Second, the complexity measure $\mathcal{C}_n(\Pi,\delta)$ is often hard to compute while our bound is tractable and the KL terms can be computed or bounded in closed-form for Gaussian and mixed-logit policies. Third, our bound is differentiable and scalable while the learning principle in (19) requires additional care in optimization (Swaminathan & Joachims, 2015a). Fourth, it is challenging to tune λ in (19) using a procedure that is aligned with online metrics. Finally, we follow the theoretically grounded approach where we optimize our bound directly instead of using a learning principle. This direct optimization of the bound does not require any additional hyper-parameters tuning.

Recently, London & Sandler (2019) elegantly made the connection between PAC-Bayes theory and OPL. As a result, they derived a generalization bound for `IPS-max` in (6) which roughly has the following form

$$R(\pi_\mu) \leq \hat{R}_n^\tau(\pi_\mu) + \mathcal{O}\left(\sqrt{\frac{(\hat{R}_n^\tau(\pi_\mu) + \frac{1}{\tau})\|\mu-\mu_0\|^2}{\tau n}} + \frac{\|\mu-\mu_0\|^2}{\tau n}\right). \tag{20}$$

Again, this bound was used to inspire a novel learning principle in the form

$$\min_{\mu} \hat{R}_n^\tau(\pi_\mu) + \lambda\|\mu-\mu_0\|^2, \tag{21}$$

where λ is a tunable hyper-parameter and $\mu_0 \in \mathbb{R}^{dK}$ is the parameter of the logging policy. This principle favors policies with low estimated cost and whose parameter is not far from that of the logging policy in terms of L_2 distance. While the bound of London & Sandler (2019) is tractable, it still has a multiplicative dependency on $1/\tau$. This makes it inapplicable to standard IPS recovered for $\tau = 0$. It is also not suitable for stochastic first-order optimization (Robbins & Monro, 1951) since the data-dependent quantity $\hat{R}_n^\tau(\pi_\mu)$ is inside a square root. Moreover, optimizing directly their bound leads to minimal improvements over the logging policy in practice. In their work, they used the learning principle in (21) instead which suffers the same issues that we discussed before for Swaminathan & Joachims (2015a), except that it is scalable. Recently, Sakhi et al. (2022) derived novel generalization bounds for a doubly robust version of the `IPS-max` estimator in (6). Sakhi et al. (2022) optimized the theoretical bound directly instead of using some form of learning principle and they showed favorable performance over existing methods. Unfortunately, their bounds have the same multiplicative dependency on $1/\tau$ which makes them vacuous for small values of τ and inapplicable to standard IPS. Moreover, we derive two-sided generalization bounds while all these works only derived one-sided generalization bounds. Unfortunately, the latter does not provide any guarantees on the expected performance of the learned policy. Also, we propose a different estimator to the clipped IPS traditionally used for OPL and we demonstrate empirically that it has better performance.

B. Bias and Variance Trade-Off

In this section, we provide additional results on how β and α control the bias and variance of $\tilde{R}_n^\beta(\cdot)$ and $\hat{R}_n^\alpha(\cdot)$, respectively. Precisely, in Propositions B.1 and B.2 we upper bound the absolute bias and variance of $\hat{R}_n^\alpha(\cdot)$ and $\tilde{R}_n^\beta(\cdot)$, respectively.

B.1. Bias and variance of IPS-α

The following proposition states the bias-variance trade-off for $\hat{R}_n^\alpha(\cdot)$.

Proposition B.1 (Bias and variance of IPS-α)**.** *Let $\alpha \in [0,1]$, the following holds for any evaluation policy $\pi \in \Pi$ that is absolutely continuous with respect to π_0*

$$|\mathbb{B}(\hat{R}_n^\alpha(\pi))| \leq \mathbb{E}_{x\sim\nu, a\sim\pi(\cdot|x)}\left[1-\pi_0(a|x)^{1-\alpha}\right],$$
$$\mathbb{V}\left[\hat{R}_n^\alpha(\pi)\right] \leq \frac{1}{n}\mathbb{E}_{x\sim\nu, a\sim\pi(\cdot|x)}\left[\frac{\pi(a|x)}{\pi_0(a|x)^{2\alpha-1}}\right].$$

Proof. We first bound the bias as

$$\begin{aligned}
\mathbb{B}(\hat{R}_n^\alpha(\pi)) &= \mathbb{E}\left[\hat{R}_n^\alpha(\pi)\right] - R(\pi), \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{x_i\sim\nu, a_i\sim\pi_0(\cdot|x_i), c_i\sim p(\cdot|x_i,a_i)}\left[c_i\frac{\pi(a_i|x_i)}{\pi_0(a_i|x_i)^\alpha}\right] - R(\pi), \\
&\overset{(i)}{=} \mathbb{E}_{(x,a,c)\sim\mu_{\pi_0}}\left[c\frac{\pi(a|x)}{\pi_0(a|x)^\alpha}\right] - R(\pi), \\
&= \mathbb{E}_{x\sim\nu}\left[\sum_{a\in\mathcal{A}} c(x,a)\frac{\pi(a|x)}{\pi_0(a|x)^{\alpha-1}}\right] - \mathbb{E}_{x\sim\nu}\left[\sum_{a\in\mathcal{A}} c(x,a)\pi(a|x)\right], \\
&= \mathbb{E}_{x\sim\nu}\left[\sum_{a\in\mathcal{A}} c(x,a)\pi(a|x)(\pi_0(a|x)^{1-\alpha}-1)\right], \\
&= \mathbb{E}_{x\sim\nu, a\sim\pi(\cdot|x)}\left[c(x,a)(\pi_0(a|x)^{1-\alpha}-1)\right],
\end{aligned}$$

where (i) follows from the i.i.d. assumption. Since $\pi_0(a|x)^{1-\alpha} \leq 1$ for any $x \in \mathcal{X}$ and $a \in \mathcal{A}$, we have that

$$\begin{aligned}
|\mathbb{B}(\hat{R}_n^\alpha(\pi))| &\leq \mathbb{E}_{x\sim\nu, a\sim\pi(\cdot|x)}\left[|c(x,a)||\pi_0(a|x)^{1-\alpha}-1|\right], \\
&\leq \mathbb{E}_{x\sim\nu, a\sim\pi(\cdot|x)}\left[1-\pi_0(a|x)^{1-\alpha}\right].
\end{aligned}$$

The variance is bounded as

$$\begin{aligned}
\mathbb{V}\left[\tilde{R}_n^\beta(\pi)\right] &= \frac{1}{n^2}\sum_{i=1}^n \mathbb{V}_{x_i\sim\nu, a_i\sim\pi_0(\cdot|x_i), c_i\sim p(\cdot|x_i,a_i)}\left[c_i\frac{\pi(a_i|x_i)}{\pi_0(a_i|x_i)^\alpha}\right], \\
&= \frac{1}{n}\mathbb{V}_{(x,a,c)\sim\mu_{\pi_0}}\left[c\frac{\pi(a|x)}{\pi_0(a|x)^\alpha}\right], \\
&\leq \frac{1}{n}\mathbb{E}_{(x,a,c)\sim\mu_{\pi_0}}\left[c^2\frac{\pi(a|x)^2}{\pi_0(a|x)^{2\alpha}}\right], \\
&\leq \frac{1}{n}\mathbb{E}_{x\sim\nu, a\sim\pi_0(\cdot|x)}\left[\frac{\pi(a|x)^2}{\pi_0(a|x)^{2\alpha}}\right], \\
&= \frac{1}{n}\mathbb{E}_{x\sim\nu}\left[\sum_{a\in\mathcal{A}}\frac{\pi(a|x)^2}{\pi_0(a|x)^{2\alpha-1}}\right], \\
&= \frac{1}{n}\mathbb{E}_{x\sim\nu, a\sim\pi(\cdot|x)}\left[\frac{\pi(a|x)}{\pi_0(a|x)^{2\alpha-1}}\right].
\end{aligned}$$

□

B.2. Bias and variance of IPS-β

The following proposition states the bias-variance trade-off for $\tilde{R}_n^\beta(\cdot)$.

Proposition B.2 (Bias and variance of IPS-β)**.** *Let $\beta \in [0,1]$, the following holds for any evaluation policy $\pi \in \Pi$ that is absolutely continuous with respect to π_0*

$$|\mathbb{B}(\tilde{R}_n^\beta(\pi))| \leq \mathbb{E}_{x\sim\nu, a\sim\pi(\cdot|x)}\big[\big|\big(\frac{\pi(a|x)}{\pi_0(a|x)}\big)^{\beta-1} - 1\big|\big]\,,$$

$$\mathbb{V}\left[\tilde{R}_n^\beta(\pi)\right] \leq \frac{1}{n}\mathbb{E}_{x\sim\nu, a\sim\pi(\cdot|x)}\big[\big(\frac{\pi(a|x)}{\pi_0(a|x)}\big)^{2\beta-1}\big]\,.$$

Proof. We first bound the bias as

$$\begin{aligned}
\mathbb{B}(\tilde{R}_n^\beta(\pi)) &= \mathbb{E}\left[\tilde{R}_n^\beta(\pi)\right] - R(\pi) \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{x_i\sim\nu, a_i\sim\pi_0(\cdot|x_i), c_i\sim p(\cdot|x_i,a_i)}\left[c_i\Big(\frac{\pi(a_i|x_i)}{\pi_0(a_i|x_i)}\Big)^\beta\right] - R(\pi)\,, \\
&\overset{(i)}{=} \mathbb{E}_{(x,a,c)\sim\mu_{\pi_0}}\left[c\Big(\frac{\pi(a|x)}{\pi_0(a|x)}\Big)^\beta\right] - R(\pi)\,, \\
&= \mathbb{E}_{x\sim\nu, a\sim\pi_0(\cdot|x)}\left[c(x,a)\Big(\frac{\pi(a|x)}{\pi_0(a|x)}\Big)^\beta\right] - \mathbb{E}_{x\sim\nu, a\sim\pi(\cdot|x)}\left[c(x,a)\right]\,, \\
&= \mathbb{E}_{x\sim\nu}\left[\sum_{a\in\mathcal{A}} c(x,a)\frac{\pi(a|x)^\beta}{\pi_0(a|x)^{\beta-1}}\right] - \mathbb{E}_{x\sim\nu}\left[\sum_{a\in\mathcal{A}} c(x,a)\pi(a|x)\right]\,, \\
&= \mathbb{E}_{x\sim\nu}\left[\sum_{a\in\mathcal{A}} c(x,a)\pi(a|x)\Big(\Big(\frac{\pi_0(a|x)}{\pi(a|x)}\Big)^{1-\beta} - 1\Big)\right]\,,
\end{aligned}$$

where (i) follows from the i.i.d. assumption. It follows that

$$\begin{aligned}
|\mathbb{B}(\tilde{R}_n^\beta(\pi))| &\leq \mathbb{E}_{x\sim\nu}\left[\sum_{a\in\mathcal{A}} |c(x,a)|\pi(a|x)\Big|\Big(\frac{\pi_0(a|x)}{\pi(a|x)}\Big)^{1-\beta} - 1\Big|\right]\,, \\
&\leq \mathbb{E}_{x\sim\nu}\left[\sum_{a\in\mathcal{A}} \pi(a|x)\Big|\Big(\frac{\pi_0(a|x)}{\pi(a|x)}\Big)^{1-\beta} - 1\Big|\right]\,, \\
&= \mathbb{E}_{x\sim\nu, a\sim\pi(\cdot|x)}\left[\Big|\Big(\frac{\pi_0(a|x)}{\pi(a|x)}\Big)^{1-\beta} - 1\Big|\right]\,.
\end{aligned}$$

The variance is bounded as

$$\begin{aligned}
\mathbb{V}\left[\tilde{R}_n^\beta(\pi)\right] &= \frac{1}{n^2}\sum_{i=1}^n \mathbb{V}_{x_i\sim\nu, a_i\sim\pi_0(\cdot|x_i), c_i\sim p(\cdot|x_i,a_i)}\Big[c_i\Big(\frac{\pi(a_i|x_i)}{\pi_0(a_i|x_i)}\Big)^\beta\Big]\,, \\
&= \frac{1}{n}\mathbb{V}_{(x,a,c)\sim\mu_{\pi_0}}\Big[c\Big(\frac{\pi(a|x)}{\pi_0(a|x)}\Big)^\beta\Big]\,, \\
&\leq \frac{1}{n}\mathbb{E}_{(x,a,c)\sim\mu_{\pi_0}}\Big[c^2\Big(\frac{\pi(a|x)}{\pi_0(a|x)}\Big)^{2\beta}\Big]\,, \\
&\leq \frac{1}{n}\mathbb{E}_{x\sim\nu, a\sim\pi_0(\cdot|x)}\Big[\Big(\frac{\pi(a|x)}{\pi_0(a|x)}\Big)^{2\beta}\Big]\,, \\
&= \frac{1}{n}\mathbb{E}_{x\sim\nu}\Big[\sum_{a\in\mathcal{A}}\frac{\pi(a|x)^{2\beta}}{\pi_0(a|x)^{2\beta-1}}\Big]\,, \\
&= \frac{1}{n}\mathbb{E}_{x\sim\nu, a\sim\pi(\cdot|x)}\Big[\frac{\pi(a|x)^{2\beta-1}}{\pi_0(a|x)^{2\beta-1}}\Big]\,.
\end{aligned}$$

□

B.3. Discussion

Here we show using Propositions B.1 and B.2 how α and β trade the bias and variance of $\hat{R}_n^\alpha(\pi)$ and $\tilde{R}_n^\beta(\pi)$, respectively. Let us start with $\hat{R}_n^\alpha(\pi)$, from Proposition B.1, the bound on the bias is minimized in $\alpha = 1$; in which case it is equal to 0. In contrast, the bound on the variance is minimized in $\alpha = 0$; in which case the variance is bounded by $1/n$. Let α_* be the minimizer of the corresponding bound of the MSE

$$\alpha_* = \operatorname{argmin}_{\alpha\in[0,1]} \mathbb{E}_{x\sim\nu, a\sim\pi(\cdot|x)} \left[1 - \pi_0(a|x)^{1-\alpha}\right]^2 + \mathbb{E}_{x\sim\nu, a\sim\pi(\cdot|x)} \left[\frac{\pi(a|x)}{\pi_0(a|x)^{2\alpha-1}}\right]/n\,.$$

We observe that when the variance is small or n is large enough such that $\mathbb{E}_{x\sim\nu, a\sim\pi(\cdot|x)} \left[\frac{\pi(a|x)}{\pi_0(a|x)^{2\alpha-1}}\right]/n \to 0$, then we have that $\alpha_* \to 1$. Thus it is better to use the standard IPS in this case. Otherwise, we have $\alpha_* \to 0$ and this is when regularization helps; basically when we have few samples or when the evaluation policy induces high variance. This demonstrates that the choice of α matters as it trades the bias and variance of $\hat{R}_n^\alpha$.

