Online Learning with Off-Policy Feedback

Germano Gabbianelli GERMANO.GABBIANELLI@UPF.EDU
Gergely Neu GERGELY.NEU@GMAIL.COM
Matteo Papini MATTEO.PAPINI@UPF.EDU
Universitat Pompeu Fabra, Barcelona, Spain

Editors: Shipra Agrawal and Francesco Orabona

Abstract

We study the problem of online learning in adversarial bandit problems under a partial observability model called off-policy feedback. In this sequential decision making problem, the learner cannot directly observe its rewards, but instead sees the ones obtained by another unknown policy run in parallel (behavior policy). Instead of a standard exploration-exploitation dilemma, the learner has to face another challenge in this setting: due to limited observations outside of their control, the learner may not be able to estimate the value of each policy equally well. To address this issue, we propose a set of algorithms that guarantee regret bounds that scale with a natural notion of mismatch between any comparator policy and the behavior policy, achieving improved performance against comparators that are well-covered by the observations. We also provide an extension to the setting of adversarial linear contextual bandits, and verify the theoretical guarantees via a set of experiments. Our key algorithmic idea is adapting the notion of pessimistic reward estimators that has been recently popular in the context of off-policy reinforcement learning.

Keywords: online learning, off-policy, partial monitoring, bandit problems

1. Introduction

Off-policy learning is one of the most fundamental concepts in reinforcement learning, concerned with the problem of learning an optimal behavior policy given sample observations generated by a (most likely suboptimal) behavior policy. This setting comes with a unique set of challenges arising from the fact that the learning agent has no influence over the observed data, and thus classical methods for reducing uncertainty via exploration do not directly apply. The inability to explore may suggest that off-policy learning is better approached as a simple "pure exploitation" problem and can be potentially solved by a greedy approach—however, more thought reveals that an effective learning method should also attempt to account for the uncertainty of the random observations. Indeed, the problem setting comes with multiple layers of uncertainty: one layer being the potentially random choices made by the behavior policy, and another being the randomness in the observed rewards. In the present paper, we study a setting where the two uncertainties can be decoupled and addressed individually: the setting of online learning against an adversarial sequence of rewards, with off-policy feedback revealed by a stationary random policy.

The setting we study lies in the intersection of two distinct paradigms of sequential decision making: adversarial online learning and off-policy reinforcement learning. Concretely, we study a sequential decision making problem where in each round, the learner has to pick one of K actions in order to maximize its total rewards. The sequence of reward assignments to actions are decided by an adversary, with each reward function determined the moment before the learner selects its action. The unique feature of the setting is that the learner does not get to observe its reward. However,

the learner does observe the reward of another action that has been randomly sampled according to a behavior policy that remains fixed during the learning process. The goal of the learner is then to gain nearly as much reward as the best fixed comparator policy.

A concrete motivating example is the following. Consider running a large online advertisement company with a well-established system that is deployed on most of the traffic. The infrastructure of the company allows real-time measurements of the clickthrough rates generated by this system. Now, imagine that the research division is given access to some small amount of traffic where a new recommendation method can be deployed, but real-time logging is not reliable due to the lower volume of traffic assigned for experimentation. Thus, the decisions of the experimental recommendation system have to be driven by the real-time logs obtained from the original system on the main traffic, which may have poor coverage of some good actions that the new system can implement. In this example, the original system corresponds to the behavior policy and the experimental system corresponds to the policy of the learner.

While there is indeed no exploration-exploitation dilemma that the learner has to address in this setting, some precaution in selecting the actions is still needed due to the potentially malicious choices of the adversary. Another, perhaps bigger, challenge is that the learner has to estimate the rewards of its own actions from observations made by another policy. The rewards of actions that are observed less frequently are obviously more difficult to estimate. It is thus a sensible requirement to have better performance guarantees against actions that have been more frequently played than against less well-understood actions. More generally, we aim to obtain guarantees that depend on how well the comparator policy is "covered" by the behavior policy.

Our main contribution is an online learning algorithm that guarantees a total expected regret against any comparator policy π^* that is of order $\sqrt{n}\sum_a \frac{\pi^*(a)}{\pi_B(a)}$, where π_B is the behavior policy and $\pi(a)$ denotes the probability that policy π plays action a. Our method makes use of a slight *pessimistic* adjustment to the classic importance-weighted reward estimators commonly used in the adversarial bandit literature. We refer to the problem-dependent factor appearing in the bound as the *coverage ratio* and denote it by $C(\pi^*;\pi_B)$. The coverage ratio quantifies the overlap between the comparator and behavior policies: it is of order K when the two policies closely match each other, but it blows up quickly as the two policies start to differ. Notably, our bounds can be orders of magnitude better than what one would obtain by adapting a standard adversarial bandit method without adjustments. For instance, a naïve analysis of the classic EXP3 method only gives a regret bound of order $\sqrt{n/\min_a \pi_B(a)}$ against all comparator policies—even against ones that are actually well covered by the behavior policy. Besides providing theoretical results, we also confirm empirically that the performance of these two methods can be quite different, and in particular that EXP3 can indeed fail to take advantage of the comparator policy being well covered by the behavior policy.

Moreover, our contributions naturally fit in the broader context of online learning under partial monitoring, which generally considers situations where the observations made by the learner are decoupled from its rewards (Rustichini, 1999; Bartók et al., 2014; Lattimore and Szepesvári, 2019). In a general partial monitoring scenario, the learner receives an observation that depends on its action but may be insufficient to reconstruct the obtained reward. A well-studied special case of partial monitoring problems is online learning with feedback graphs (Mannor and Shamir, 2011; Alon et al., 2015, 2017; Kocák et al., 2014, 2016). In this setting, the set of observations associated with each action are given by a directed graph whose nodes are the actions: if actions a and a' are connected with an arc pointing from a to a', the learner observes the reward of action a when it plays action a'. The graph may not have self-loops for every action, which allows the possibility

that the learner will not observe its own reward. Clearly, our setting can be embedded in this class of problems by considering a sequence of randomly generated star graphs where the action taken by the behavior policy is connected with all other actions. However, the graph does not contain self-loops which renders all existing methods for this problem unsuitable for our problem. In this sense, our contribution sheds some new light on the hardness of learning with feedback graphs without self-loops, and can potentially inspire future work in this domain.