Similarly, from Proposition B.2, we define β_* as

$$\beta_* = \operatorname{argmin}_{\beta\in[0,1]} \mathbb{E}_{x\sim\nu, a\sim\pi(\cdot|x)} \left[\left|\left(\frac{\pi(a|x)}{\pi_0(a|x)}\right)^{\beta-1} - 1\right|\right] + \frac{1}{n}\mathbb{E}_{x\sim\nu, a\sim\pi(\cdot|x)} \left[\left(\frac{\pi(a|x)}{\pi_0(a|x)}\right)^{2\beta-1}\right].$$

Again, we observe that if $\mathbb{E}_{x\sim\nu, a\sim\pi(\cdot|x)} \left[\left(\frac{\pi(a|x)}{\pi_0(a|x)}\right)^{2\beta-1}\right]/n \to 0$, then $\beta_* \to 1$; in which case it is better to use standard IPS. Otherwise, we have $\beta_* \to 0$ to regularize the importance weights.

C. Proofs for Off-Policy Learning

In this section, we provide the complete proofs for our OPL results in Section 4. We start with proving Theorem 4.1 in Appendix C.1. We then state the extension of Theorem 4.1 along with its proof in Appendix C.2. After that, in Appendix C.3, we provide the proof for Proposition 5.1. Finally, in Appendix C.4, we discuss in detail and prove our claims regarding the number of samples needed so that the performance of the learned policy is close to that of the optimal policy.

C.1. Proof of Theorem 4.1

In this section, we prove Theorem 4.1.

Proof. First, we decompose the difference $R(\pi_\mathbb{Q}) - \hat{R}_n^\alpha(\pi_\mathbb{Q})$ as

$$R(\pi_\mathbb{Q}) - \hat{R}_n^\alpha(\pi_\mathbb{Q}) = \underbrace{R(\pi_\mathbb{Q}) - \frac{1}{n}\sum_{i=1}^n R(\pi_\mathbb{Q}|x_i)}_{I_1} + \underbrace{\frac{1}{n}\sum_{i=1}^n R(\pi_\mathbb{Q}|x_i) - \frac{1}{n}\sum_{i=1}^n R^\alpha(\pi_\mathbb{Q}|x_i)}_{I_2} + \underbrace{\frac{1}{n}\sum_{i=1}^n R^\alpha(\pi_\mathbb{Q}|x_i) - \hat{R}_n^\alpha(\pi_\mathbb{Q})}_{I_3}\,,$$

where

$$\begin{aligned}
R(\pi_\mathbb{Q}) &= \mathbb{E}_{x\sim\nu\,, a\sim\pi_\mathbb{Q}(\cdot|x)}\left[c(x,a)\right]\,,\\
R(\pi_\mathbb{Q}|x_i) &= \mathbb{E}_{a\sim\pi_\mathbb{Q}(\cdot|x_i)}\left[c(x_i,a)\right]\,,\\
R^\alpha(\pi_\mathbb{Q}|x_i) &= \mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{\pi_\mathbb{Q}(a|x_i)}{\pi_0(a|x_i)^\alpha}c(x_i,a)\right]\,,\\
\hat{R}_n^\alpha(\pi) &= \frac{1}{n}\sum_{i=1}^n \frac{\pi(a_i|x_i)}{\pi_0(a_i|x_i)^\alpha}c_i\,.
\end{aligned}$$

Our goal is to bound $|R(\pi_\mathbb{Q}) - \hat{R}_n^\alpha(\pi_\mathbb{Q})|$ and thus we need to bound $|I_1| + |I_2| + |I_3|$. We start with $|I_1|$, Alquier (2021, Theorem 3.3) yields that following inequality holds with probability at least $1 - \delta/2$ for any distribution $\mathbb{Q}$ on $\mathcal{H}$

$$|I_1| \le \sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{4\sqrt{n}}{\delta}}{2n}}\,. \tag{22}$$

Moreover, $|I_2|$ can be bounded by decomposing it as

$$
\begin{aligned}
|I_2| &= \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{a\sim\pi_{\mathbb{Q}}(\cdot|x_i)} \left[c(x_i,a)\right] - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{a\sim\pi_0(\cdot|x_i)} \left[\frac{\pi_{\mathbb{Q}}(a|x_i)}{\pi_0^{\alpha}(a|x_i)} c(x_i,a)\right] \right| \\
&= \left| \frac{1}{n} \sum_{i=1}^{n} \sum_{a\in\mathcal{A}} \pi_{\mathbb{Q}}(a|x_i) c(x_i,a) - \pi_0(a|x_i) \frac{\pi_{\mathbb{Q}}(a|x_i)}{\pi_0^{\alpha}(a|x_i)} c(x_i,a) \right| \\
&= \left| \frac{1}{n} \sum_{i=1}^{n} \sum_{a\in\mathcal{A}} \Big(\pi_{\mathbb{Q}}(a|x_i) - \frac{\pi_{\mathbb{Q}}(a|x_i)}{\pi_0^{\alpha-1}(a|x_i)}\Big) c(x_i,a) \right| \\
&= \left| \frac{1}{n} \sum_{i=1}^{n} \sum_{a\in\mathcal{A}} \Big(1 - \pi_0^{1-\alpha}(a|x_i)\Big) \pi_{\mathbb{Q}}(a|x_i) c(x_i,a) \right| , \\
&\leq \frac{1}{n} \sum_{i=1}^{n} \sum_{a\in\mathcal{A}} \left|1 - \pi_0^{1-\alpha}(a|x_i)\right| \pi_{\mathbb{Q}}(a|x_i) \left|c(x_i,a)\right| .
\end{aligned}
$$

But $1 - \pi_0^{1-\alpha}(a|x) \geq 0$ and $|c(x,a)| \leq 1$ for any $a \in \mathcal{A}$ and $x \in \mathcal{X}$. Thus

$$
|I_2| \leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{a\sim\pi_{\mathbb{Q}}(\cdot|x_i)} \left[1 - \pi_0^{1-\alpha}(a|x_i)\right] . \tag{23}
$$

Finally, we need to bound the main term $|I_3|$. To achieve this, we borrow the following technical lemma from Haddouche & Guedj (2022). It is slightly different from the one in Haddouche & Guedj (2022); their result holds for any $n \geq 1$ while we state a simpler version where n is fixed in advance.

Lemma C.1. *Let $\mathcal{Z}$ be an instance space and let $S_n = (z_i)_{i\in[n]}$ be an n-sized dataset for some $n \geq 1$. Let $(\mathcal{F}_i)_{i\in\{0\}\cup[n]}$ be a filtration adapted to S_n. Also, let $\mathcal{H}$ be a hypothesis space and $(f_i(S_i,h))_{i\in[n]}$ be a martingale difference sequence for any $h \in \mathcal{H}$, that is for any $i \in [n]$, and $h \in \mathcal{H}$, we have that $\mathbb{E}\left[f_i(S_i,h)\,|\mathcal{F}_{i-1}\right] = 0$. Moreover, for any $h \in \mathcal{H}$, let $M_n(h) = \sum_{i=1}^{n} f_i(S_i,h)$. Then for any fixed prior, $\mathbb{P}$, on $\mathcal{H}$, any $\lambda > 0$, the following holds with probability $1-\delta$ over the sample S_n, simultaneously for any $\mathbb{Q}$, on $\mathcal{H}$*

$$
\left|\mathbb{E}_{h\sim\mathbb{Q}}\left[M_n(h)\right]\right| \leq \frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log(2/\delta)}{\lambda} + \frac{\lambda}{2}\left(\mathbb{E}_{h\sim\mathbb{Q}}\left[\langle M\rangle_n(h) + [M]_n(h)\right]\right) ,
$$

where $\langle M\rangle_n(h) = \sum_{i=1}^{n} \mathbb{E}\left[f_i(S_i,h)^2\,|\mathcal{F}_{i-1}\right]$ and $[M]_n(h) = \sum_{i=1}^{n} f_i(S_i,h)^2$.

To apply Lemma C.1, we need to construct an adequate martingale difference sequence $(f_i(S_i,h))_{i\in[n]}$ for $h \in \mathcal{H}$ that allows us to retrieve $|I_3|$. To achieve this, we define $S_n = (a_i)_{i\in[n]}$ as the set of n taken actions. Also, we let $(\mathcal{F}_i)_{i\in\{0\}\cup[n]}$ be a filtration adapted to S_n. For $h \in \mathcal{H}$, we define $f_i(S_i,h)$ as

$$
f_i(S_i,h) = f_i(a_i,h) = \mathbb{E}_{a\sim\pi_0(\cdot|x_i)} \left[\frac{\mathbb{I}_{\{h(x_i)=a\}}}{\pi_0(a|x_i)^{\alpha}} c(x_i,a)\right] - \frac{\mathbb{I}_{\{h(x_i)=a_i\}}}{\pi_0(a_i|x_i)^{\alpha}} c(x_i,a_i) .
$$

We stress that $f_i(S_i,h)$ only depends on the last action in S_i, a_i, and the predictor h. For this reason, we denote it by $f_i(a_i,h)$. The function f_i is indexed by i since it depends on the fixed i-th context, x_i. The context x_i is fixed and thus randomness only comes from $a_i \sim \pi_0(\cdot|x_i)$. It follows that the expectations are under $a_i \sim \pi_0(\cdot|x_i)$. First, we have that $\mathbb{E}\left[f_i(a_i,h)\,|\mathcal{F}_{i-1}\right] = 0$ for any $i \in [n]$, $h \in \mathcal{H}$. This follows from

$$
\begin{aligned}
\mathbb{E}\left[f_i(a_i,h)\,|\mathcal{F}_{i-1}\right] &= \mathbb{E}_{a_i\sim\pi_0(\cdot|x_i)} \left[f_i(a_i,h)\,\middle|a_1,\ldots,a_{i-1}\right] , \\
&= \mathbb{E}_{a_i\sim\pi_0(\cdot|x_i)} \left[\mathbb{E}_{a\sim\pi_0(\cdot|x_i)} \left[\frac{\mathbb{I}_{\{h(x_i)=a\}}}{\pi_0(a|x_i)^{\alpha}} c(x_i,a)\right] - \frac{\mathbb{I}_{\{h(x_i)=a_i\}}}{\pi_0(a_i|x_i)^{\alpha}} c(x_i,a_i)\,\middle|a_1,\ldots,a_{i-1}\right] , \\
&\overset{(i)}{=} \mathbb{E}_{a\sim\pi_0(\cdot|x_i)} \left[\frac{\mathbb{I}_{\{h(x_i)=a\}}}{\pi_0(a|x_i)^{\alpha}} c(x_i,a)\right] - \mathbb{E}_{a_i\sim\pi_0(\cdot|x_i)} \left[\frac{\mathbb{I}_{\{h(x_i)=a_i\}}}{\pi_0(a_i|x_i)^{\alpha}} c(x_i,a_i)\,\middle|a_1,\ldots,a_{i-1}\right] .
\end{aligned}
$$

In (i) we use the fact that given x_i, $\mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{\mathbb{I}_{\{h(x_i)=a\}}}{\pi_0(a|x_i)^\alpha}c(x_i,a)\right]$ is deterministic. Now a_i does not depend on $a_1,\dots,a_{i-1}$ since logged data is i.d.d. Hence

$$\begin{aligned}\mathbb{E}_{a_i\sim\pi_0(\cdot|x_i)}\left[\frac{\mathbb{I}_{\{h(x_i)=a_i\}}}{\pi_0(a_i|x_i)^\alpha}c(x_i,a_i)\Big|a_1,\dots,a_{i-1}\right] &= \mathbb{E}_{a_i\sim\pi_0(\cdot|x_i)}\left[\frac{\mathbb{I}_{\{h(x_i)=a_i\}}}{\pi_0(a_i|x_i)^\alpha}c(x_i,a_i)\right]\,,\\ &= \mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{\mathbb{I}_{\{h(x_i)=a\}}}{\pi_0(a|x_i)^\alpha}c(x_i,a)\right]\,.\end{aligned}$$

It follows that

$$\begin{aligned}\mathbb{E}\left[f_i\left(a_i,h\right)|\mathcal{F}_{i-1}\right] &= \mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{\mathbb{I}_{\{h(x_i)=a\}}}{\pi_0(a|x_i)^\alpha}c(x_i,a)\right]-\mathbb{E}_{a_i\sim\pi_0(\cdot|x_i)}\left[\frac{\mathbb{I}_{\{h(x_i)=a_i\}}}{\pi_0(a_i|x_i)^\alpha}c(x_i,a_i)\Big|a_1,\dots,a_{i-1}\right]\,,\\ &= \mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{\mathbb{I}_{\{h(x_i)=a\}}}{\pi_0(a|x_i)^\alpha}c(x_i,a)\right]-\mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{\mathbb{I}_{\{h(x_i)=a\}}}{\pi_0(a|x_i)^\alpha}c(x_i,a)\right]\,,\\ &= 0\,.\end{aligned}$$

Therefore, for any $h\in\mathcal{H}$, $(f_i(a_i,h))_{i\in[n]}$ is a martingale difference sequence. Hence we apply Lemma C.1 and obtain that the following inequality holds with probability at least $1-\delta/2$ for any $\mathbb{Q}$ on $\mathcal{H}$

$$|\mathbb{E}_{h\sim\mathbb{Q}}\left[M_n(h)\right]|\leq\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P})+\log(4/\delta)}{\lambda}+\frac{\lambda}{2}\left(\mathbb{E}_{h\sim\mathbb{Q}}\left[\langle M\rangle_n(h)+[M]_n(h)\right]\right)\,, \tag{24}$$

where

$$\begin{aligned}M_n(h) &= \sum_{i=1}^n f_i\left(a_i,h\right)\,,\\ \langle M\rangle_n(h) &= \sum_{i=1}^n\mathbb{E}\left[f_i\left(a_i,h\right)^2|\mathcal{F}_{i-1}\right]\,,\\ [M]_n(h) &= \sum_{i=1}^n f_i\left(a_i,h\right)^2\end{aligned}$$

Now these terms can be decomposed as

$$\begin{aligned}\mathbb{E}_{h\sim\mathbb{Q}}\left[M_n(h)\right] &= \sum_{i=1}^n\mathbb{E}_{h\sim\mathbb{Q}}\left[f_i\left(a_i,h\right)\right]\,,\\ &= \sum_{i=1}^n\mathbb{E}_{h\sim\mathbb{Q}}\left[\mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{\mathbb{I}_{\{h(x_i)=a\}}}{\pi_0(a|x_i)^\alpha}c(x_i,a)\right]-\frac{\mathbb{I}_{\{h(x_i)=a_i\}}}{\pi_0(a_i|x_i)^\alpha}c(x_i,a_i)\right]\,,\\ &\overset{(i)}{=} \sum_{i=1}^n\mathbb{E}_{h\sim\mathbb{Q}}\left[\mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{\mathbb{I}_{\{h(x_i)=a\}}}{\pi_0(a|x_i)^\alpha}c(x_i,a)\right]\right]-\mathbb{E}_{h\sim\mathbb{Q}}\left[\frac{\mathbb{I}_{\{h(x_i)=a_i\}}}{\pi_0(a_i|x_i)^\alpha}c(x_i,a_i)\right]\,,\\ &\overset{(ii)}{=} \sum_{i=1}^n\mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{\mathbb{E}_{h\sim\mathbb{Q}}\left[\mathbb{I}_{\{h(x_i)=a\}}\right]}{\pi_0(a|x_i)^\alpha}c(x_i,a)\right]-\frac{\mathbb{E}_{h\sim\mathbb{Q}}\left[\mathbb{I}_{\{h(x_i)=a_i\}}\right]}{\pi_0(a_i|x_i)^\alpha}c(x_i,a_i)\,,\\ &\overset{(iii)}{=} \sum_{i=1}^n\mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{\pi_{\mathbb{Q}}(a|x_i)}{\pi_0(a|x_i)^\alpha}c(x_i,a)\right]-\sum_{i=1}^n\frac{\pi_{\mathbb{Q}}(a_i|x_i)}{\pi_0(a_i|x_i)^\alpha}c(x_i,a_i)\,,\end{aligned}$$

where we use the linearity of the expectation in both (i) and (ii). In (iii), we use our definition of policies in (9). Therefore, we have that

$$\begin{aligned}\mathbb{E}_{h\sim\mathbb{Q}}\left[M_n(h)\right] &= \sum_{i=1}^n\mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{\pi_{\mathbb{Q}}(a|x_i)}{\pi_0(a|x_i)^\alpha}c(x_i,a)\right]-\sum_{i=1}^n\frac{\pi_{\mathbb{Q}}(a_i|x_i)}{\pi_0(a_i|x_i)^\alpha}c(x_i,a_i)\,,\\ &\overset{(i)}{=} \sum_{i=1}^n R^\alpha(\pi_{\mathbb{Q}}|x_i)-n\hat{R}_n^\alpha(\pi_{\mathbb{Q}})\,,\\ &= nI_3\,,\end{aligned} \tag{25}$$

where we used the fact that $c_i = c(a_i, x_i)$ for any $i \in [n]$ in (i).