Another line of work closely related to ours is the literature on offline reinforcement learning, where the learner cannot interact with the environment and has instead only access to a fixed dataset gathered by a behavior policy (Levine et al., 2020). In this context, the idea of of employing some form of pessimism has been extremely popular in the last few years, and pessimism has been purported to come with many desirable properties (Jin et al., 2021; Buckman et al., 2021; Uehara and Sun, 2021; Rashidinejad et al., 2021; Xie et al., 2021). One of these is that pessimistic offline RL methods can overcome the typical limitation of requiring the behavior policy to sufficiently explore the *whole* state-action space, which many previous results suffer from (Antos et al., 2008; Munos and Szepesvári, 2008; Chen and Jiang, 2019; Xie and Jiang, 2021). This assumption is very strong and often not verified in practice. However, a series of recent works show that, via an appropriate use of pessimism, it is possible to obtain bounds which scale with the coverage with respect to a comparator policy, instead of the whole state-action space. Many of these results are surveyed in the work of Xiao et al. (2021), who show that pessimistic policies are minimax optimal with respect to a special objective that weighs problem instances with a notion of inherent difficulty of estimating the value of the optimal policy. On the other hand, they show that without such weighting, pessimism is in fact only one of many possible heuristics that are all minimax optimal when considering the natural version of the optimization objective. This highlights that pessimism may not necessarily play a special role in offline optimization, and that the quest to understand the complexity of offline reinforcement learning is far from being over.

While our results definitely do not settle the debate of whether or not pessimism is the best way to deal with off-policy observations, they do provide some new insights. Most importantly, our findings highlight that pessimism remains an effective method for obtaining comparator-dependent guarantees. Such guarantees have attracted quite some interest in the literature on online learning with fully observable outcomes Chaudhuri et al. (2009); Koolen (2013); Luo and Schapire (2015); Koolen and van Erven (2015); Orabona and Pál (2016); Cutkosky and Orabona (2018). One common building block of parameter-free methods in this context is the PROD algorithm of Cesa-Bianchi et al. (2007), used, for instance, in the algorithm designs of Sani et al. (2014); Gaillard et al. (2014); Koolen and van Erven (2015). Interestingly, our analysis also leans heavily on the tools developed by Cesa-Bianchi et al. (2007). When it comes to the bandit setting, comparator-dependent results are apparently much more sparse and in fact we are only aware of the work of Lattimore (2015) that studies the possibility of guaranteeing better performance against certain comparators. As for our specific problem, we are not aware of any existing method that would be able to guarantee meaningful instance-dependent performance bounds.

Notation. We denote the set of probability distributions over a set $\mathcal{S}$ as $\Delta_{\mathcal{S}}$. We denote the scalar product of $x, y \in \mathbb{R}^d$ as $\langle x, y \rangle$ and use $\|\cdot\|_2$ to denote the Euclidean norm. For a positive semi-definite matrix $A \in \mathbb{R}^{d\times d}$, we write $\lambda_{\min}(A)$ and $\operatorname{Tr}(A)$ to denote respectively its smallest eigenvalue and its trace. Finally, we use the conventions that $\prod_{k=i}^{j} = 1$ and $\sum_{k=i}^{j} = 0$ when $j < i$.

2. Preliminaries

We study an n-rounds sequential-decision game between a *learner* (who has a finite set of actions $\mathcal{A}$) and an *adversary*, where the following steps are repeated in each round $t \in [n]$:

1. The adversary picks a reward function $r_t : \mathcal{A} \to [0, 1]$, mapping each action to a nonnegative numerical reward,
2. an action $A_t \in \mathcal{A}$ is selected by the learner,
3. another action $A_t^B \in \mathcal{A}$ is selected according to a fixed *behavior policy* π_B,
4. the learner gains reward $R_t = r_t(A_t)$ and observes $R_t^B = r_t(A_t^B)$ along with A_t^B.

Notably, the learner does not get to observe its own reward R_t but has to make do with the reward R_t^B gained by the behavior policy. We allow the adversary to be adaptive in the sense of being able to take into account all past actions of the learner and the behavior policy when selecting the reward function. Also, the learner is allowed to use randomization for selecting its action. Precisely, we will denote the history of interactions up to the end of round t by $\mathcal{F}_t = \sigma\left(r_t, A_t^B, A_t, \ldots, r_1, A_1^B, A_1\right)$, and respectively denote conditional expectations and probabilities with respect to the induced filtration by $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot|\,\mathcal{F}_{t-1}]$ and $\mathbb{P}_t[\cdot] = \mathbb{P}[\cdot|\,\mathcal{F}_{t-1}]$. With this notation, we define the *policy* of the learner as $\pi_t(a) = \mathbb{P}_t[A_t = a]$ for all actions $a \in \mathcal{A}$.

The objective of the learner is to minimize the *regret*, with respect to any time-invariant comparator policy $\pi^* \in \Delta_{\mathcal{A}}$, defined as

$$\mathcal{R}(\pi^*) = \sum_{t=1}^{n} \sum_{a} \big(\pi^*(a) - \pi_t(a)\big) r_t(a). \tag{1}$$

In particular we will be interested in providing bounds on $\mathbb{E}[\mathcal{R}(\pi^*)]$, where the expectation integrates over all the randomness involved in selecting the random actions A_t and A_t^B and the reward functions r_t. In words, the expected regret measures the expected gap between the total rewards gained by the learner and the amount gained by a fixed comparator policy π^*.

The most common definition of regret compares the learner's performance to the optimal policy π^* that selects the action $a^* = \arg\max_a \sum_{t=1}^{n} r_t(a)$. However, it is easy to see that this comparator strategy may be unsuitable for measuring performance in the setting we consider. Specifically, it is unreasonable to expect strong guarantees against the optimal policy when the behavior policy selects the optimal actions very rarely. Specifically, the adversary can take advantage of the behavior policy covering the action space only partially, and hide the best rewards among the least-frequently sampled actions. In the most extreme case, the behavior policy may not select some actions at all, which clearly makes it impossible for the learner to compete with the optimal policy. Thus, we aim to achieve regret guarantees that scale with the level of mismatch between the behavior and comparator policies, capturing the intuition that comparator strategies that are well covered by the data should be easier to compete with. Concretely, we will define the coverage ratio between π^* and π_B as

$$C(\pi^*; \pi_B) = \sum_{a} \frac{\pi^*(a)}{\pi_B(a)}, \tag{2}$$

and aim to provide regret bounds that scale with this quantity. The intuitive significance of this coverage ratio is that it roughly captures the hardness of estimating the value of the comparator

policy π^* using only data from π_B. Indeed, a simple argument reveals that the estimation error of the total reward of any given action a scales as $\sqrt{n/\pi_B(a)}$ in the worst case. Thus, we set out to prove regret guarantees against each comparator π^* that scale proportionally to the worst-case estimation error of order $\sqrt{C(\pi^*;\pi_B)n}$.

3. Algorithm and main results

This section presents our main contributions: a set of algorithms for online off-policy learning and their comparator-dependent performance guarantees that scale with the coverage ratio between the comparator policy and the behavior policy. For the sake of clarity of exposition, we first describe our approach in a relatively simple setting where the number of actions is finite and the behavior policy is known. We then extend the algorithm to be able to deal with unknown behavior policies in Section 3.2 and to linear contextual bandit problems in Section 3.3.

Algorithm 1 EXP3-IX for Off-Policy Learning

Input: learning rate η, IX parameters $(\gamma_t)_{t=1}^n$
for $t \leftarrow 1, \dots, n$ **do**
  compute $w_t(a) = \exp(\eta \sum_{k=1}^{t-1} \widetilde{r}_k(a)) \quad \forall a \in \mathcal{A}$
  play A_t according to $\pi_t(\cdot) = w_t(\cdot) / \sum_{a\in\mathcal{A}} w_t(a)$
  observe A_t^B and R_t^B
  compute $\widetilde{r}_t(A_t^B) = R_t^B / \left(\pi_B(A_t^B) + \gamma_t\right)$
end for

3.1. Known behavior policy

Let us first consider the case where the learner has full prior knowledge of π_B. The algorithm we propose is an adaptation of the EXP3-IX algorithm first proposed by Kocák et al. (2014) and later analyzed more generally by Neu (2015). At each time-step t the algorithm computes the weights

$$w_1(a) = 1,$$
$$w_t(a) = w_{t-1}(a) e^{\eta \widetilde{r}_{t-1}(a)} = \exp\left(\eta \sum_{k=1}^{t-1} \widetilde{r}_k(a)\right),$$

and the normalization factors $W_t = \sum_a w_t(a)$, and uses them to draw the action A_t according to $\pi_t(a) = \frac{w_t(a)}{W_t}$. Here, η is a positive learning-rate parameter and $\widetilde{r}$ is the *Implicit eXploration* (IX) estimate of the reward function r_t, modified to use the rewards obtained by the behavior policy π_B, since the learner cannot see its own rewards:

$$\widetilde{r}_t(a) = \frac{R_t^B \mathbb{1}\{A_t^B = a\}}{\pi_B(a) + \gamma_t} = \frac{r_t(a)\mathbb{1}\{A_t^B = a\}}{\pi_B(a) + \gamma_t}, \tag{3}$$

where $\gamma_t \geq 0$ is an appropriately chosen parameter. The full algorithm is shown as Algorithm 1 .

When setting $\gamma_t = 0$, $\widetilde{r}_t$ is clearly an unbiased estimator of r_t since $\mathbb{E}_t\left[\mathbb{1}\{A_t^B = a\}\right] = \pi_B(a)$. Otherwise, for $\gamma_t > 0$, the estimator is biased towards zero which can be seen as a *pessimistic* bias in the sense that it underestimates the true rewards: $\mathbb{E}_t\left[\widetilde{r}_t(a)\right] \leq r_t(a)$. This property is crucially

important to achieve our goal to obtain performance guarantees that scale with the mismatch between π^* and π_B. We believe that this use of the IX estimator with positive rewards is novel. In previous work, the IX estimator has been used with losses, resulting in an optimistic bias, exploited, for example, by the high-probability analysis of Neu (2015). It is far from obvious that our alternative usage of the IX estimator would induce the right notion of pessimism needed for achieving coverage-dependent results in the off-policy setting. This is established in the following:

Theorem 1 *For any comparator policy π^*, the expected regret of* EXP3-IX *initialized with any positive learning rate η and $\gamma_t = \frac{\eta}{2}$, is bounded as*

$$\mathbb{E}[\mathcal{R}(\pi^*)] \leq \frac{\log K}{\eta} + \frac{\eta}{2} \sum_{t=1}^{n} \sum_{a} \frac{r_t(a)\pi^*(a)}{\pi_B(a) + \frac{\eta}{2}}. \tag{4}$$

Setting the learning rate to $\eta = \sqrt{\frac{\log K}{n}}$ and to $\eta = \sqrt{\frac{\log K}{C(\pi^*;\pi_B)n}}$ respectively gives

$$\mathbb{E}[\mathcal{R}(\pi^*)] \leq \sqrt{n \log K} \left(1 + \frac{1}{2} C(\pi^*; \pi_B)\right) \tag{5}$$

$$\mathbb{E}[\mathcal{R}(\pi^*)] \leq \sqrt{2C(\pi^*; \pi_B) n \log K}. \tag{6}$$

The proof is based on a set of small but important changes made to the standard EXP3 analysis originally due to Auer et al. (2002), and is deferred to Section 4. The bound above successfully achieves our goal of guaranteeing better regret against comparator policies that are well-covered by the behavior policy. In particular, the first bound of Equation (5) provides a bound that holds uniformly for all behavior policies without requiring prior commitment to any coverage level, whereas the second bound guarantees improved guarantees against policies with a given coverage level at the price of using a learning-rate parameter that is specific to the desired coverage. Notably, the coverage ratio is of the order K when the comparator policy closely matches the behavior policy, but the actual bound of Equation (4) can be much smaller when there are many actions that the behavior policy selects with probability much smaller than γ.

It is worthwhile to compare this result with what one would obtain by a straightforward adaptation of a standard adversarial bandit algorithm like EXP3 (Auer et al., 2002)—which essentially corresponds to our algorithm with the choice $\gamma = 0$. A standard calculation shows that the regret of this strategy can be upper bounded by

$$\mathbb{E}[\mathcal{R}(\pi^*)] \leq \frac{\log K}{\eta} + \eta \sum_{t=1}^{n} \sum_{a} \mathbb{E}\left[\frac{\pi_t(a)}{\pi_B(a)}\right].$$

Notice that the right-hand side of this bound does not depend on the comparator policy, which suggests that this method is not quite suitable for achieving our goal. Even worse, the only way to bound the second term in the bound seems to be by $n/\min_a \pi_B(a)$, which scales inversely with the coverage of the least well-covered action. A pessimistic interpretation of this argument suggests that EXP3 may have huge regret when some actions are not covered appropriately. A more charitable reading is that EXP3 may not be able to take advantage of situations where the comparator policy is well-covered by the behavior policy. We set out to understand this phenomenon empirically in Section 5.

The results of Theorem 1 could be extended to deal with a nonstationary sequence of behavior policies, and the regret bound can be shown to scale with the average of the coverage ratios, as long as the behavior policies are revealed to the learner.

3.2. Unknown behavior policy

In the previous section we assumed to have full prior knowledge of the behavior policy π_B in order to compute our reward estimator $\widehat{r}_t$. In this section, we show that this is not an inherent limitation of our technique and that it can be easily addressed by using a simple plugin estimator $\widehat{\pi}_t$ of the behavior policy, which is then used in the definition of $\widetilde{r}_t$:

$$\widehat{\pi}_1(\cdot) = 0, \qquad \widehat{\pi}_t(a) = \frac{1}{t-1}\sum_{k=1}^{t-1} \mathbb{1}\{A_k^B = a\}, \tag{7}$$

$$\widetilde{r}_t(a) = \frac{r_t(a)\mathbb{1}\{A_t^B = a\}}{\widehat{\pi}_t(a) + \gamma_t}. \tag{8}$$

We then feed these reward estimates to the exponential-weights procedure described in the previous section. As the following theorem shows, the resulting algorithm satisfies essentially the same regret bound as the method that has full knowledge of π_B.