Now we focus on the terms $\langle M \rangle_n(h)$ and $[M]_n(h)$. First, we have that

$$\begin{aligned} f_i\left(a_i, h\right)^2 &= \left(\mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{\mathbb{I}_{\{h(x_i)=a\}}}{\pi_0(a|x_i)^\alpha}c(x_i,a)\right] - \frac{\mathbb{I}_{\{h(x_i)=a_i\}}}{\pi_0(a_i|x_i)^\alpha}c(x_i,a_i)\right)^2, \\ &= \mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{\mathbb{I}_{\{h(x_i)=a\}}}{\pi_0(a|x_i)^\alpha}c(x_i,a)\right]^2 + \left(\frac{\mathbb{I}_{\{h(x_i)=a_i\}}}{\pi_0(a_i|x_i)^\alpha}c(x_i,a_i)\right)^2 \\ &\qquad - 2\mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{\mathbb{I}_{\{h(x_i)=a\}}}{\pi_0(a|x_i)^\alpha}c(x_i,a)\right]\frac{\mathbb{I}_{\{h(x_i)=a_i\}}}{\pi_0(a_i|x_i)^\alpha}c(x_i,a_i), \\ &= \mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{\mathbb{I}_{\{h(x_i)=a\}}}{\pi_0(a|x_i)^\alpha}c(x_i,a)\right]^2 + \frac{\mathbb{I}_{\{h(x_i)=a_i\}}}{\pi_0(a_i|x_i)^{2\alpha}}c(x_i,a_i)^2 \\ &\qquad - 2\mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{\mathbb{I}_{\{h(x_i)=a\}}}{\pi_0(a|x_i)^\alpha}c(x_i,a)\right]\frac{\mathbb{I}_{\{h(x_i)=a_i\}}}{\pi_0(a_i|x_i)^\alpha}c(x_i,a_i). \end{aligned} \tag{26}$$

Moreover, $f_i\left(a_i, h\right)^2$ does not depend on $a_1, \ldots, a_{i-1}$. Thus

$$\mathbb{E}\left[f_i\left(a_i,h\right)^2|\mathcal{F}_{i-1}\right] = \mathbb{E}_{a_i\sim\pi_0(\cdot|x_i)}\left[f_i\left(a_i,h\right)^2|\mathcal{F}_{i-1}\right] = \mathbb{E}_{a_i\sim\pi_0(\cdot|x_i)}\left[f_i\left(a_i,h\right)^2\right] = \mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[f_i\left(a,h\right)^2\right].$$

Computing $\mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[f_i\left(a,h\right)^2\right]$ using the decomposition in (26) yields

$$\begin{aligned} \mathbb{E}\left[f_i\left(a_i,h\right)^2|\mathcal{F}_{i-1}\right] &= \mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[f_i\left(a,h\right)^2\right], \\ &= -\mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{\mathbb{I}_{\{h(x_i)=a\}}}{\pi_0(a|x_i)^\alpha}c(x_i,a)\right]^2 + \mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{\mathbb{I}_{\{h(x_i)=a\}}}{\pi_0(a|x_i)^{2\alpha}}c(x_i,a)^2\right] \end{aligned} \tag{27}$$

Combining (26) and (27) leads to

$$\begin{aligned} \mathbb{E}\left[f_i\left(a_i,h\right)^2|\mathcal{F}_{i-1}\right] + f_i\left(a_i,h\right)^2 &= \mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{\mathbb{I}_{\{h(x_i)=a\}}}{\pi_0(a|x_i)^{2\alpha}}c(x_i,a)^2\right] + \frac{\mathbb{I}_{\{h(x_i)=a_i\}}}{\pi_0(a_i|x_i)^{2\alpha}}c(x_i,a_i)^2 \\ &\qquad - 2\mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{\mathbb{I}_{\{h(x_i)=a\}}}{\pi_0(a|x_i)^\alpha}c(x_i,a)\right]\frac{\mathbb{I}_{\{h(x_i)=a_i\}}}{\pi_0(a_i|x_i)^\alpha}c(x_i,a_i), \\ &\overset{(i)}{\leq} \mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{\mathbb{I}_{\{h(x_i)=a\}}}{\pi_0(a|x_i)^{2\alpha}}c(x_i,a)^2\right] + \frac{\mathbb{I}_{\{h(x_i)=a_i\}}}{\pi_0(a_i|x_i)^{2\alpha}}c(x_i,a_i)^2. \end{aligned} \tag{28}$$

The inequality in (i) holds because $-2\mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{\mathbb{I}_{\{h(x_i)=a\}}}{\pi_0(a|x_i)^\alpha}c(x_i,a)\right]\frac{\mathbb{I}_{\{h(x_i)=a_i\}}}{\pi_0(a_i|x_i)^\alpha}c(x_i,a_i) \leq 0$. Therefore, we have that

$$\langle M \rangle_n(h) + [M]_n(h) \leq \sum_{i=1}^n \mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{\mathbb{I}_{\{h(x_i)=a\}}}{\pi_0(a|x_i)^{2\alpha}}c(x_i,a)^2\right] + \frac{\mathbb{I}_{\{h(x_i)=a_i\}}}{\pi_0(a_i|x_i)^{2\alpha}}c(x_i,a_i)^2.$$

Finally, by using the linearity of the expectation and the definition of policies in (9), we get that

$$\begin{aligned} \mathbb{E}_{h\sim\mathbb{Q}}\left[\langle M \rangle_n(h) + [M]_n(h)\right] &\leq \sum_{i=1}^n \mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{\mathbb{E}_{h\sim\mathbb{Q}}\left[\mathbb{I}_{\{h(x_i)=a\}}\right]}{\pi_0(a|x_i)^{2\alpha}}c(x_i,a)^2\right] + \frac{\mathbb{E}_{h\sim\mathbb{Q}}\left[\mathbb{I}_{\{h(x_i)=a_i\}}\right]}{\pi_0(a_i|x_i)^{2\alpha}}c(x_i,a_i)^2, \\ &= \sum_{i=1}^n \mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{\pi_{\mathbb{Q}}(a|x_i)}{\pi_0(a|x_i)^{2\alpha}}c(x_i,a)^2\right] + \frac{\pi_{\mathbb{Q}}(a_i|x_i)}{\pi_0(a_i|x_i)^{2\alpha}}c(x_i,a_i)^2. \end{aligned} \tag{29}$$

Combining (24) and (29) yields

$$\begin{aligned} n|I_3| &= |\sum_{i=1}^n R^\alpha(\pi_{\mathbb{Q}}|x_i) - n\hat{R}_n^\alpha(\pi_{\mathbb{Q}})| \\ &\leq \frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log(4/\delta)}{\lambda} + \frac{\lambda}{2}\sum_{i=1}^n \mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{\pi_{\mathbb{Q}}(a|x_i)}{\pi_0(a|x_i)^{2\alpha}}c(x_i,a)^2\right] + \frac{\pi_{\mathbb{Q}}(a_i|x_i)}{\pi_0(a_i|x_i)^{2\alpha}}c(x_i,a_i)^2. \end{aligned} \tag{30}$$

This means that the following inequality holds with probability at least $1-\delta/2$ for any distribution $\mathbb{Q}$ on $\mathcal{H}$

$$|I_3| \leq \frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log(4/\delta)}{n\lambda} + \frac{\lambda}{2n}\sum_{i=1}^{n}\mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{\pi_{\mathbb{Q}}(a|x_i)}{\pi_0(a|x_i)^{2\alpha}}c(x_i,a)^2\right] + \frac{\lambda}{2n}\sum_{i=1}^{n}\frac{\pi_{\mathbb{Q}}(a_i|x_i)}{\pi_0(a_i|x_i)^{2\alpha}}c(x_i,a_i)^2\,. \tag{31}$$

However we know that $c(x,a)^2 \leq 1$ for any $x \in \mathcal{X}$ and $a \in \mathcal{A}$ and that $c(x_i,a_i) = c_i$ for any $i \in [n]$. Thus the following inequality holds with probability at least $1-\delta/2$ for any distribution $\mathbb{Q}$ on $\mathcal{H}$

$$|I_3| \leq \frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log(4/\delta)}{n\lambda} + \frac{\lambda}{2n}\sum_{i=1}^{n}\mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{\pi_{\mathbb{Q}}(a|x_i)}{\pi_0(a|x_i)^{2\alpha}}\right] + \frac{\lambda}{2n}\sum_{i=1}^{n}\frac{\pi_{\mathbb{Q}}(a_i|x_i)}{\pi_0(a_i|x_i)^{2\alpha}}c_i^2\,. \tag{32}$$

The union bound of (22) and (32) combined with the deterministic result in (23) yields that the following inequality holds with probability at least $1-\delta$ for any distribution $\mathbb{Q}$ on $\mathcal{H}$

$$\begin{aligned}|R(\pi_{\mathbb{Q}}) - \hat{R}_n^{\alpha}(\pi_{\mathbb{Q}})| \leq \sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{4\sqrt{n}}{\delta}}{2n}} + \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{a\sim\pi_{\mathbb{Q}}(\cdot|x_i)}\left[1-\pi_0^{1-\alpha}(a|x_i)\right] + \frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log(4/\delta)}{n\lambda}\\ + \frac{\lambda}{2n}\sum_{i=1}^{n}\mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{\pi_{\mathbb{Q}}(a|x_i)}{\pi_0(a|x_i)^{2\alpha}}\right] + \frac{\lambda}{2n}\sum_{i=1}^{n}\frac{\pi_{\mathbb{Q}}(a_i|x_i)}{\pi_0(a_i|x_i)^{2\alpha}}c_i^2\,.\end{aligned} \tag{33}$$

□

C.2. Extensions of Theorem 4.1

Proposition C.2 (Extension of Theorem 4.1 to hold simultaneously for any $\lambda \in (0,1)$)**.** *Let $n \geq 1$, $\delta \in [0,1]$, $\alpha \in [0,1]$, and let $\mathbb{P}$ be a fixed prior on $\mathcal{H}$, then with probability at least $1-\delta$ over draws $\mathcal{D}_n \sim \mu_{\pi_0}^n$, the following holds simultaneously for any posterior $\mathbb{Q}$ on $\mathcal{H}$, and for any $\lambda \in (0,1)$ that*

$$|R(\pi_{\mathbb{Q}}) - \hat{R}_n^{\alpha}(\pi_{\mathbb{Q}})| \leq \sqrt{\frac{\mathrm{KL}'_1(\pi_{\mathbb{Q}},\lambda)}{2n}} + B_n^{\alpha}(\pi_{\mathbb{Q}}) + \frac{\mathrm{KL}'_2(\pi_{\mathbb{Q}},\lambda)}{n\lambda} + \frac{\lambda}{2}\bar{V}_n^{\alpha}(\pi_{\mathbb{Q}})\,.$$

where

$$\begin{aligned}\mathrm{KL}'_1(\pi_{\mathbb{Q}},\lambda) &= D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{8\sqrt{n}}{\delta\lambda}\,,\\ \mathrm{KL}'_2(\pi_{\mathbb{Q}},\lambda) &= 2\big(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{8}{\delta\lambda}\big)\,,\\ B_n^{\alpha}(\pi_{\mathbb{Q}}) &= 1 - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{a\sim\pi_{\mathbb{Q}}(\cdot|x_i)}\left[\pi_0^{1-\alpha}(a|x_i)\right]\,,\\ \bar{V}_n^{\alpha}(\pi_{\mathbb{Q}}) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{\pi_{\mathbb{Q}}(a|x_i)}{\pi_0(a|x_i)^{2\alpha}}\right] + \frac{\pi_{\mathbb{Q}}(a_i|x_i)}{\pi_0(a_i|x_i)^{2\alpha}}c_i^2.\end{aligned}$$

Proof. Let $\delta \in (0,1)$. For any $i \geq 1$, we define $\lambda_i = 2^{-i}$ and let $\delta_i = \delta\lambda_i$. Then Theorem 4.1 yields that for any $i \geq 1$, the following inequality holds with probability at least $1-\delta_i$ for any $\mathbb{Q}$ on $\mathcal{H}$

$$|R(\pi_{\mathbb{Q}}) - \hat{R}_n^{\alpha}(\pi_{\mathbb{Q}})| \leq \sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{4\sqrt{n}}{\delta_i}}{2n}} + B_n^{\alpha}(\pi_{\mathbb{Q}}) + \frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{4}{\delta_i}}{n\lambda_i} + \frac{\lambda_i}{2}\bar{V}_n^{\alpha}(\pi_{\mathbb{Q}})\,.$$

Now notice that $\sum_{i=1}^{\infty}\lambda_i = 1$, and hence $\sum_{i=1}^{\infty}\delta_i = \delta$. Therefore, the union bound of the above inequalities over $i \geq 1$ yields that with probability at least $1-\delta$, the following inequality holds with probability at least $1-\delta$ for any $\mathbb{Q}$ on $\mathcal{H}$ and for any $i \geq 1$

$$|R(\pi_{\mathbb{Q}}) - \hat{R}_n^{\alpha}(\pi_{\mathbb{Q}})| \leq \sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{4\sqrt{n}}{\delta_i}}{2n}} + B_n^{\alpha}(\pi_{\mathbb{Q}}) + \frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{4}{\delta_i}}{n\lambda_i} + \frac{\lambda_i}{2}\bar{V}_n^{\alpha}(\pi_{\mathbb{Q}})\,. \tag{34}$$

Let $\lceil\cdot\rceil$ denote the ceiling function, then we have that for any $\lambda \in (0,1)$, there exists $j = \lceil \frac{-\log\lambda}{\log 2} \rceil \geq 1$ such that $\lambda/2 \leq \lambda_j \leq \lambda$. Since (34) holds for any $i \geq 1$, it holds in particular for j. In addition to this, we have that $\frac{1}{\lambda_j} \leq \frac{2}{\lambda}$, that $\lambda_j \leq \lambda$ and that $\frac{1}{\delta_j} = \frac{1}{\lambda_j \delta} \leq \frac{2}{\delta\lambda}$. This yields that the following inequality holds with probability at least $1-\delta$ for any $\mathbb{Q}$ on $\mathcal{H}$ and for any $\lambda \in (0,1)$

$$|R(\pi_{\mathbb{Q}}) - \hat{R}_n^{\alpha}(\pi_{\mathbb{Q}})| \leq \sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{8\sqrt{n}}{\delta\lambda}}{2n}} + B_n^{\alpha}(\pi_{\mathbb{Q}}) + 2\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{8}{\delta\lambda}}{n\lambda} + \frac{\lambda}{2}\bar{V}_n^{\alpha}(\pi_{\mathbb{Q}})\,. \tag{35}$$

The additional 2 in $2\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P})+\log\frac{8}{\delta\lambda}}{n\lambda}$ appears since we used that $\frac{1}{\lambda_j} \leq \frac{2}{\lambda}$. Similarly, the additional $\frac{2}{\lambda}$ in the logarithmic terms is due to the fact that $\frac{1}{\delta_j} \leq \frac{2}{\delta\lambda}$. Finally, setting

$$\begin{aligned}
\mathrm{KL}'_1(\pi_{\mathbb{Q}}, \lambda) &= D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{8\sqrt{n}}{\delta\lambda}\,,\\
\mathrm{KL}'_2(\pi_{\mathbb{Q}}, \lambda) &= 2\big(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{8}{\delta\lambda}\big)\,,
\end{aligned}$$

concludes the proof. □

Next, we provide a similar proof to extend Theorem 4.1 to any $\alpha \in (0,1]$. While we only provide a one-sided inequality, the same covering technique can be used to obtain the other side of the inequality.