Theorem 2 *For any comparator policy π^*, the expected regret of* EXP3-IX *with learning rate $\eta = \sqrt{\log(K)/n}$ and parameter sequence $\gamma_1 = 1 + \frac{\eta}{2}$, $\gamma_t = \frac{\eta}{2} + \sqrt{\log(K(t-1)^2)/(2t-2)}$, and estimates as in Equation (7), is bounded as*

$$\mathbb{E}[\mathcal{R}(\pi^*)] = \mathcal{O}\left(C(\pi^*;\pi_B)\sqrt{n\log(Kn)}\right). \tag{9}$$

The parameter tuning achieving the above bound is similar to what is used in the previous theorem, and does not require the learner to have any problem-specific information that would be difficult to acquire. Details are relegated to Appendix B along with the proof of the theorem.

3.3. Linear contextual bandits

We now switch gears and provide an extension to a significantly more advanced setup: that of adversarial linear contextual bandits, first studied by Neu and Olkhovskaya (2020). In each round t of this sequential game, the learner first observes a context X_t before making its decision, and the reward function r_t is assumed to be an adversarially chosen function of the context X_t and the action A_t taken by the learner. In particular, the adversary chooses a *reward vector* $\theta_t \in \mathbb{R}^d$ at each step, which determines the rewards for each context-action pair as $r_t(x,a) = \langle \theta_t, \varphi(x,a)\rangle$, where $\varphi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ is a *feature map* known to both the learner and the adversary. We assume that the contexts live in an abstract space $\mathcal{X}$ and are drawn i.i.d. according to a fixed probability distribution for all t. On the other hand, the adversary has full freedom in choosing the reward functions, as long as it only depends on past observations and in particular does not depend on X_t or A_t. The only restriction we put on the adversary is that we continue to require the rewards to be in the interval $[0,1]$. Moreover, as in the previous section the learner is not allowed to see its own rewards, but only the ones of an other policy π_B running in parallel. In this setting, a policy π is a mapping from contexts to probability distributions over the space of actions.

The objective of the learner is to minimize the regret defined with respect to any time-invariant comparator policy $\pi^* : \mathcal{X} \to \Delta_{\mathcal{A}}$ as:

$$\mathcal{R}(\pi^*) = \sum_{t=1}^{n}\sum_{a}\left(\pi^*(a|X_t) - \pi_t(a|X_t)\right) r_t(X_t,a).$$

Our algorithm for this setting is a combination of the context-wise exponential weights method proposed by Neu and Olkhovskaya (2020) with the ideas developed in the previous section. The algorithm design is complicated by the fact that the implicit exploration estimator is not very straightforward to extend to this setting, which necessitates an alternative, but closely related, approach. In particular, we will define an *unbiased* estimator of the reward vector θ_t and feed the resulting reward estimates to calculate policy updates via the PROD update rule proposed by Cesa-Bianchi et al. (2007) (see also Cesa-Bianchi and Lugosi (2006), Section 2.7).

Concretely, following the algorithm design of Neu and Olkhovskaya (2020), we define the matrix $\overline{V}(\pi) = \mathbb{E}_t\left[\sum_a \pi(a|X_t)\varphi(X_t,a)\varphi(X_t,a)^\intercal\right]$ and the estimator

$$\widehat{\theta}_t = \left(\overline{V}(\pi_B)\right)^{-1}\varphi(X_t, A_t^B)R_t^B. \tag{10}$$

Since $\varphi(X_t, A_t^B)R_t^B = \varphi(X_t, A_t^B)\varphi(X_t, A_t^B)^\intercal\theta_t$, it is easy to see that $\widehat{\theta}_t$ is an unbiased estimator of θ_t. These estimators are then used to update a set of weights w_t defined for each context-action pair as

$$w_t(x,a) = \prod_{k=1}^{t-1}(1+\eta\langle\widehat{\theta}_k, \varphi(x,a)\rangle), \tag{11}$$

$$W_t(x) = \sum_a w_t(x,a), \tag{12}$$

and the policy is then given as $\pi_t(a|x) = \frac{w_t(x,a)}{W_t(x)}$. Notice that this policy can be also seen as another form of pessimistic reward estimation. Infact, letting $\hat{r}_t$ be an unbiased reward estimator, the PROD-style update of Equation (11) can be seen as an EXP3 update with the modified reward estimator $\widetilde{r}_t = \frac{1}{\eta}\log(1+\eta\hat{r}_t)$, which clearly lower bounds the original reward estimator throught the inequality $\log(1+z) \le z$, corresponding to a form of pessimism. Moreover, the policy can be easily implemented without explicitly keeping track of the weights for all (x,a) pairs, as they are well-defined through the sequence of reward-estimate vectors $\widehat{\theta}_1, \dots, \widehat{\theta}_{t-1}$. For simplicity[1], we assume that the matrix $\overline{V}(\pi_B)$ is known for the learner and has uniformly lower-bounded eigenvalues so that its inverse exists. Following the naming convention of Neu and Olkhovskaya (2020), we call the resulting algorithm LINPROD and show its pseudocode in Algorithm 2. Similarly to the previous sections, we are aiming for a comparator-dependent performance guarantee that depends on the

Algorithm 2 LINPROD for off-policy learning

Input: learning rate η
for $t \leftarrow 1, \dots, n$ **do**
  observe X_t
  compute $w_t(X_t, \cdot) = \prod_{k=1}^{t-1}\left(1+\eta\langle\widehat{\theta}_k, \varphi(X_t,\cdot)\rangle\right)$
  draw A_t from $\pi_t(\cdot|X_t) = w_t(X_t,\cdot)/\sum_a w_t(X_t,a)$
  observe R_t^B and $\varphi(X_t, A_t^B)$
  compute $\widehat{\theta}_t$ as in Equation (10)
end for

1. These restrictions can be removed using the techniques developed in the previous section, although at the price of a significantly more technical analysis. We opted to preserve clarity of presentation instead.

mismatch of the comparator and the behavior policy. However, this quantity is not straightforward to define in the case that we consider, due to the fact that we consider a potentially infinite space of contexts. In particular, the natural idea of considering $\mathbb{E}\left[\sum_a \frac{\pi^*(a|X_t)}{\pi_B(a|X_t)}\right]$ as a measure of mismatch is problematic as it can blow up when there exists even a tiny set of contexts where the two policies have no overlap. Intuitively, it should be possible to estimate the reward vector even when there are states where the policies pick different actions, as long as they are aligned in the feature space in an appropriate sense. To make this intuition formal, we will consider the following alternative notion of *feature coverage ratio*:

$$C_\varphi(\pi^*;\pi_B) = \mathrm{Tr}\left[\left(\overline{V}(\pi_B)\right)^{-1}\overline{V}(\pi^*)\right]. \tag{13}$$

This notion of coverage appropriately measures the extent to which the feature vectors $\varphi(X_t, A_t^*)$ generated by the comparator policy line up with the features excited by the behavior policy. Similar distribution-mismatch measures are common in the offline RL literature, and in particular the results of Jin et al. (2021) are stated in terms of the same quantity. The following theorem gives a performance guarantee stated in terms of this measure of distribution mismatch.