Proposition C.3 (One-sided extension of Theorem 4.1 to hold simultaneously for any $\alpha \in (0,1) \cup \{1\}$)**.** *Let $n \geq 1$, $\delta \in [0,1]$, $\lambda > 0$, and let $\mathbb{P}$ be a fixed prior on $\mathcal{H}$, then with probability at least $1-\delta$ over draws $\mathcal{D}_n \sim \mu_{\pi_0}^n$, the following holds simultaneously for any posterior $\mathbb{Q}$ on $\mathcal{H}$, and for any $\alpha \in (0,1]$ that*

$$R(\pi_{\mathbb{Q}}) \leq \hat{R}_n^{\alpha}(\pi_{\mathbb{Q}}) + \sqrt{\frac{\mathrm{KL}''_1(\pi_{\mathbb{Q}}, \alpha)}{2n}} + B_n^{\alpha}(\pi_{\mathbb{Q}}) + \frac{\mathrm{KL}''_2(\pi_{\mathbb{Q}}, \alpha)}{n\lambda} + \frac{\lambda}{2}\bar{V}_n^{\alpha}(\pi_{\mathbb{Q}})\,.$$

where

$$\begin{aligned}
\mathrm{KL}''_1(\pi_{\mathbb{Q}}, \alpha) &= D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{8\sqrt{n}}{\delta\alpha}\,,\\
\mathrm{KL}''_2(\pi_{\mathbb{Q}}, \alpha) &= D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{8}{\delta\alpha}\,,\\
B_n^{\alpha}(\pi_{\mathbb{Q}}) &= 1 - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{a\sim\pi_{\mathbb{Q}}(\cdot|x_i)}\left[\pi_0^{1-\alpha}(a|x_i)\right]\,,\\
\bar{V}_n^{\alpha}(\pi_{\mathbb{Q}}) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{\pi_{\mathbb{Q}}(a|x_i)}{\pi_0(a|x_i)^{2\alpha}}\right] + \frac{\pi_{\mathbb{Q}}(a_i|x_i)}{\pi_0(a_i|x_i)^{2\alpha}}c_i^2.
\end{aligned}$$

Proof. Let $\delta \in (0,1)$. For any $i \geq 0$, we define $\alpha_i = 2^{-i}$ and let $\delta_i = \delta\alpha_i/2$. Then Theorem 4.1 yields that for any $i \geq 0$, the following inequality holds with probability at least $1-\delta_i$ for any $\mathbb{Q}$ on $\mathcal{H}$

$$|R(\pi_{\mathbb{Q}}) - \hat{R}_n^{\alpha_i}(\pi_{\mathbb{Q}})| \leq \sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{4\sqrt{n}}{\delta_i}}{2n}} + B_n^{\alpha_i}(\pi_{\mathbb{Q}}) + \frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{4}{\delta_i}}{n\lambda} + \frac{\lambda}{2}\bar{V}_n^{\alpha_i}(\pi_{\mathbb{Q}})\,.$$

Now notice that $\sum_{i=0}^{\infty}\alpha_i = 2$, and hence by definition of δ_i, we have $\sum_{i=0}^{\infty}\delta_i = \delta$. Therefore, the union bound of the above inequalities over $i \geq 0$ yields that with probability at least $1-\delta$, the following inequality holds with probability at least $1-\delta$ for any $\mathbb{Q}$ on $\mathcal{H}$ and for any $i \geq 0$

$$|R(\pi_{\mathbb{Q}}) - \hat{R}_n^{\alpha_i}(\pi_{\mathbb{Q}})| \leq \sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{4\sqrt{n}}{\delta_i}}{2n}} + B_n^{\alpha_i}(\pi_{\mathbb{Q}}) + \frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{4}{\delta_i}}{n\lambda} + \frac{\lambda}{2}\bar{V}_n^{\alpha_i}(\pi_{\mathbb{Q}})\,. \tag{36}$$

Let $\lfloor\cdot\rfloor$ denote the floor function, then we have that for any $\alpha \in (0,1]$, there exists $j = \lfloor \frac{-\log\alpha}{\log 2} \rfloor \geq 0$ such that $\alpha \leq \alpha_j \leq 2\alpha$. Since (34) holds for any $i \geq 0$, it holds in particular for j. In addition, we have that $B_n^{\alpha}(\pi_{\mathbb{Q}})$ and $\hat{R}_n^{\alpha}(\pi_{\mathbb{Q}})$ are

decreasing in α while $\bar{V}_n^{\alpha}(\pi_{\mathbb{Q}})$ is increasing in α. Therefore, we have that $\hat{R}_n^{\alpha_j}(\pi_{\mathbb{Q}}) \leq \hat{R}_n^{\alpha}(\pi_{\mathbb{Q}})$, $B_n^{\alpha_j}(\pi_{\mathbb{Q}}) \leq B_n^{\alpha}(\pi_{\mathbb{Q}})$, and $\bar{V}_n^{\alpha_j}(\pi_{\mathbb{Q}}) \leq \bar{V}_n^{2\alpha}(\pi_{\mathbb{Q}})$. Moreover, we have that $\frac{1}{\delta_j} \leq \frac{2}{\delta\alpha}$. This yields that the following inequality holds with probability at least $1-\delta$ for any $\mathbb{Q}$ on $\mathcal{H}$ and for any $\alpha \in (0,1]$

$$R(\pi_{\mathbb{Q}}) \leq \hat{R}_n^{\alpha}(\pi_{\mathbb{Q}}) + \sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{8\sqrt{n}}{\delta\alpha}}{2n}} + B_n^{\alpha}(\pi_{\mathbb{Q}}) + \frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{8}{\delta\alpha}}{n\lambda} + \frac{\lambda}{2}\bar{V}_n^{2\alpha}(\pi_{\mathbb{Q}}) \,. \tag{37}$$

Finally, setting

$$\mathrm{KL}''_1(\pi_{\mathbb{Q}}, \alpha) = D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{8\sqrt{n}}{\delta\alpha} \,,$$
$$\mathrm{KL}''_2(\pi_{\mathbb{Q}}, \alpha) = D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{8}{\delta\alpha} \,,$$

concludes the proof. □

C.3. Proof of Proposition 5.1

Haddouche & Guedj (2022, Theorem 7) provides an application of Lemma C.1 to the general PAC-Bayes learning problems in Section 4.1. We cannot apply their theorem directly to get Proposition 5.1 for two reasons. They assume that the loss function is non-negative and they derive a one-sided generalization bound. In our case, the loss function is negative and we want to derive a two-sided generalization bound. Fortunately, we show with a slight modification of their proof that the result can be extended to two-sided inequalities with negative losses. In fact, the only requirement is that the sign of loss is fixed. We show next how this is achieved.

Proof. First, note that Lemma C.1 does not make any assumption on the sign of the martingale difference sequence $(f_i(S_i,h))_{i\in[n]}$ nor on the sign of the terms that decompose it. Now similarly to the proof in Appendix C.1, we define $S_n = (x_i, a_i)_{i\in[n]}$ as the set of n observed contexts and taken actions. Also, we let $(\mathcal{F}_i)_{i\in\{0\}\cup[n]}$ be a filtration adapted to S_n. For $h \in \mathcal{H}$, we define $f_i(S_i, h)$ as

$$f_i(S_i,h) = f_i(x_i,a_i,h) = f(x_i,a_i,h) = \mathbb{E}_{x\sim\nu, a\sim\pi_0(\cdot|x)}\left[\frac{\mathbb{I}_{\{h(x)=a\}}}{\pi_0(a|x)^{\alpha}}c(x,a)\right] - \frac{\mathbb{I}_{\{h(x_i)=a_i\}}}{\pi_0(a_i|x_i)^{\alpha}}c(x_i,a_i) \,.$$

Here $f_i(S_i,h)$ only depends on the last samples x_i, a_i and the predictor h. For this reason, we denote it by $f_i(x_i,a_i,h)$. Also, the function f_i does not depend on i and this is why we simplify the notation as $f_i(x_i,a_i,h) = f(x_i,a_i,h)$. Moreover, the randomness in $f(x_i,a_i,h)$ is only due $x_i \sim \nu$ and $a_i \sim \pi_0(\cdot|x_i)$; all other terms are deterministic. Thus the expectations are under $x_i \sim \nu, a_i \sim \pi_0(\cdot|x_i)$. Now similarly to the proof in Appendix C.1, we have that $\mathbb{E}[f(x_i,a_i,h)|\mathcal{F}_{i-1}] = 0$ for any $i \in [n], h \in \mathcal{H}$. Therefore, $(f(x_i,a_i,h))_{i\in[n]}$ is a martingale difference sequence for any $h \in \mathcal{H}$. Thus we apply Lemma C.1 and get that that with probability at least $1-\delta$, the following holds simultaneously for any distribution $\mathbb{Q}$ on $\mathcal{H}$

$$|\mathbb{E}_{h\sim\mathbb{Q}}[M_n(h)]| \leq \frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log(2/\delta)}{\lambda} + \frac{\lambda}{2}\left(\mathbb{E}_{h\sim\mathbb{Q}}\left[\langle M\rangle_n(h) + [M]_n(h)\right]\right) \,, \tag{38}$$

where

$$M_n(h) = \sum_{i=1}^{n} f(x_i,a_i,h) \,,$$
$$\langle M\rangle_n(h) = \sum_{i=1}^{n} \mathbb{E}\left[f(x_i,a_i,h)^2 |\mathcal{F}_{i-1}\right] \,,$$
$$[M]_n(h) = \sum_{i=1}^{n} f(x_i,a_i,h)^2 \,.$$

Now we compute $\mathbb{E}_{h\sim\mathbb{Q}}\left[M_n(h)\right]$ as

$$\begin{aligned}\mathbb{E}_{h\sim\mathbb{Q}}\left[M_n(h)\right] &= \sum_{i=1}^{n}\mathbb{E}_{x\sim\nu,a\sim\pi_0(\cdot|x)}\left[\frac{\pi_{\mathbb{Q}}(a|x)}{\pi_0(a|x)^{\alpha}}c(x,a)\right]-\frac{\pi_{\mathbb{Q}}(a_i|x_i)}{\pi_0(a_i|x_i)^{\alpha}}c(x_i,a_i)\,,\\ &= n\mathbb{E}_{x\sim\nu,a\sim\pi_0(\cdot|x)}\left[\frac{\pi_{\mathbb{Q}}(a|x)}{\pi_0(a|x)^{\alpha}}c(x,a)\right]-\sum_{i=1}^{n}\frac{\pi_{\mathbb{Q}}(a_i|x_i)}{\pi_0(a_i|x_i)^{\alpha}}c(x_i,a_i)\,,\end{aligned}\tag{39}$$

where we used the linearity of the expectation $\mathbb{E}_{h\sim\mathbb{Q}}\left[\cdot\right]$ and the definition of policies in (9). Moreover, similarly to the proof in Appendix C.1, we have that

$$\begin{aligned}\langle M\rangle_n(h)+[M]_n(h) &= \sum_{i=1}^{n}\mathbb{E}\left[f\left(x_i,a_i,h\right)^2|\mathcal{F}_{i-1}\right]+f\left(x_i,a_i,h\right)^2\\ &= \sum_{i=1}^{n}\mathbb{E}_{x\sim\nu,a\sim\pi_0(\cdot|x)}\left[\frac{\mathbb{I}_{\{h(x)=a\}}}{\pi_0(a|x)^{2\alpha}}c(x,a)^2\right]+\frac{\mathbb{I}_{\{h(x_i)=a_i\}}}{\pi_0(a_i|x_i)^{2\alpha}}c(x_i,a_i)^2\\ &\qquad -2\mathbb{E}_{x\sim\nu,a\sim\pi_0(\cdot|x)}\left[\frac{\mathbb{I}_{\{h(x)=a\}}}{\pi_0(a|x)^{\alpha}}c(x,a)\right]\frac{\mathbb{I}_{\{h(x_i)=a_i\}}}{\pi_0(a_i|x_i)^{\alpha}}c(x_i,a_i)\,,\\ &\overset{(i)}{\leq} n\mathbb{E}_{x\sim\nu,a\sim\pi_0(\cdot|x)}\left[\frac{\mathbb{I}_{\{h(x)=a\}}}{\pi_0(a|x)^{2\alpha}}c(x,a)^2\right]+\sum_{i=1}^{n}\frac{\mathbb{I}_{\{h(x_i)=a_i\}}}{\pi_0(a_i|x_i)^{2\alpha}}c(x_i,a_i)^2\,,\end{aligned}\tag{40}$$

where (i) holds since $-2\mathbb{E}_{x\sim\nu,a\sim\pi_0(\cdot|x)}\left[\frac{\mathbb{I}_{\{h(x)=a\}}}{\pi_0(a|x)^{\alpha}}c(x,a)\right]\frac{\mathbb{I}_{\{h(x_i)=a_i\}}}{\pi_0(a_i|x_i)^{\alpha}}c(x_i,a_i)\leq 0$ for any $i\in[n]$. This is where the non-negative loss assumption is not needed. Our loss $L_\alpha(h,x,a,c)=\frac{\mathbb{I}_{\{h(x)=a\}}}{\pi_0(a|x)^{\alpha}}c$ is negative since $c\in[-1,0]$. However, we only need the product between the loss and its expectation to be non-negative. This holds in particular when the loss has a fixed sign. In that case, the expectation of the loss and the loss itself will have the same sign and thus their product will be non-negative. In our case, the loss has a fixed negative sign and this is all we needed. Now notice that

$$\begin{aligned}&n\mathbb{E}_{x\sim\nu,a\sim\pi_0(\cdot|x)}\left[\frac{\pi_{\mathbb{Q}}(a|x)}{\pi_0(a|x)^{\alpha}}c(x,a)\right]=nR^{\alpha}(\pi_{\mathbb{Q}})\,,\\ &\sum_{i=1}^{n}\frac{\pi_{\mathbb{Q}}(a_i|x_i)}{\pi_0(a_i|x_i)^{\alpha}}c(x_i,a_i)=n\hat{R}_n^{\alpha}(\pi_{\mathbb{Q}})\,,\end{aligned}$$

where we used that $c(x_i,a_i)=c_i$ for any $i\in[n]$ in the second equality. Using these two equalities and plugging (39) and (40) in (38) yields that with probability at least $1-\delta$, the following holds simultaneously for any distribution $\mathbb{Q}$ on $\mathcal{H}$

$$\begin{aligned}n\left|R^{\alpha}(\pi_{\mathbb{Q}})-\hat{R}_n^{\alpha}(\pi_{\mathbb{Q}})\right| \leq \frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P})+\log(2/\delta)}{\lambda}+\frac{\lambda}{2}\Big(&n\mathbb{E}_{x\sim\nu,a\sim\pi_0(\cdot|x)}\left[\frac{\pi_{\mathbb{Q}}(a|x)}{\pi_0(a|x)^{2\alpha}}c(x,a)^2\right]\\ &+\sum_{i=1}^{n}\frac{\pi_{\mathbb{Q}}(a_i|x_i)}{\pi_0(a_i|x_i)^{2\alpha}}c(x_i,a_i)^2\Big)\,.\end{aligned}\tag{41}$$