Theorem 3 *Let η be any positive learning rate and suppose that it is small enough so that $\lambda_{\min}(\overline{V}(\pi_B)) \geq 2\eta \sup_{x,a} \|\phi(x,a)\|_2^2$ holds. Then, for any comparator policy π^* the expected regret of* LINPROD *is upper-bounded by*

$$\mathbb{E}\left[\mathcal{R}(\pi^*)\right] \leq \frac{\log K}{\eta} + \eta n C_\varphi(\pi^*;\pi_B),$$

Setting $\eta = \sqrt{\frac{\log K}{n}}$ and $\eta = \sqrt{\frac{\log K}{C_\varphi(\pi^*;\pi_B)n}}$ and supposing that n is large enough so that η satisfies the condition, the regret can be further bounded respectively as

$$\mathbb{E}[\mathcal{R}(\pi^*)] \leq \sqrt{n\log K}\left(1 + C_\varphi(\pi^*;\pi_B)\right),$$
$$\mathbb{E}[\mathcal{R}(\pi^*)] \leq 2\sqrt{C_\varphi(\pi^*;\pi_B) n \log K}.$$

The bound mirrors the qualities of Theorem 1, and in particular it implies good performance when the comparator policy is well-covered by the behavior policy. Under ideal conditions where these policies are close enough, the coverage ratio is of order d, which essentially matches the rate proved by Neu and Olkhovskaya (2020) for the case of standard bandit feedback. The bound then degrades as the two policies drift apart. We recover the best-known bounds for the stochastic setting (Jin et al., 2021). The latter were stated for the setting of off-policy learning in linear MDPs, which includes the stochastic version of our problem as a special case. Note that our algorithm requires knowledge of $\overline{V}(\pi_B)$. However, provided that the context distribution is known, it is possible to use instead an estimate based on matrix geometric resampling, as proposed by Neu and Olkhovskaya (2020).

4. Analysis

This section provides the key ideas required for proving our main results. Due to space restrictions, we will only prove Theorem 1 here and defer the proof of the other two theorems to Appendices B and C.

For the analysis, it will be useful to define the unbiased reward estimator $\widehat{r}_t(a) = \frac{r_t(a)\mathbb{1}\{A_t^B = a\}}{\pi_B(a)}$, which essentially corresponds to the biased IX estimator $\widetilde{r}_t$ when setting $\gamma_t = 0$. One of the

most important properties of the IX estimator that we will repeatedly use is stated in the following inequality:

$$\frac{r_t(a)\mathbb{1}\{A_t^B = a\}}{\pi_B(a) + \gamma_t} \leq \frac{1}{2\gamma_t} \log\left(1 + 2\gamma_t \widehat{r}_t(a)\right). \tag{14}$$

The result follows from a simple calculation in the proof of Lemma 1 of Neu (2015) that we reproduce in Appendix A.1 for the convenience of the reader. Notably, the term on the right hand side can be thought of as a reward estimator itself. Combining this reward estimator with the exponential weights policy with $\eta = \gamma$ gives rise to the PROD algorithm of Cesa-Bianchi et al. (2007), which is a fact that some of our proofs will implicitly take advantage of. This observation also motivates our algorithm design for the contextual bandit setting in Section 3.3.

The proof of Theorem 1 The proof builds on the classical analysis of exponential weights algorithm originally due to Vovk (1990), Littlestone and Warmuth (1994) and Freund and Schapire (1997), and its extension to adversarial bandit problems by Auer et al. (2002). In particular, our starting point is the following lemma that can be proved directly with arguments borrowed from any of these past works:

Lemma 4

$$\sum_{t=1}^{n} \sum_{a} \pi^*(a)\widetilde{r}_t(a) \leq \frac{\log K}{\eta} + \frac{1}{\eta} \sum_{t=1}^{n} \log \sum_{a} \pi_t(a) \exp\left(\eta \widetilde{r}_t(a)\right).$$

We include the proof for the sake of completeness in Appendix A.2. To proceed, notice that the above bound can be combined with Equation (14) to obtain

$$\begin{aligned}
\sum_{t=1}^{n} \sum_{a} \pi^*(a)\widetilde{r}_t(a) &\leq \frac{\log K}{\eta} + \frac{1}{\eta} \sum_{t=1}^{n} \log \sum_{a} \pi_t(a) \exp\left(\frac{\eta}{2\gamma} \log\left(1 + 2\gamma\, \widehat{r}_t(a)\right)\right) \\
&= \frac{\log K}{\eta} + \frac{1}{\eta} \sum_{t=1}^{n} \log \sum_{a} \pi_t(a) \left(1 + \eta \widehat{r}_t(a)\right) \\
&\leq \frac{\log K}{\eta} + \sum_{t=1}^{n} \sum_{a} \pi_t(a) \widehat{r}_t(a),
\end{aligned} \tag{15}$$

where we used the choice $\gamma = \eta/2$ in the second line and the inequality $\log(1 + x) \leq x$ that holds for all $x > -1$ in the last line.

It remains to relate the two sums in the above expression to the total reward of the learner and the comparator policy. To this end, we first notice that for any given action a, we have

$$\begin{aligned}
\mathbb{E}_t\left[\widetilde{r}_t(a)\right] &= \mathbb{E}_t\left[\frac{r_t(a)\mathbb{1}\{A_t^B = a\}}{\pi_B(a) + \gamma}\right] \\
&= \frac{r_t(a)\pi_B(a)}{\pi_B(a) + \gamma} = r_t(a) - \frac{\gamma r_t(a)}{\pi_B(a) + \gamma}.
\end{aligned} \tag{16}$$

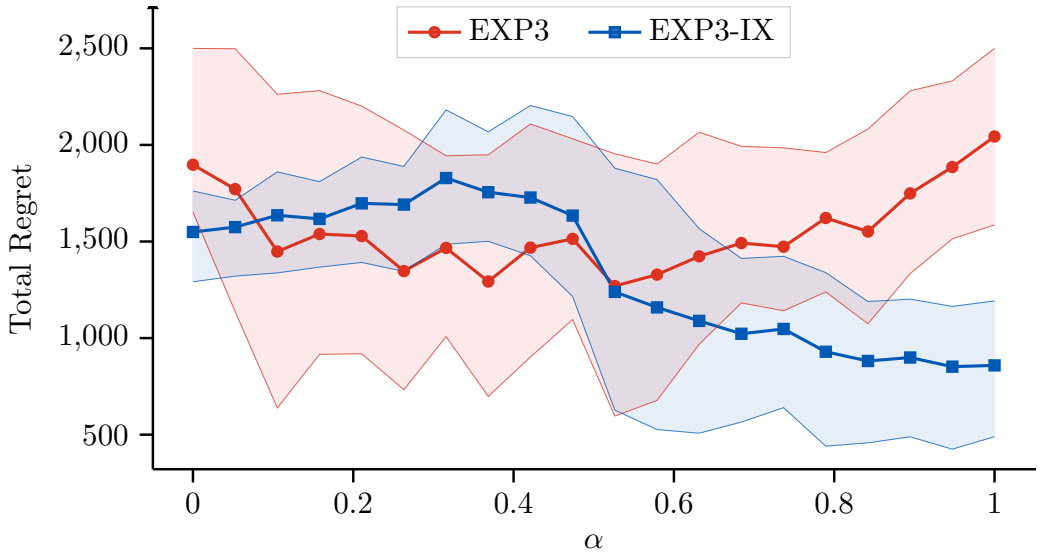

Figure 1: Total regret after 1000 steps for different values of the interpolation parameter α. Thick lines represent the mean regret over 100 independent runs, while the shaded area represents the interval between the 25% and 75% quantiles.