Again we used the linearity of the expectation $\mathbb{E}_{h\sim\mathbb{Q}}\left[\cdot\right]$ and the definition of policies in (9). Finally, we have that $c(x_i,a_i)=c_i$ for any $i\in[n]$. Thus with probability at least $1-\delta$ the following inequality holds for any distribution $\mathbb{Q}$ on $\mathcal{H}$

$$\begin{aligned}\left|R^{\alpha}(\pi_{\mathbb{Q}})-\hat{R}_n^{\alpha}(\pi_{\mathbb{Q}})\right| \leq \frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P})+\log(2/\delta)}{n\lambda}+&\frac{\lambda}{2}\mathbb{E}_{x\sim\nu,a\sim\pi_0(\cdot|x)}\left[\frac{\pi_{\mathbb{Q}}(a|x)}{\pi_0(a|x)^{2\alpha}}c(x,a)^2\right]\\ &+\frac{\lambda}{2n}\sum_{i=1}^{n}\frac{\pi_{\mathbb{Q}}(a_i|x_i)}{\pi_0(a_i|x_i)^{2\alpha}}c_i^2\,.\end{aligned}\tag{42}$$

This concludes the proof. □

C.4. Sample Complexity

Proposition C.4. *Let $\mathcal{M}_1(\mathcal{H})$ be the set of probability distributions on the hypothesis space $\mathcal{H}$, and let $\lambda > 0$, $n \geq 1$, $\delta \in [0,1]$, $\alpha \in [0,1]$, and let $\mathbb{P}$ be a fixed prior on $\mathcal{H}$, then with probability at least $1-\delta$ over draws $\mathcal{D}_n \sim \mu^n_{\pi_0}$, we have*

$$R(\pi_{\hat{\mathbb{Q}}_n}) \leq R(\pi_{\mathbb{Q}_*}) + 2\sqrt{\frac{\mathrm{KL}_1(\pi_{\mathbb{Q}_*})}{2n}} + 2B^\alpha_n(\pi_{\mathbb{Q}_*}) + 2\frac{\mathrm{KL}_2(\pi_{\mathbb{Q}_*})}{n\lambda} + \lambda \bar{V}^\alpha_n(\pi_{\mathbb{Q}_*})\,.$$

where $\pi_{\hat{\mathbb{Q}}_n}$ is the learned policy with $\hat{\mathbb{Q}}_n = \mathrm{argmin}_{\mathbb{Q}\in\mathcal{M}_1(\mathcal{H})} \hat{R}^\alpha_n(\pi_{\mathbb{Q}}) + \sqrt{\frac{\mathrm{KL}_1(\pi_{\mathbb{Q}})}{2n}} + B^\alpha_n(\pi_{\mathbb{Q}}) + \frac{\mathrm{KL}_2(\pi_{\mathbb{Q}})}{n\lambda} + \frac{\lambda}{2}\bar{V}^\alpha_n(\pi_{\mathbb{Q}})$, $\mathbb{Q}_* = \mathrm{argmin}_{\mathbb{Q}\in\mathcal{M}_1(\mathcal{H})} R(\pi_{\mathbb{Q}})$, and

$$\mathrm{KL}_1(\pi_{\mathbb{Q}}) = D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{4\sqrt{n}}{\delta}\,, \qquad \mathrm{KL}_2(\pi_{\mathbb{Q}}) = D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{4}{\delta}\,,$$

$$B^\alpha_n(\pi_{\mathbb{Q}}) = 1 - \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{a\sim\pi_{\mathbb{Q}}(\cdot|x_i)}\left[\pi_0^{1-\alpha}(a|x_i)\right]\,, \quad \bar{V}^\alpha_n(\pi_{\mathbb{Q}}) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{a\sim\pi_0(\cdot|x_i)}\left[\frac{\pi_{\mathbb{Q}}(a|x_i)}{\pi_0(a|x_i)^{2\alpha}}\right] + \frac{\pi_{\mathbb{Q}}(a_i|x_i)c_i^2}{\pi_0(a_i|x_i)^{2\alpha}}\,.$$

Proof. First, Theorem 4.1 holds for any potentially data dependent distribution $\mathbb{Q}$ on $\mathcal{H}$. In particular, we have that with probability at least $1-\delta$ the following inequalities hold simultaneously for $\hat{\mathbb{Q}}_n$ and $\mathbb{Q}_*$

$$|R(\pi_{\hat{\mathbb{Q}}_n}) - \hat{R}^\alpha_n(\pi_{\hat{\mathbb{Q}}_n})| \leq \sqrt{\frac{\mathrm{KL}_1(\pi_{\hat{\mathbb{Q}}_n})}{2n}} + B^\alpha_n(\pi_{\hat{\mathbb{Q}}_n}) + \frac{\mathrm{KL}_2(\pi_{\hat{\mathbb{Q}}_n})}{n\lambda} + \frac{\lambda}{2}\bar{V}^\alpha_n(\pi_{\hat{\mathbb{Q}}_n})\,,$$

$$|R(\pi_{\mathbb{Q}_*}) - \hat{R}^\alpha_n(\pi_{\mathbb{Q}_*})| \leq \sqrt{\frac{\mathrm{KL}_1(\pi_{\mathbb{Q}_*})}{2n}} + B^\alpha_n(\pi_{\mathbb{Q}_*}) + \frac{\mathrm{KL}_2(\pi_{\mathbb{Q}_*})}{n\lambda} + \frac{\lambda}{2}\bar{V}^\alpha_n(\pi_{\mathbb{Q}_*})\,.$$

Taking only one side of these inequalities yields that with probability at least $1-\delta$ the following inequalities hold simultaneously for $\hat{\mathbb{Q}}_n$ and $\mathbb{Q}_*$

$$R(\pi_{\hat{\mathbb{Q}}_n}) \leq \underbrace{\hat{R}^\alpha_n(\pi_{\hat{\mathbb{Q}}_n}) + \sqrt{\frac{\mathrm{KL}_1(\pi_{\hat{\mathbb{Q}}_n})}{2n}} + B^\alpha_n(\pi_{\hat{\mathbb{Q}}_n}) + \frac{\mathrm{KL}_2(\pi_{\hat{\mathbb{Q}}_n})}{n\lambda} + \frac{\lambda}{2}\bar{V}^\alpha_n(\pi_{\hat{\mathbb{Q}}_n})}_{(I)}\,,$$

$$\hat{R}^\alpha_n(\pi_{\mathbb{Q}_*}) \leq R(\pi_{\mathbb{Q}_*}) + \sqrt{\frac{\mathrm{KL}_1(\pi_{\mathbb{Q}_*})}{2n}} + B^\alpha_n(\pi_{\mathbb{Q}_*}) + \frac{\mathrm{KL}_2(\pi_{\mathbb{Q}_*})}{n\lambda} + \frac{\lambda}{2}\bar{V}^\alpha_n(\pi_{\mathbb{Q}_*})\,.$$

Now using the definition of $\pi_{\hat{\mathbb{Q}}_n}$, we know that

$$I \leq \hat{R}^\alpha_n(\pi_{\mathbb{Q}_*}) + \sqrt{\frac{\mathrm{KL}_1(\pi_{\mathbb{Q}_*})}{2n}} + B^\alpha_n(\pi_{\mathbb{Q}_*}) + \frac{\mathrm{KL}_2(\pi_{\mathbb{Q}_*})}{n\lambda} + \frac{\lambda}{2}\bar{V}^\alpha_n(\pi_{\mathbb{Q}_*})\,.$$

This yields that with probability at least $1-\delta$ the following inequalities hold simultaneously for $\hat{\mathbb{Q}}_n$ and $\mathbb{Q}_*$

$$R(\pi_{\hat{\mathbb{Q}}_n}) \leq \hat{R}^\alpha_n(\pi_{\mathbb{Q}_*}) + \sqrt{\frac{\mathrm{KL}_1(\pi_{\mathbb{Q}_*})}{2n}} + B^\alpha_n(\pi_{\mathbb{Q}_*}) + \frac{\mathrm{KL}_2(\pi_{\mathbb{Q}_*})}{n\lambda} + \frac{\lambda}{2}\bar{V}^\alpha_n(\pi_{\mathbb{Q}_*})\,,$$

$$\hat{R}^\alpha_n(\pi_{\mathbb{Q}_*}) \leq R(\pi_{\mathbb{Q}_*}) + \sqrt{\frac{\mathrm{KL}_1(\pi_{\mathbb{Q}_*})}{2n}} + B^\alpha_n(\pi_{\mathbb{Q}_*}) + \frac{\mathrm{KL}_2(\pi_{\mathbb{Q}_*})}{n\lambda} + \frac{\lambda}{2}\bar{V}^\alpha_n(\pi_{\mathbb{Q}_*})\,.$$

Computing the sum of these two inequalities concludes the proof. □

Corollary C.5 (Special case of Proposition C.4)**.** *Let $\mathcal{H} = \left\{h_\theta\,; \theta \in \mathbb{R}^{dK}\right\}$ of mappings $h_\theta(x) = \mathrm{argmax}_{a\in\mathcal{A}}\, \phi(x)^\top\theta_a$ for any $x \in \mathcal{X}$. Let $n \geq 1$, $\delta \in [0,1]$, $\alpha \in [0,1]$, and let $\mathbb{P} = \mathcal{N}(\mu_0, I_{dK})$ be a fixed prior on $\mathcal{H}$, then with probability at least $1-\delta$ over draws $\mathcal{D}_n \sim \mu^n_{\pi_0}$, we have that*

$$R(\pi_{\hat{\mathbb{Q}}_n}) \leq R(\pi_{\mathbb{Q}_*}) + \frac{\sqrt{\|\mu_* - \mu_0\|^2 + 2\log\frac{4\sqrt{n}}{\delta}}}{\sqrt{n}} + 2(1-K^{\alpha-1}) + \frac{\|\mu_* - \mu_0\|^2 + 2\log\frac{4}{\delta}}{\sqrt{n}} + \frac{K^{2\alpha-1} + K^{2\alpha}}{\sqrt{n}}\,.$$

where $\pi_{\hat{\mathbb{Q}}_n}$ is the learned policy with $\hat{\mathbb{Q}}_n = \mathrm{argmin}_{\mathbb{Q}=\mathcal{N}(\mu,I_{dK})} \hat{R}^\alpha_n(\pi_{\mathbb{Q}}) + \sqrt{\frac{\mathrm{KL}_1(\pi_{\mathbb{Q}})}{2n}} + B^\alpha_n(\pi_{\mathbb{Q}}) + \frac{\mathrm{KL}_2(\pi_{\mathbb{Q}})}{n\lambda} + \frac{\lambda}{2}\bar{V}^\alpha_n(\pi_{\mathbb{Q}})$, $\mathbb{Q}_* = \mathrm{argmin}_{\mathbb{Q}=\mathcal{N}(\mu,I_{dK})} R(\pi_{\mathbb{Q}})$.

Proof. This result follows from the general Proposition C.4 by simply setting $\mathbb{P} = \mathcal{N}(\mu_0, I_{dK})$ and $\mathbb{Q}_* = \mathcal{N}(\mu_*, I_{dK})$. First, since the covariance matrices of both distributions are I_{dK}, their KL divergence is $D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) = \|\mu_* - \mu_0\|^2/2$. Moreover, since the logging policy is uniform then $B_n^\alpha(\pi_\mathbb{Q}) = (1 - K^{\alpha-1})$ and $\bar{V}_n^\alpha(\pi_\mathbb{Q}) \leq K^{2\alpha-1} + K^{2\alpha}$. Using these quantities, setting $\lambda = 1/\sqrt{n}$ and applying Proposition C.4 yields that with probability at least $1-\delta$ over draws $\mathcal{D}_n \sim \mu_{\pi_0}^n$, we have that

$$R(\pi_{\hat{\mathbb{Q}}_n}) \leq R(\pi_{\mathbb{Q}_*}) + \frac{\sqrt{\|\mu_* - \mu_0\|^2 + 2\log\frac{4\sqrt{n}}{\delta}}}{\sqrt{n}} + 2(1 - K^{\alpha-1}) + \frac{\|\mu_* - \mu_0\|^2 + 2\log\frac{4}{\delta}}{\sqrt{n}} + \frac{K^{2\alpha-1} + K^{2\alpha}}{\sqrt{n}}\,.$$

This concludes the proof. □

The above corollary allows us to give insights into the sample complexity of our procedure. That is, the number of samples needed so that the performance of the learned policy $\pi_{\hat{\mathbb{Q}}_n}$ is close to that of the optimal one. Let $\epsilon > 2(1 - K^{\alpha-1})$ for $\alpha \in [1 - \log 2/\log K, 1]$. This condition on α ensures that $\epsilon \in [0, 1]$ and it is mild as α is often close to 1. Let δ, then the following implication holds

$$\epsilon \geq \frac{\sqrt{\|\mu_* - \mu_0\|^2 + 2\log\frac{4\sqrt{n}}{\delta}}}{\sqrt{n}} + 2(1 - K^{\alpha-1}) + \frac{\|\mu_* - \mu_0\|^2 + 2\log\frac{4}{\delta}}{\sqrt{n}} + \frac{K^{2\alpha-1} + K^{2\alpha}}{\sqrt{n}}$$
$$\implies \mathbb{P}(R(\pi_{\hat{\mathbb{Q}}_n}) \leq R(\pi_{\mathbb{Q}_*}) + \epsilon) \geq 1 - \delta\,. \tag{43}$$

First, we use that $\sqrt{\|\mu_* - \mu_0\|^2 + 2\log\frac{4\sqrt{n}}{\delta}} \leq \|\mu_* - \mu_0\| + \sqrt{2\log\frac{4\sqrt{n}}{\delta}}$. Moreover we bound $K^{2\alpha-1} + K^{2\alpha} \leq 2K^{2\alpha}$. Then the implication in (43) becomes

$$\sqrt{n} \geq \frac{\|\mu_* - \mu_0\| + \|\mu_* - \mu_0\|^2 + 2\log\frac{4}{\delta} + \sqrt{2\log\frac{4\sqrt{n}}{\delta}} + 2K^{2\alpha}}{\epsilon - 2(1 - K^{\alpha-1})} \implies \mathbb{P}(R(\pi_{\hat{\mathbb{Q}}_n}) \leq R(\pi_{\mathbb{Q}_*}) + \epsilon) \geq 1 - \delta\,. \tag{44}$$

We only provide intuition on the sample complexity and aim at having easy-to-interpret terms. Thus we omit the logarithmic terms in (44) and assume that $\|\mu_* - \mu_0\|^2 \geq \|\mu_* - \mu_0\|$. This leads to the claim made in Section 5.1. Of course, a more precise sample complexity analysis can be made by studying the function $h(x) = \sqrt{x} - \sqrt{2\log\frac{4\sqrt{x}}{\delta}}/(\epsilon - 2(1 - K^{\alpha-1}))$ and finding x such that $f(x) \geq \frac{\|\mu_* - \mu_0\| + \|\mu_* - \mu_0\|^2 + 2\log\frac{4}{\delta} + 2K^{2\alpha}}{\epsilon - 2(1 - K^{\alpha-1})}$.