Via the tower rule of expectation, this implies

$$\mathbb{E}\left[\sum_{t=1}^{n}\sum_{a}\pi^*(a)\widetilde{r}_t(a)\right] = \mathbb{E}\left[\sum_{t=1}^{n}\sum_{a}\pi^*(a)r_t(a)\right] - \gamma\sum_{t=1}^{n}\sum_{a}\frac{\pi^*(a)r_t(a)}{\pi_B(a)+\gamma}.$$

Similarly, since $\mathbb{E}_t\left[\widehat{r}_t(a)\right] = r_t(a)$, we also have

$$\mathbb{E}\left[\sum_{t=1}^{n}\sum_{a}\pi_t(a)\widehat{r}_t(a)\right] = \mathbb{E}\left[\sum_{t=1}^{n}\sum_{a}\pi_t(a)r_t(a)\right].$$

Putting these two facts together with Equation (15), we obtain the result claimed in the theorem.

5. Empirical Results

The goal of this section is to compare the performances of EXP3 and EXP3-IX under different levels of coverage, and verify if indeed our method outperforms EXP3 in situations where the behavior policy is well-aligned with the comparator, as the theory suggests. As pointed out earlier, a naïve analysis of EXP3 suggests that its regret may scale as $\sqrt{n/\min_a \pi_B(a)}$ in the worst case, and therefore may not be able to take advantage of situations where the behavior policy is well aligned with the comparator. Our proposed method, instead, should be able to do so since it enjoys comparator-dependent bounds.

We instantiate a 100-armed bandit, with Bernoulli rewards for all arms. By default, all rewards have mean 0.5. However, for the first half of the game ($t \leq n/2$), we change the mean reward of the last arm to 0.8, and for the remaining half, the mean of the first arm to 1. Suboptimal arms always have the default mean reward of 0.5. This means that arm 100 is the best for the first half of the game, but eventually gets outperformed by arm 1. We set the number of rounds n to 10000, the learning rate η of both algorithms to the recommended $\sqrt{\log(K)/n}$ and $\gamma_t = \eta/2$. We repeat the game for a range of behavior policies defined for each α as $\pi_{B,\alpha}(i) \propto (1-\alpha)\frac{i}{K} + \alpha(1 - \frac{i-1}{K})$, for $i \in [0, \dots, K]$, where α varies from 0 to 1. Hence, α closer to 1 means the behavior policy puts large probability mass on the first action, which we use as the comparator in our experiment. We plot the results of the experiment on Figure 5.

The results clearly match the intuitions that one can derive from our performance guarantees: the regret of EXP3 indeed deteriorates as $\min_a \pi_B(a)$ approaches 0 at the two extremes $\alpha = 0$ and $\alpha = 1$. In particular, EXP3 fails to take advantage of the favorable case where the optimal policy is well covered, while EXP3-IX performs significantly better in the latter case, as predicted by our theory.

Moreover, it is worth to note that EXP3-IX was originally proposed, in the adversarial bandit literature, as a variant of EXP3 with lower variance, allowing to bound regret with high probability instead of merely in expectation. This variance reduction effect clearly carries over to our setting. However, we were not able to establish high-probability bounds for the adversarial-off-policy setting so far, and leave this question open for future research.

6. Conclusion

We introduced a new online learning setting where the learner is only allowed to observe off-policy feedback generated by a fixed behavior policy. We have proposed an algorithm with comparator-dependent regret bounds of order $C(\pi^*; \pi_B)\sqrt{n}$, depending on a naturally defined coverage ratio parameter $C(\pi^*; \pi_B)$ that characterizes the mismatch between the behavior and the comparator policies. Many questions remain open regarding the potential tightness of this result. First, we have shown that the bounds can be improved to $O(\sqrt{C(\pi^*; \pi_B)n})$, if one wishes to restrict their attention to comparators whose coverage level is at a fixed level $C(\pi^*; \pi_B)$. However, the tuning required for achieving this result depends on the desired coverage level. It is an interesting open problem to find out if this requirement can be relaxed, and bounds of order $\sqrt{C(\pi^*; \pi_B)n}$ can be simultaneously achieved for all comparators π^* by a single algorithm. We conjecture that this question can be addressed by a careful adaptation of existing techniques for adaptive online learning, and in particular we believe that adapting the methodology of Koolen and van Erven (2015) should be especially suitable for achieving this goal.

Questions regarding the best achievable performances for our newly defined problem are even more exciting. As an adaptation of the results of Xiao et al. (2021) show via an online-to-batch reduction, the minimax regret of any algorithm for this setting has to scale as $\sqrt{n/\min_a \pi_B(a)}$, suggesting that our naïve adaptation of EXP3 is already minimax optimal. In our view, this makes it all the more interesting to identify characteristics of individual problem instances that make faster learning possible, and we believe that comparator-dependent regret bounds scaling with the coverage ratio are only one of many possible flavors of adaptive performance guarantees. One concrete question that we are particularly interested in is a better understanding of the "Pareto regret frontier" of achievable regrets, roughly corresponding to the set of comparator-dependent regret bounds that

are achievable by any algorithm. Clearly, the bounds we achieve are just singular elements of this set. We conjecture that bounds of order $\sqrt{C(\pi^*;\pi_B)n}$ are indeed on the regret frontier. Whether this is indeed true or if there are other distinguished entries on the Pareto frontier with desirable properties remains to be seen. All in all, our results highlight that off-policy learning is a field of study that's ripe with open questions that can be interesting for the online learning community that is typically very keen on instance-dependent analysis.

A more ambitious question for future research is if our techniques can be extended to more challenging settings, and especially online learning in Markov decision processes (Even-Dar et al., 2009; Neu et al., 2010, 2014). We think that an extension to this setting would be particularly valuable, given the recent flurry of interest in offline reinforcement learning. In this context, we could potentially exploit the unique feature of our algorithm design that, unlike all other methods, it does not rely on explicit uncertainty quantification for calculating its pessimistic updates. This could mean a major advantage over traditional off-policy RL methods that rely on uncertainty quantification to build confidence sets over abstract objects (like the entire transition function of the Markov process), which is a notoriously hard problem, especially in the infinite-horizon setting. In contrast, as our results in Section 3.3 highlight, the pessimistic nature of our method is realized through an update rule that is slightly more conservative than the standard exponential-weights update rule. We believe that this insight can be very useful for developing new methods for offline RL, even more so since they appear to be directly compatible with the primal-dual off-policy learning methods of Nachum et al. (2019a,b); Uehara et al. (2020).

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 950180).

References

Noga Alon, Nicolò Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In *COLT*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 23–35. JMLR.org, 2015.

Noga Alon, Nicolò Cesa-Bianchi, Claudio Gentile, Shie Mannor, Yishay Mansour, and Ohad Shamir. Nonstochastic multi-armed bandits with graph-structured feedback. *SIAM J. Comput.*, 46(6):1785–1826, 2017.

András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Mach. Learn.*, 71(1): 89–129, 2008.

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

Gábor Bartók, Dean P. Foster, Dávid Pál, Alexander Rakhlin, and Csaba Szepesvári. Partial monitoring - classification, regret bounds, and algorithms. *Math. Oper. Res.*, 39(4):967–997, 2014.

Jacob Buckman, Carles Gelada, and Marc G. Bellemare. The importance of pessimism in fixed-dataset policy optimization. In *ICLR*. OpenReview.net, 2021.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

Nicolò Cesa-Bianchi, Yishay Mansour, and Gilles Stoltz. Improved second-order bounds for prediction with expert advice. *Mach. Learn.*, 66(2-3):321–352, 2007.

Kamalika Chaudhuri, Yoav Freund, and Daniel J. Hsu. A parameter-free hedging algorithm. In *NeurIPS*, pages 297–305. Curran Associates, Inc., 2009.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 1042–1051. PMLR, 2019.

Ashok Cutkosky and Francesco Orabona. Black-box reductions for parameter-free online learning in banach spaces. In *COLT*, volume 75 of *Proceedings of Machine Learning Research*, pages 1493–1529. PMLR, 2018.

Eyal Even-Dar, Sham M. Kakade, and Yishay Mansour. Online Markov decision processes. *Math. Oper. Res.*, 34(3):726–736, 2009.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, 1997.

Pierre Gaillard, Gilles Stoltz, and Tim van Erven. A second-order bound with excess losses. In *COLT*, volume 35 of *JMLR Workshop and Conference Proceedings*, pages 176–196. JMLR.org, 2014.

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

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

Tomáš Kocák, Gergely Neu, and Michal Valko. Online learning with noisy side observations. In *AISTATS*, volume 51 of *JMLR Workshop and Conference Proceedings*, pages 1186–1194. JMLR.org, 2016.

Wouter M. Koolen. The Pareto regret frontier. In *NeurIPS*, pages 863–871, 2013.

Wouter M. Koolen and Tim van Erven. Second-order quantile methods for experts and combinatorial games. In *COLT*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 1155–1175. JMLR.org, 2015.

Tor Lattimore. The Pareto regret frontier for bandits. In *NeurIPS*, pages 208–216, 2015.

Tor Lattimore and Csaba Szepesvári. Cleaning up the neighborhood: A full classification for adversarial partial monitoring. In *ALT*, volume 98 of *Proceedings of Machine Learning Research*, pages 529–556. PMLR, 2019.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *CoRR*, abs/2005.01643, 2020.

Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Inf. Comput.*, 108(2): 212–261, 1994.

Haipeng Luo and Robert E. Schapire. Achieving all with no parameters: Adanormalhedge. In *COLT*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 1286–1304. JMLR.org, 2015.

Shie Mannor and Ohad Shamir. From bandits to experts: On the value of side-observations. In *NeurIPS*, pages 684–692, 2011.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *J. Mach. Learn. Res.*, 9:815–857, 2008.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. DualDICE: Behavior-agnostic estimation of discounted stationary distribution corrections. pages 2315–2325, 2019a.

Ofir Nachum, Bo Dai, Ilya Kostrikov, Yinlam Chow, Lihong Li, and Dale Schuurmans. AlgaeDICE: Policy gradient from arbitrary experience. *CoRR*, abs/1912.02074, 2019b.

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

Gergely Neu and Julia Olkhovskaya. Efficient and robust algorithms for adversarial linear contextual bandits. In *COLT*, volume 125 of *Proceedings of Machine Learning Research*, pages 3049–3068. PMLR, 2020.

Gergely Neu, András György, and Csaba Szepesvári. The online loop-free stochastic shortest-path problem. In *COLT*, pages 231–243. Omnipress, 2010.

Gergely Neu, András György, Csaba Szepesvári, and András Antos. Online Markov decision processes under bandit feedback. *IEEE Trans. Autom. Control.*, 59(3):676–691, 2014.

Francesco Orabona and Dávid Pál. Coin betting and parameter-free online learning. In *NeurIPS*, pages 577–585, 2016.

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

Aldo Rustichini. Minimizing regret: The general case. *Games and Economic Behavior*, 29(1-2): 224–243, 1999.

Amir Sani, Gergely Neu, and Alessandro Lazaric. Exploiting easy data in online optimization. In *NeurIPS*, pages 810–818, 2014.

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*, 2021.

Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and q-function learning for off-policy evaluation. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 9659–9668. PMLR, 2020.

Vladimir G. Vovk. Aggregating strategies. In *COLT*, pages 371–386. Morgan Kaufmann, 1990.

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

Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 11404–11413. PMLR, 2021.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. In *NeurIPS*, pages 6683–6694, 2021.

Appendix A. Omitted details from the proof of Theorem 1

A.1. Proof of the bound of Equation (14)

This proof is extracted from the proof of Lemma 1 of Neu (2015). Let $c \in \mathbb{R}_+$ be any non-negative constant. Then,

$$\frac{r_t(a)\mathbb{1}\{A_t = a\}}{\pi_B(a) + c} \le \frac{r_t(a)\mathbb{1}\{A_t = a\}}{\pi_B(a) + c\, r_t(a)} = \frac{\mathbb{1}\{A_t = a\}}{2c} \cdot \frac{2c\, r_t(a)/\pi_B(a)}{1 + c\, r_t(a)/\pi_B(a)} \le \frac{1}{2c} \log(1 + 2c\widehat{r}_t(a))$$

where the first step follows from $r_t(a) \in [0,1]$ and the last one from the inequality $\frac{x}{1+x/2} \le \log(1+x)$, which holds for all $x \ge 0$.