D. Experiments

D.1. Setup

We consider the standard supervised-to-bandit conversion (Agarwal et al., 2014). Precisely, let $\mathcal{S}_n^{\mathrm{TR}}$ and $\mathcal{S}_{n_{\mathrm{TS}}}^{\mathrm{TS}}$ be the training and testing set of a classification dataset, respectively. First, we transform the training set $\mathcal{S}_n^{\mathrm{TR}}$ to a logged bandit data $\mathcal{D}_n$ as described in Algorithm 1. The resulting logged data $\mathcal{D}_n$ is then used to train our policies. After that, the learned policies are tested on $\mathcal{S}_{n_{\mathrm{TS}}}^{\mathrm{TS}}$ as described in Algorithm 2. We consider that the resulting reward in Algorithm 2 is a good proxy for the unknown true reward of the learned policies. This will be our performance metric, the higher the better.

In our experiments, we use the following image classification datasets `MNIST` (LeCun et al., 1998), `FashionMNIST` (Xiao et al., 2017), `EMNIST` (Cohen et al., 2017) and `CIFAR100` (Krizhevsky et al., 2009). We provide a summary of the statistics of these datasets in Table 1. Algorithm 1 takes as input a logging policy π_0 which we define as

$$\pi_0(a|x) = \frac{\exp(\eta_0 \cdot \phi(x)^\top \mu_{0,a})}{\sum_{a' \in \mathcal{A}} \exp(\eta_0 \cdot \phi(x)^\top \mu_{0,a'})}\,, \qquad \forall (x, a) \in \mathcal{X} \times \mathcal{A}\,. \tag{45}$$

Here $\phi(x) \in \mathbb{R}^d$ is the feature transformation function that outputs a d-dimensional vector, $\mu_0 = (\mu_{0,a})_{a \in \mathcal{A}} \in \mathbb{R}^{dK}$ are learnable parameters and η_0 is an inverse-temperature parameter for the `softmax` in (45). We explain next how these quantities are derived in detail.

The feature transformation function $\phi(x) \in \mathbb{R}^d$: for all the datasets, except `CIFAR100`, the feature transformation function $\phi(\cdot)$ is defined as $\phi(x) = \frac{x}{\|x\|}$ for any $x \in \mathcal{X}$. That is, we simply normalize the features $x \in \mathcal{X}$ by their L_2

Table 1. Statistics of the datasets used in our experiments.

DATA SET	NBR. TRAIN SAMPLES n	NBR. TEST SAMPLES n_{TS}	NBR. ACTIONS K	DIMENSION d
MNIST	60000	10000	10	784
FASHIONMNIST	60000	10000	10	784
EMNIST	112800	18800	47	784
CIFAR100	50000	10000	100	2048

norm $\|x\|$. In contrast, CIFAR100 is a more challenging problem. Thus we use transfer learning to extract features $\phi(x)$ expressive enough so that a linear softmax model would enjoy a reasonable performance. Precisely, we retrieve the last hidden layer of a ResNet-50 network, pre-trained on the ImageNet dataset, to output 2048-dimensional features. Finally, the obtained features are normalized as $\frac{x}{\|x\|}$ and this whole process (ResNet-50 + normalization) corresponds to $\phi(\cdot)$ for CIFAR100.

The parameters $\mu_0 = (\mu_{0,a})_{a\in\mathcal{A}} \in \mathbb{R}^{dK}$: we learn the parameters μ_0 using 5% of the training set $\mathcal{S}_n^{\text{TR}}$. Precisely, we use the cross-entropy loss with an L_2 regularization of 10^{-6} to prevent the logging policy π_0 from being degenerate. This ensures that the learning policies are absolutely continuous with respect to the logging policy π_0, a condition under which standard IPS is unbiased. In optimization, we use Adam (Kingma & Ba, 2014) with a learning rate of 0.1 for 10 epochs. In all the experiments, we set the prior $\mathbb{P} = \mathcal{N}(\eta_0\mu_0, I_{dK})$ for the Gaussian policies in (12) and we set it as $\mathbb{P} = \mathcal{N}(\eta_0\mu_0, I_{dK}) \times \mathrm{G}(0,1)^K$ for the mixed-logit policies in (11). Our theory requires that the prior does not depend on data. Given that μ_0 is learned on the 5% portion of data, we only train our learning policies on the remaining 95% portion of the data to match our theoretical requirements.

The inverse-temperature parameter $\eta_0 \in \mathbb{R}$: this controls the performance of the logging policy. A high positive value of η_0 leads to a well-performing logging policy, while a negative one leads to a low-performing logging policy. When $\eta_0 = 0$, π_0 is identical to the uniform policy. In our experiments η_0 varies between 0 and 1.

Algorithm 1 Supervised-to-bandit: creating logged data

Input: training classification set $\mathcal{S}_n^{\text{TR}} = \{(x_i, y_i)\}_{i=1}^n$, logging policy π_0.
Output: logged bandit data $\mathcal{D}_n = (x_i, a_i, c_i)_{i\in[n]}$.
Initialize $\mathcal{D}_n = \{\}$
for $i = 1, \dots, n$ **do**
  $a_i \sim \pi_0(\cdot|x_i)$
  $c_i = -\mathbb{I}_{\{a_i = y_i\}}$
  $\mathcal{D}_n \leftarrow \mathcal{D}_n \cup \{(x_i, a_i, c_i)\}$.

Algorithm 2 Supervised-to-bandit: testing policies

Input: image classification dataset $\mathcal{S}_{n_{\text{TS}}}^{\text{TS}} = \{(x_i, y_i)\}_{i=1}^{n_{\text{TS}}}$, learned policy $\hat{\pi}_n$.
Output: reward r.
for $i = 1, \dots, n_{\text{TS}}$ **do**
  $a_i \sim \hat{\pi}_n(\cdot|x_i)$
  $r_i = \mathbb{I}_{\{a_i = y_i\}}$
$r = \frac{1}{n_{\text{TS}}} \sum_{i=1}^{n_{\text{TS}}} r_i$.

Now it remains to explain the learning policies $\pi_{\mathbb{Q}}$ and the corresponding closed-form bounds using either our results or those in existing works (London & Sandler, 2019; Sakhi et al., 2022).

D.2. Policies

Here we present the two families of policies that we use in our experiments, Gaussian and mixed-logit policies.

D.2.1. MIXED-LOGIT

Let $\mathcal{H} = \left\{h_{\theta,\gamma}\,;\theta \in \mathbb{R}^{dK}, \gamma \in \mathbb{R}^K\right\}$ be a hypothesis space of mappings $h_{\theta,\gamma}(x) = \operatorname{argmax}_{a\in\mathcal{A}} \phi(x)^\top\theta_a + \gamma_a$ for any $x \in \mathcal{X}$. Here $\phi(x)$ outputs a d-dimensional representation of context $x \in \mathcal{X}$. Now assume that for any $a \in \mathcal{A}$, γ_a is a standard Gumbel perturbation, $\gamma_a \sim \mathrm{G}(0,1)$, then we have that

$$\begin{aligned}\pi_\theta^{\mathrm{SOF}}(a|x) &= \frac{\exp(\phi(x)^\top\theta_a)}{\sum_{a'\in\mathcal{A}}\exp(\phi(x)^\top\theta_{a'})}\,,\\ &= \mathbb{E}_{\gamma\sim\mathrm{G}(0,1)^K}\left[\mathbb{I}_{\{h_{\theta,\gamma}(x)=a\}}\right]\,.\end{aligned} \tag{46}$$

In addition, we randomize θ such as $\theta \sim \mathcal{N}(\mu, \sigma^2 I_{dK})$ where $\mu \in \mathbb{R}^{dK}$ and $\sigma > 0$. It follows that the posterior $\mathbb{Q}$ is a multivariate Gaussian $\mathcal{N}(\mu, \sigma^2 I_{dK})$ over the parameters θ with standard Gumbel perturbations $\gamma \sim \mathrm{G}(0,1)^K$. We denote such policies by $\pi_{\mu,\sigma}^{\mathrm{MIXL}}$ and they are defined as

$$\begin{aligned}\pi_{\mu,\sigma}^{\mathrm{MIXL}}(a|x) &= \mathbb{E}_{\theta\sim\mathcal{N}(\mu,\sigma^2 I_{dK})}\left[\frac{\exp(\phi(x)^\top\theta_a)}{\sum_{a'\in\mathcal{A}}\exp(\phi(x)^\top\theta_{a'})}\right]\,,\\ &= \mathbb{E}_{\theta\sim\mathcal{N}(\mu,\sigma^2 I_{dK})}\left[\pi_\theta^{\mathrm{SOF}}(a|x)\right]\,,\\ &= \mathbb{E}_{\theta\sim\mathcal{N}(\mu,\sigma^2 I_{dK}),\gamma\sim\mathrm{G}(0,1)^K}\left[\mathbb{I}_{\{h_{\theta,\gamma}(x)=a\}}\right]\,.\end{aligned} \tag{47}$$

To sample from the mixed-logit policies $\pi_{\mu,\sigma}^{\mathrm{MIXL}}$, we first sample $\theta \sim \mathcal{N}(\mu, \sigma^2 I_{dK})$ and $\gamma \sim \mathrm{G}(0,1)^K$ and then set the sampled action as $a \leftarrow h_{\theta,\gamma}(x)$. Now we also need to compute the gradient of the expectation in (47). This needs additional care since the distribution under which we take the expectation depends on the parameters μ, σ. Fortunately, the reparameterization trick can be used in this case. Roughly speaking, it allows us to express a gradient of the expectation in (47) as an expectation of a gradient. In our case, we use the *local* reparameterizaton trick (Kingma et al., 2015) which is known for reducing the variance of stochastic gradients. Precisely, we rewrite (47) as

$$\begin{aligned}\pi_{\mu,\sigma}^{\mathrm{MIXL}}(a|x) &= \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\|\phi(x)\|^2 I_K)}\left[\frac{\exp(\phi(x)^\top\mu_a + \sigma\epsilon_a)}{\sum_{a'\in\mathcal{A}}\exp(\phi(x)^\top\mu_{a'} + \sigma\epsilon_{a'})}\right]\,.\\ &= \mathbb{E}_{\epsilon\sim\mathcal{N}(0,I_K)}\left[\frac{\exp(\phi(x)^\top\mu_a + \sigma\epsilon_a)}{\sum_{a'\in\mathcal{A}}\exp(\phi(x)^\top\mu_{a'} + \sigma\epsilon_{a'})}\right]\,,\end{aligned}$$

where we used that $\|\phi(x)\|^2 = 1$ since features are normalized. It follows that gradients read

$$\nabla_{\mu,\sigma}\pi_{\mu,\sigma}^{\mathrm{MIXL}}(a|x) = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,I_K)}\left[\nabla_{\mu,\sigma}\frac{\exp(\phi(x)^\top\mu_a + \sigma\epsilon_a)}{\sum_{a'\in\mathcal{A}}\exp(\phi(x)^\top\mu_{a'} + \sigma\epsilon_{a'})}\right]\,.$$

Moreover, the propensities are approximated as

$$\pi_{\mu,\sigma}^{\mathrm{MIXL}}(a|x) \approx \frac{1}{S}\sum_{i\in[S]}\frac{\exp(\phi(x)^\top\mu_a + \sigma\epsilon_{i,a})}{\sum_{a'\in\mathcal{A}}\exp(\phi(x)^\top\mu_{a'} + \sigma\epsilon_{i,a'})}\,, \qquad \epsilon_i \sim \mathcal{N}(0, I_K)\,, \forall i \in [S]\,. \tag{48}$$

In all our experiments, we set $S = 32$.

D.2.2. GAUSSIAN

We define the hypothesis space $\mathcal{H} = \left\{h_\theta\,;\theta \in \mathbb{R}^{dK}\right\}$ of mappings $h_\theta(x) = \operatorname{argmax}_{a\in\mathcal{A}} \phi(x)^\top\theta_a$ for any $x \in \mathcal{X}$. It follows that the learning policies $\pi_{\mathbb{Q}} = \pi_{\mu,\sigma}^{\mathrm{GAUS}}$ read

$$\pi_{\mu,\sigma}^{\mathrm{GAUS}}(a|x) = \mathbb{E}_{\theta\sim\mathcal{N}(\mu,\sigma^2 I_{dK})}\left[\mathbb{I}_{\{h_\theta(x)=a\}}\right]\,. \tag{49}$$

To see why this can be beneficial (Sakhi et al., 2022), let π_* be the optimal policy. Given $x \in \mathcal{X}$, $\pi_*(\cdot|x)$ should be deterministic; it chooses the best action for context x with probability 1. That is, there exists $\mu_* \in \mathbb{R}^{dK}$ such that $\pi_* = \mathbb{I}_{\{h_{\mu_*}(x)=a\}}$. When $\mu \to \mu_*$ and $\sigma \to 0$, the Gaussian policy in (49) approaches π_*. In contrast, the mixed-logit policy in (47) approaches $\pi_{\mu_*}^{\mathrm{SOF}}$. However, $\pi_{\mu_*}^{\mathrm{SOF}}$ is not deterministic due to the additional randomness in γ and is equal to π_* only if

$\phi(x)^\top \mu_{*,a_*(x)} \to \infty$. This explains the choice of removing the Gumbel noise. First, Sakhi et al. (2022) showed that (49) can be written as

$$\pi^{\text{GAUS}}_{\mu,\sigma}(a|x) = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1)}\Big[\prod_{a'\neq a}\Phi\big(\epsilon + \frac{\phi(x)^\top(\mu_a - \mu_{a'})}{\sigma\|\phi(x)\|}\big)\Big],$$

where Φ is the cumulative distribution function of a standard normal variable. But $\|\phi(x)\| = 1$ in all our experiments. Thus

$$\pi^{\text{GAUS}}_{\mu,\sigma}(a|x) = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1)}\Big[\prod_{a'\neq a}\Phi\big(\epsilon + \frac{\phi(x)^\top(\mu_a - \mu_{a'})}{\sigma}\big)\Big].$$

Then similarly to mixed-logit policies, the gradient reads

$$\nabla_{\mu,\sigma}\pi^{\text{GAUS}}_{\mu,\sigma}(a|x) = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1)}\Big[\nabla_{\mu,\sigma}\prod_{a'\neq a}\Phi\big(\epsilon + \frac{\phi(x)^\top(\mu_a - \mu_{a'})}{\sigma}\big)\Big].$$

Moreover, the propensities are approximated as

$$\pi^{\text{GAUS}}_{\mu,\sigma}(a|x) \approx \frac{1}{S}\sum_{i\in[S]}\prod_{a'\neq a}\Phi\big(\epsilon_i + \frac{\phi(x)^\top(\mu_a - \mu_{a'})}{\sigma}\big), \qquad \epsilon_i \sim \mathcal{N}(0,1), \forall i \in [S]. \tag{50}$$

In all our experiments, we set $S = 32$.

D.3. Baselines

Here we present all the methods that we use in our experiments. For each method, we state the result that holds for any learning policy π. After that, we derive the corresponding closed-form bounds for Gaussian and mixed-logit policies that we presented previously. All the baselines require computing the KL divergence between the prior $\mathbb{P}$ and the posterior $\mathbb{Q}$. Thus before presenting them, we state the following lemma that allows bounding the KL divergence between the prior $\mathbb{P}$ and the posterior $\mathbb{Q}$ in the cases of mixed-logit or Gaussian policies.

Lemma D.1 (KL divergence for Gaussian distributions with Gumbel noise)**.** *For distributions* $\mathbb{P} = \mathcal{N}\left(\mu_0, \sigma_0^2 I_{dK}\right) \times \mathrm{G}(0,1)^K$ *and* $\mathbb{Q} = \mathcal{N}\left(\mu, \sigma^2 I_{dK}\right) \times \mathrm{G}(0,1)^K$*, with* $\mu_0, \mu \in \mathbb{R}^{dK}$ *and* $0 < \sigma^2 \leq \sigma_0^2 < \infty$*,*

$$D_{\text{KL}}(\mathbb{Q}\|\mathbb{P}) \leq \frac{\|\mu - \mu_0\|^2}{2\sigma_0^2} + \frac{dK}{2}\log\frac{\sigma_0^2}{\sigma^2}.$$

Moreover, this result holds when the Gumbel noise is removed. That is when $\mathbb{P} = \mathcal{N}\left(\mu_0, \sigma_0^2 I_{dK}\right)$ *and* $\mathbb{Q} = \mathcal{N}\left(\mu, \sigma^2 I_{dK}\right)$*.*

We borrow this lemma from London & Sandler (2019). In particular, Lemma D.1 shows that the KL terms for both policies can be bounded by the same quantity. As a result, the corresponding bounds will be the same; the only difference is the space of learning policies on which we optimize. For completeness, however, we write these bounds for both types of policies although they are similar. Since existing approaches are not named, we name them as **(Author, Policy)** where **Author** $\in$ **{Ours, London et al., Sakhi et al. 1, Sakhi et al. 2}** and **Policy** $\in$ **{Gaussian, Mixed-Logit}** . Here **Ours**, **London et al.**, **Sakhi et al. 1** and **Sakhi et al. 2** correspond to Theorem 4.1, London & Sandler (2019, Theorem 1), Sakhi et al. (2022, Proposition 1), Sakhi et al. (2022, Proposition 3), respectively. For example, London & Sandler (2019, Theorem 1) leads to two baselines **(London et al., Gaussian)** and **(London et al., Mixed-Logit)**. In all our experiments, the learning policies are trained using Adam (Kingma & Ba, 2014) with a learning rate of 0.1 for 20 epochs.

D.3.1. Ours, Theorem 4.1

(Ours, Gaussian) Here we use the Gaussian policies in (49). Thus we only replace the term, $D_{\text{KL}}(\mathbb{Q}\|\mathbb{P})$, with its closed-form bound in Lemma D.1. This leads to the following objective.

$$\min_{\mu\in\mathbb{R}^{dK},\sigma>0}\Big(\hat{R}_n^\alpha\left(\pi^{\text{GAUS}}_{\mu,\sigma}\right) + \sqrt{\frac{\frac{\|\mu-\mu_0\|^2}{2} - \frac{dK}{2}\log\sigma^2 + \log\frac{4\sqrt{n}}{\delta}}{2n}} + B_n^\alpha(\pi^{\text{GAUS}}_{\mu,\sigma}) + \frac{\frac{\|\mu-\mu_0\|^2}{2} - \frac{dK}{2}\log\sigma^2 + \log\frac{4}{\delta}}{n\lambda} + \frac{\lambda}{2}\bar{V}_n^\alpha(\pi^{\text{GAUS}}_{\mu,\sigma})\Big),$$

where we used that $\sigma_0 = 1$ since our prior is $\mathbb{P} = \mathcal{N}(\eta_0\mu_0, I_{dK})$ for Gaussian policies. Moreover, we set $\lambda = \sqrt{2\frac{\frac{\|\mu-\mu_0\|^2}{2} - \frac{dK}{2}\log\sigma^2 + \log\frac{4}{\delta}}{n\bar{V}_n^\alpha(\pi^{\text{GAUS}}_{\mu,\sigma})}}$.

(Ours, Mixed-Logit) Here we use the mixed-logit policies in (47). Thus we only replace the terms, $D_{\text{KL}}(\mathbb{Q}\|\mathbb{P})$, with their closed-form bound in Lemma D.1. This leads to the following objective.

$$\min_{\mu\in\mathbb{R}^{dK},\sigma>0}\Bigg(\hat{R}_n^\alpha\left(\pi^{\text{MIXL}}_{\mu,\sigma}\right) + \sqrt{\frac{\frac{\|\mu-\mu_0\|^2}{2} - \frac{dK}{2}\log\sigma^2 + \log\frac{4\sqrt{n}}{\delta}}{2n}} + B_n^\alpha(\pi^{\text{MIXL}}_{\mu,\sigma}) + \frac{\frac{\|\mu-\mu_0\|^2}{2} - \frac{dK}{2}\log\sigma^2 + \log\frac{4}{\delta}}{n\lambda} + \frac{\lambda}{2}\bar{V}_n^\alpha(\pi^{\text{MIXL}}_{\mu,\sigma})\Bigg),$$

where we used that $\sigma_0 = 1$ since our prior is $\mathbb{P} = \mathcal{N}(\eta_0\mu_0, I_{dK}) \times \text{G}(0,1)^K$ for mixed-logit policies. Moreover, we set $\lambda = \sqrt{2\frac{\frac{\|\mu-\mu_0\|^2}{2} - \frac{dK}{2}\log\sigma^2 + \log\frac{4}{\delta}}{n\bar{V}_n^\alpha(\pi^{\text{MIXL}}_{\mu,\sigma})}}$.

D.3.2. London & Sandler (2019, Theorem 1)

Proposition D.2. *Let $\tau \in (0,1)$, $n \geq 1$, $\delta \in (0,1)$ and let $\mathbb{P}$ be a fixed prior on $\mathcal{H}$, then with probability at least $1-\delta$ over draws $\mathcal{D}_n \sim \mu^n_{\pi_0}$, the following holds simultaneously for all posteriors, $\mathbb{Q}$, on $\mathcal{H}$ that*

$$R(\pi_\mathbb{Q}) \leq \hat{R}_n^\tau(\pi_\mathbb{Q}) + \sqrt{\frac{2\left(\hat{R}_n^\tau(\pi_\mathbb{Q}) + \frac{1}{\tau}\right)\left(D_{\text{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{n}{\delta}\right)}{\tau(n-1)}} + \frac{2\left(D_{\text{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{n}{\delta}\right)}{\tau(n-1)}. \tag{51}$$

Baseline 1: (London et al., Gaussian) Here we use the Gaussian policies in (49). Thus we only replace the terms, $D_{\text{KL}}(\mathbb{Q}\|\mathbb{P})$, with their closed-form bound in Lemma D.1. This leads to the following objective.

$$\min_{\mu\in\mathbb{R}^{dK},\sigma>0}\Bigg(\hat{R}_n^\tau\left(\pi^{\text{GAUS}}_{\mu,\sigma}\right) + \sqrt{\frac{2\left(\hat{R}_n^\tau\left(\pi^{\text{GAUS}}_{\mu,\sigma}\right) + \frac{1}{\tau}\right)\left(\frac{\|\mu-\mu_0\|^2}{2} - \frac{dK}{2}\log\sigma^2 + \log\frac{n}{\delta}\right)}{\tau(n-1)} + \frac{2\left(\frac{\|\mu-\mu_0\|^2}{2} - \frac{dK}{2}\log\sigma^2 + \log\frac{n}{\delta}\right)}{\tau(n-1)}}\Bigg),$$

where we used that $\sigma_0 = 1$ since our prior is $\mathbb{P} = \mathcal{N}(\eta_0\mu_0, I_{dK})$ for Gaussian policies.

Baseline 2: (London et al., Mixed-Logit) Here we consider the mixed-logit policies in (47). Since the additional Gumbel noise does not affect the KL divergence (Lemma D.1), we have the same objective as in the Gaussian case. That is

$$\min_{\mu\in\mathbb{R}^{dK},\sigma>0}\Bigg(\hat{R}_n^\tau\left(\pi^{\text{MIXL}}_{\mu,\sigma}\right) + \sqrt{\frac{2\left(\hat{R}_n^\tau\left(\pi^{\text{MIXL}}_{\mu,\sigma}\right) + \frac{1}{\tau}\right)\left(\frac{\|\mu-\mu_0\|^2}{2} - \frac{dK}{2}\log\sigma^2 + \log\frac{n}{\delta}\right)}{\tau(n-1)} + \frac{2\left(\frac{\|\mu-\mu_0\|^2}{2} - \frac{dK}{2}\log\sigma^2 + \log\frac{n}{\delta}\right)}{\tau(n-1)}}\Bigg),$$

where we used that $\sigma_0 = 1$ since our prior is $\mathbb{P} = \mathcal{N}(\eta_0\mu_0, I_{dK}) \times \text{G}(0,1)^K$ for mixed-logit policies.

D.3.3. Sakhi et al. (2022, Proposition 1)

Proposition D.3. *Let $\tau \in (0,1)$, $n \geq 1$, $\delta \in (0,1)$ and let $\mathbb{P}$ be a fixed prior on $\mathcal{H}$, then with probability at least $1-\delta$ over draws $\mathcal{D}_n \sim \mu^n_{\pi_0}$, the following holds simultaneously for all posteriors, $\mathbb{Q}$, on $\mathcal{H}$ that*

$$R(\pi_\mathbb{Q}) \leq \min_{\lambda>0}\frac{1}{\tau\left(e^\lambda - 1\right)}\left(1 - e^{-\tau\lambda\hat{R}_n^\tau(\pi_\mathbb{Q}) + \frac{D_{\text{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{2\sqrt{n}}{\delta}}{n}}\right). \tag{52}$$

Baseline 3: (Sakhi et al. 1, Gaussian) Here we use the Gaussian policies in (49).

$$\min_{\mu\in\mathbb{R}^{dK},\sigma>0,\lambda>0}\left(\frac{1}{\tau\left(e^{\lambda}-1\right)}\left(1-e^{-\tau\lambda\hat{R}_n^{\tau}\left(\pi_{\mu,\sigma}^{\text{GAUS}}\right)+\frac{\frac{\|\mu-\mu_0\|^2}{2}-\frac{dK}{2}\log\sigma^2+\log\frac{2\sqrt{n}}{\delta}}{n}}\right)\right), \tag{53}$$

where we used that $\sigma_0 = 1$ since our prior is $\mathbb{P} = \mathcal{N}(\eta_0\mu_0, I_{dK})$ for Gaussian policies.

Baseline 4: (Sakhi et al. 1, Mixed-Logit) Here we consider the mixed-logit policies in (47).

$$\min_{\mu\in\mathbb{R}^{dK},\sigma>0,\lambda>0}\left(\frac{1}{\tau\left(e^{\lambda}-1\right)}\left(1-e^{-\tau\lambda\hat{R}_n^{\tau}\left(\pi_{\mu,\sigma}^{\text{MIXL}}\right)+\frac{\frac{\|\mu-\mu_0\|^2}{2}-\frac{dK}{2}\log\sigma^2+\log\frac{2\sqrt{n}}{\delta}}{n}}\right)\right). \tag{54}$$

where we used that $\sigma_0 = 1$ since our prior is $\mathbb{P} = \mathcal{N}(\eta_0\mu_0, I_{dK}) \times \mathrm{G}(0,1)^K$ for mixed-logit policies.

D.3.4. Sakhi et al. (2022, Proposition 3)

Proposition D.4. *Let $\tau \in (0,1)$, $n \geq 1$, $\delta \in (0,1)$, let $\mathbb{P}$ be a fixed prior on $\mathcal{H}$, and let $\Lambda = \{\lambda_i\}_{i\in[n_\lambda]}$ a set of n_λ positive scalars. Then with probability at least $1-\delta$ over draws $\mathcal{D}_n \sim \mu_{\pi_0}^n$, the following holds simultaneously for all posteriors, $\mathbb{Q}$, on $\mathcal{H}$ and any $\lambda_i \in \Lambda$,*

$$R\left(\pi_{\mathbb{Q}}\right) \leq \hat{R}_n^{\tau}\left(\pi_{\mathbb{Q}}\right) + \sqrt{\frac{D_{\text{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{4\sqrt{n}}{\delta}}{2n}} + \frac{D_{\text{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{2n_\lambda}{\delta}}{\lambda} + \frac{\lambda}{n}g\left(\frac{\lambda}{\tau n}\right)\mathcal{V}_n^{\tau}\left(\pi_{\mathbb{Q}}\right), \tag{55}$$

where $g : u \to \frac{\exp(u)-1-u}{u^2}$ and $\mathcal{V}_n^{\tau}(\pi_{\mathbb{Q}}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{a\sim\pi_{\mathbb{Q}}(\cdot|x_i)}\left[\frac{\pi_0(a|x_i)}{\max(\tau,\pi_0(a|x_i))^2}\right]$, .

Baseline 5: (Sakhi et al. 2, Gaussian) Here we consider the Gaussian policies in (49).

$$\min_{\mu\in\mathbb{R}^{dK},\sigma>0,\lambda\in\Lambda}\left(\hat{R}_n^{\tau}\left(\pi_{\mu,\sigma}^{\text{GAUS}}\right) + \sqrt{\frac{\frac{\|\mu-\mu_0\|^2}{2}-\frac{dK}{2}\log\sigma^2+\log\frac{4\sqrt{n}}{\delta}}{2n}} + \frac{\frac{\|\mu-\mu_0\|^2}{2}-\frac{dK}{2}\log\sigma^2+\log\frac{2n_\lambda}{\delta}}{\lambda} + \frac{\lambda}{n}g\left(\frac{\lambda}{\tau n}\right)\mathcal{V}_n^{\tau}\left(\pi_{\mu,\sigma}^{\text{GAUS}}\right)\right), \tag{56}$$

where we used that $\sigma_0 = 1$ since our prior is $\mathbb{P} = \mathcal{N}(\eta_0\mu_0, I_{dK})$ for Gaussian policies.

Baseline 6: (Sakhi et al. 2, Mixed-Logit) Here we consider the mixed-logit policies in (47).

$$\min_{\mu\in\mathbb{R}^{dK},\sigma>0,\lambda\in\Lambda}\left(\hat{R}_n^{\tau}\left(\pi_{\mu,\sigma}^{\text{MIXL}}\right) + \sqrt{\frac{\frac{\|\mu-\mu_0\|^2}{2}-\frac{dK}{2}\log\sigma^2+\log\frac{4\sqrt{n}}{\delta}}{2n}} + \frac{\frac{\|\mu-\mu_0\|^2}{2}-\frac{dK}{2}\log\sigma^2+\log\frac{2n_\lambda}{\delta}}{\lambda} + \frac{\lambda}{n}g\left(\frac{\lambda}{\tau n}\right)\mathcal{V}_n^{\tau}\left(\pi_{\mu,\sigma}^{\text{MIXL}}\right)\right), \tag{57}$$

where we used that $\sigma_0 = 1$ since our prior is $\mathbb{P} = \mathcal{N}(\eta_0\mu_0, I_{dK}) \times \mathrm{G}(0,1)^K$ for mixed-logit policies.

D.4. Additional Results and Discussion

In Figure 4, we report the reward of the learned policy using one of the considered methods. We make the following observations:

- **Choice of τ and α:** in Figure 4, we set $\tau = 1/\sqrt[4]{n} \approx 0.06$ and $\alpha = 1 - 1/\sqrt[4]{n} \approx 0.94$ so that when n is large enough, both $\hat{R}_n^{\tau}(\pi)$ and $\hat{R}_n^{\alpha}(\pi)$ approach $\hat{R}_n^{\text{IPS}}(\pi)$ (Ionides, 2008). This is because standard IPS should be preferred when $n \to \infty$. For completeness, we also show in Figure 5 that the choice of α and τ does not affect the conclusions that we make here. We also include in Figure 5 the results with an adaptive and data-dependent α obtained using (14) in Section 4.4. The results in Figure 5 will be discussed in detail after we finish analyzing the results in Figure 4.

- **Overall performance:** our method outperforms the baselines for any class of learning policies (Gaussian or mixed-logit) and any choice of logging policies. The only exception is when the logging policy is uniform.
- **Effect of the class of learning policies:** the class of policies, Gaussian or mixed-logit, affects the performance of all the baselines. In general, Gaussian policies behave better than mixed-logit policies. However, this is less significant for our method; the performance of both Gaussian and mixed-logit policies are comparable, and in both cases, our method outperforms the baselines with Gaussian policies. Therefore, in general, Gaussian policies should be preferred over mixed-logit policies. But in case engineering constraints impose the choice of mixed-logit or softmax policies, then the performance of our method is robust to this choice.
- **Effect of the logging policy:** our method reaches the maximum reward even when the logging policy is not performing well. In contrast, the baselines only reach their best reward when the logging policy is already well-performing ($\eta_0 \approx 1$), in which case minor to no improvements are made. Note that the baselines have a better reward than ours when the logging policy is uniform. But our method has better reward when the logging policy is not uniform, that is when $\eta_0 > 0$. This is more common in practice since the logging policy is deployed in production and thus it is expected to perform better than the uniform policy.

In Figure 5, we compare our method to **(Sakhi et al. 2)** with Gaussian policies since this was the best-performing baseline in our experiments in Figure 4. Note that we did not include `CIFAR100` in Figure 4 as it was computationally heavy to run these experiments with varying η_0, α and τ for a very high-dimensional dataset such as `CIFAR100`. We consider 20 varying values of τ and α evenly spaced in $(0, 1)$. We also include the results using the adaptive tuning procedure of α described in Section 4.4 (green curve). We make the following observations:

- **Adaptive and data-dependent α:** This procedure is reliable since the performance with an adaptive α (green curve) is comparable with the best possible choice of α. This is consistent for the three datasets.
- **Effect of the choice α:** as we observed before, the only case where the choice of α may lead to bad-performing policies is when the logging policy is uniform. When the logging policy is not uniform, our method outperforms the best baseline with the best τ for a wide range of values of α. Also, note that there is no very bad choice of α, in contrast with $\tau \approx 0$ that led to a very bad performing policy that slightly improved upon the logging policy. This attests to the robustness of our method to the choice of α. Moreover, our bound regularizes better α; it contains a bias-variance trade-off term for α. Also, the bound of **(Sakhi et al. 2)** has a $1/\tau$ making it vacuous for small values of τ.
- **Best choice of α:** To see the effect of α for varying problems, we consider the following experiment. We split the logging policies into two groups. The first is *modest logging* which corresponds to logging policies whose η_0 is between 0 and 0.5. This includes uniform logging policies and other average-performing logging policies. The second is *good logging* which corresponds to logging policies whose η_0 is between 0.5 and 1. After that, for each α, we compute the average reward of the learned policy across either the group of modest or good logging policies. For each dataset, this leads to the two red and green curves in the second row of Figure 5. Overall, we observe that $\alpha \approx 0.7$ leads to the best performance for the *modest logging* group. Thus when the performance of the logging policy is average, regularizing the importance weights can be critical. In contrast, when the performance of the logging policy is already good, regularization is less needed and we can set $\alpha \approx 1$. Fortunately, one of the main strengths of this work is that our bound also holds for standard IPS recovered for $\alpha = 1$. The bounds in all prior works cannot provide good performance for standard IPS due to their dependency on $1/\tau$.

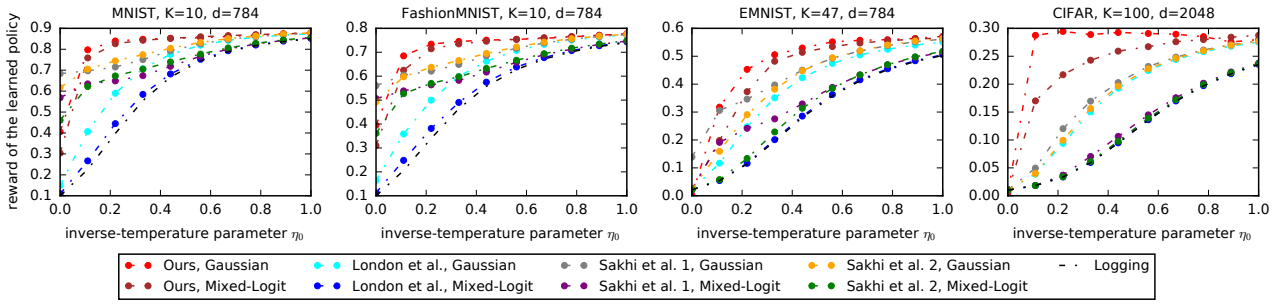

Figure 4. The reward of the learned policy for four datasets with varying quality of the logging policy $\eta_0 \in [0, 1]$.

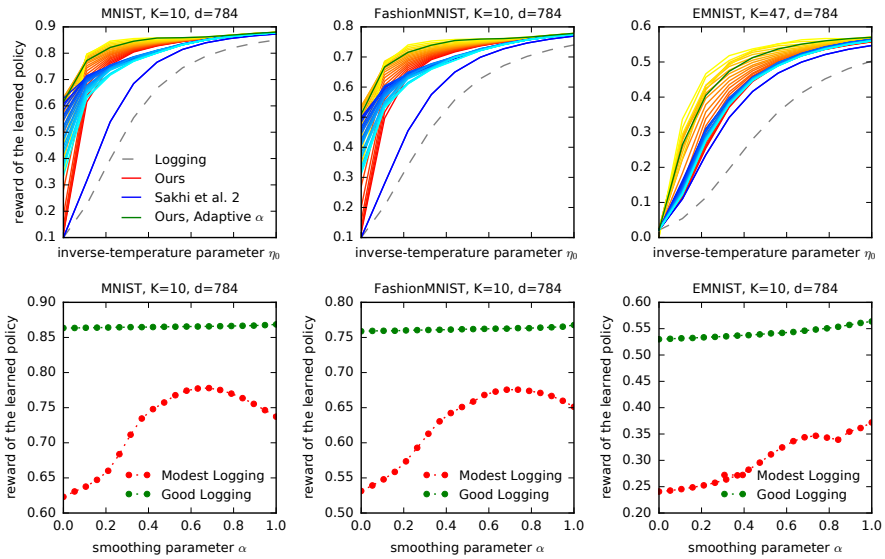

Figure 5. In the first row, we report the reward of the learned policy with 20 evenly space values of $\tau \in (0, 1)$ and $\alpha \in (0, 1)$ and varying $\eta_0 \in [0, 1]$, and for an adaptive and data-dependent α obtained using (14) in Section 4.4. The blue-to-cyan colors correspond to different values of τ. The lighter the color, the higher the value of τ. For instance, the cyan lines correspond to high values of τ while the blue ones correspond to very small values of τ. Similarly, the red-to-yellow colors correspond to different values α. The lighter the color, the higher the value of α. For instance, the yellow lines correspond to high values of α while the red ones correspond to very small values of α. Finally, the green curve corresponds to the reward of the learned policy using an adaptive and data-dependent α described in (14) (Section 4.4). In the second row, we report the *average* reward of the learned policies using our method across the modest logging group ($\eta_0 \in [0, 0.5]$ in red) and the good logging group ($\eta_0 \in [0.5, 1]$ in green).

D.5. Learning Principles

Here we compare our bound in Theorem 4.1 and our learning principle in (16) to the one in London & Sandler (2019). We do not include the learning principle in Swaminathan & Joachims (2015a) since the one in London & Sandler (2019) enjoys similar performance and is far more scalable. The learning principle of London & Sandler (2019) is defined as

$$\min_{\mu} \hat{R}_n^{\tau}(\pi_\mu) + \lambda \|\mu - \mu_0\|^2 \,. \tag{58}$$

where λ is a tunable hyper-parameters, π_μ is the softmax policy defined in (46) and $\mu \in \mathbb{R}^{dK}$ is its parameter vector. This learning principle is referred to as **(London et al., LP)**. In contrast, our learning principle is defined as

$$\hat{R}_n^{\alpha}(\pi_\mu) + \lambda_1 \|\mu - \mu_0\|^2 + \lambda_2 \bar{V}_n^{\alpha}(\pi_\mu) + \lambda_3 B_n^{\alpha}(\pi_\mu) \,, \tag{59}$$

where λ_1, λ_2 and λ_3 are tunable hyper-parameters and π_μ is the Gaussian policy in (12) with a fixed $\sigma = 1$. Our learning principle is referred to as **(Ours, LP)**. Finally, our bound in Theorem 4.1 with Gaussian policies is referred to as **(Ours, Bound)**. Similarly to the previous experiments, we set $\tau = 1/\sqrt[4]{n} \approx 0.06$ and $\alpha = 1 - 1/\sqrt[4]{n} \approx 0.94$ so that when n is large enough, both $\hat{R}_n^{\tau}(\pi)$ and $\hat{R}_n^{\alpha}(\pi)$ approach $\hat{R}_n^{\text{IPS}}(\pi)$ (Ionides, 2008). For the learning principles, we tried multiple values of hyper-parameters $\lambda, \lambda_1, \lambda_2$ and λ_3, all between 10^{-5} and 10^{-1}. For instance, we found that the best hyper-parameter for London & Sandler (2019) is $\lambda = 10^{-5}$ which matches the value they found in their `FashionMNIST` experiments. For our learning principle, the best hyper-parameters were $\lambda_1 = 10^{-5}, \lambda_2 = 10^{-5}$ and $\lambda_3 = 10^{-5}$. In contrast, our bound does not require hyper-parameter tuning. We report in Figure 6 the reward of the learned policy on the `FashionMNIST` for all these methods with varying values of hyper-parameters. To reduce clutter, we only report the reward for good choices of hyper-parameters $\lambda, \lambda_1, \lambda_2$ and λ_3. We observe that for a wide range of hyper-parameters, our learning principle outperforms the one in London & Sandler (2019). However, both learning principles are sensitive to the choice of hyper-parameters. In contrast, our bound does not require the tuning of any additional hyper-parameter and it achieves the best performance except for the uniform logging policy. In addition to being more theoretically grounded, this approach also enjoys favorable empirical performance without additional hyper-parameter tuning, an important practical consideration.

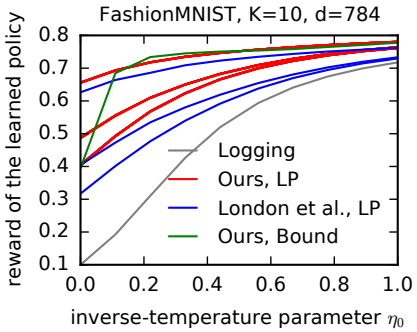

Figure 6. The reward of the learned policy using either our bound in Theorem 4.1 (referred to as **(Ours, Bound)** in green), our learning principle in (16) (referred to as **(Ours, LP)** in red for multiple values of hyper-parameters) or the learning principle in London & Sandler (2019) (referred to as **(London et al., LP)** in blue) for multiple values of hyper-parameters).

D.6. Other Importance Weight Corrections

Su et al. (2020); Metelli et al. (2021) also proposed corrections that are different from hard clipping (a detailed comparison is given in Section 3). However, they were not included in our main experiments since they do not provide generalization guarantees; they focus on OPE and only propose a heuristic for OPL in their Appendix B.2 and Section 6.1.2, respectively. Those heuristics are not based on theory, in contrast with ours which is directly derived from our generalization bound. However, for completeness, we also compare our regularization of importance weights to theirs. To make such a comparison, we use the hyper-parameters and tuning procedures provided in Section 6 and Appendix B.2 for Metelli et al. (2021) and Sections 5 and 6.1.2 for Su et al. (2020). Overall, we observe in Figure 7 that our method outperforms these baselines in OPL and the gap is more significant when the logging policy is not performing well.

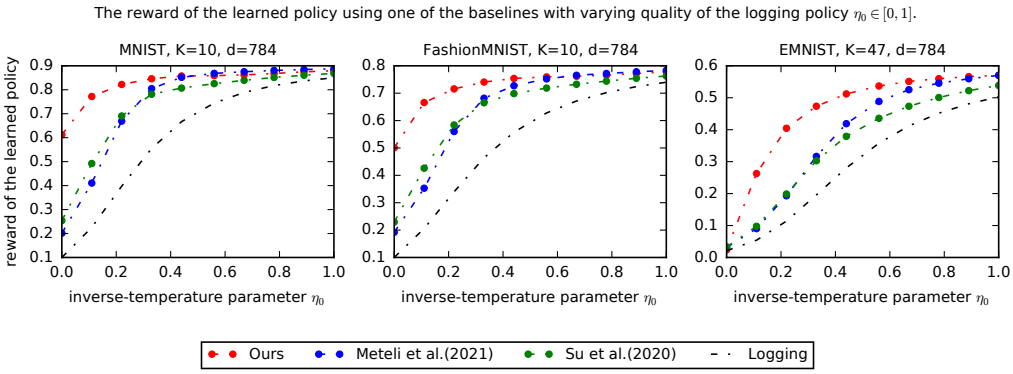

Figure 7. The reward of the learned policy with varying quality of the logging policy $\eta_0 \in [0, 1]$ using either our regularization (α-IPS) or the ones in Su et al. (2020); Metelli et al. (2021).