A.2. The proof of Lemma 4

We study the evolution of the potential function $\frac{1}{\eta}\log\frac{W_{n+1}}{W_1}$. On the one hand, we have for any action $\bar{a}$ that

$$\frac{1}{\eta}\log\frac{W_{n+1}}{W_1} = \frac{1}{\eta}\log\left(\frac{1}{K}\sum_a w_{n+1}(a)\right) \ge \frac{1}{\eta}\log\left(\frac{1}{K}\, w_{n+1}(\bar{a})\right) = \sum_{t=1}^n \widetilde{r}_t(a) - \frac{\log K}{\eta}. \quad (17)$$

Multiplying this bound with $\pi^*(\bar{a})$ and summing up over actions gives the lower bound

$$\frac{1}{\eta}\log\frac{W_{n+1}}{W_1} \ge \sum_{t=1}^n \sum_a \pi^*(a)\widetilde{r}_t(a) - \frac{1}{\eta}\log K. \quad (18)$$

On the other hand, the potential can be rewritten as follows:

$$\frac{1}{\eta}\log\frac{W_{n+1}}{W_1} = \frac{1}{\eta}\sum_{t=1}^n \log\frac{W_{t+1}}{W_t} = \frac{1}{\eta}\sum_{t=1}^n \log\frac{\sum_a w_t(a)e^{\eta\widetilde{r}_t(a)}}{W_t} = \frac{1}{\eta}\sum_{t=1}^n \log\sum_a \pi_t(a)e^{\eta\widetilde{r}_t(a)}.$$

Putting the two expressions together concludes the proof.

Appendix B. The proof of Theorem 2

In this proof, we have to face the added challenge of having to account for the possible inaccuracy of our estimator of π_B. To this end, we define a sequence of "good events" under which the policy estimate is well-concentrated and analyze the regret under this event and its complement, using that the good event should hold with high probability. Concretely, we define the failure probability $\delta_t \in (0,1)$, the tolerance parameter ε_t, and the t-th good event as follows:

$$\varepsilon_1 = 1\,, \quad \varepsilon_t = \sqrt{\frac{\log(K/\delta_t)}{2(t-1)}} \qquad E_t = \big\{|\widehat{\pi}_t(a) - \pi_B(a)| \le \varepsilon_t \ (\forall a \in \mathcal{A})\big\}. \quad (19)$$

An application of Hoeffding's inequality shows that E_t holds with probability at least $1 - \delta_t$. Now, setting $\gamma_t = \varepsilon_t + \eta/2$, we can observe that under event E_t, we have

$$\widetilde{r}_t(a) = \frac{r_t(a)\mathbb{1}\{A_t^B = a\}}{\widehat{\pi}_t + \gamma_t} \le \frac{r_t(a)\mathbb{1}\{A_t^B = a\}}{\pi_B(a) + \eta/2} \le \frac{1}{\eta}\log(1 + \eta\widehat{r}_t(a)). \quad (20)$$

We proceed by noticing that the bound of Lemma 4 continues to apply, and that we can bound the term appearing on the right-hand side as follows:

$$\mathbb{E}_t\left[\frac{1}{\eta}\log\sum_a \pi_t(a)\exp\left(\eta\widetilde{r}_t(a)\right)\right] \le \mathbb{1}\{E_t\}\mathbb{E}_t\left[\frac{1}{\eta}\log\sum_a \pi_t(a)\exp\left(\eta\widetilde{r}_t(a)\right)\right] + \mathbb{1}\{\overline{E}_t\}\frac{2}{\eta}$$
$$\le \mathbb{1}\{E_t\}\mathbb{E}_t\left[\sum_a \pi_t(a)\widehat{r}_t(a)\right] + \mathbb{1}\{\overline{E}_t\}\frac{2}{\eta} \le \sum_a \pi_t(a)r_t(a) + \mathbb{1}\{\overline{E}_t\}\frac{2}{\eta},$$

where in the first line we used that $e^{\eta\widetilde{r}_t(a)} \le e^{\eta/\gamma_t} \le e^2$ and in the second line we used the bound of Equation (20), the fact that E_t is $\mathcal{F}_{t-1}$-measurable, that $\mathbb{E}_t\left[\widehat{r}_t(a)\right] = r_t(a)$, and finally upper bounded the indicator $\mathbb{1}\{E_t\}$ by one. Taking marginal expectations and summing up for all t, we get

$$\mathbb{E}\left[\frac{1}{\eta}\sum_{t=1}^n\log\sum_a \pi_t(a)\exp\left(\eta\widetilde{r}_t(a)\right)\right] \le \mathbb{E}\left[\sum_{t=1}^n\sum_a \pi_t(a)r_t(a)\right] + \frac{2}{\eta}\sum_{t=1}^n\delta_t,$$

where we used $\mathbb{E}\left[\mathbb{1}\{\overline{E}_t\}\right] \le \delta_t$.

It thus remains to relate the term on the left-hand side of the bound of Lemma 4. To do this, we similarly write

$$\mathbb{E}_t\left[\widetilde{r}_t(a)\right] \ge \mathbb{1}\{E_t\}\mathbb{E}_t\left[\widetilde{r}_t(a)\right] = \mathbb{1}\{E_t\}\mathbb{E}_t\left[\frac{r_t(a)\mathbb{1}\{A_t^B = a\}}{\widehat{\pi}_t(a)+\gamma_t}\right]$$
$$\ge \mathbb{1}\{E_t\}\cdot\frac{r_t(a)\pi_B(a)}{\pi_B(a)+\varepsilon_t+\gamma_t} \ge \mathbb{1}\{E_t\}r_t(a) - \frac{\varepsilon_t+\gamma_t}{\pi_B(a)},$$

where in the first inequality we exploited that $\widetilde{r}_t(a)$ is nonnegative, in the second one we used that E_t is $\mathcal{F}_{t-1}$-measurable and the defining property of the good event, and in the last one we simplified some expressions. Thus, we have

$$\mathbb{E}\left[\sum_{t=1}^n\sum_a \pi^*(a)r_t(a)\right] \le \mathbb{E}\left[\sum_{t=1}^n\sum_a \pi^*(a)\left(\widetilde{r}_t(a) + (1-\mathbb{1}\{E_t\})r_t(a) + \frac{\varepsilon_t+\gamma_t}{\pi_B(a)}\right)\right]$$
$$= \mathbb{E}\left[\sum_{t=1}^n\sum_a \pi^*(a)\widetilde{r}_t(a)\right] + \sum_{t=1}^n\delta_t + \sum_a\frac{\pi^*(a)}{\pi_B(a)}\sum_{t=1}^n\left(2\varepsilon_t+\frac{\eta}{2}\right),$$

where in the last line we recalled that $\gamma_t = \varepsilon_t + \eta/2$. Putting the two bounds together, we arrive to

$$\mathbb{E}\left[\sum_{t=1}^n\sum_a\left(\pi^*(a)r_t(a) - \pi_t(a)r_t(a)\right)\right] \le \frac{\log K}{\eta} + \left(1+\frac{2}{\eta}\right)\sum_{t=1}^n\delta_t + \left(\frac{\eta n}{2} + 2\sum_{t=1}^n\varepsilon_t\right)\sum_a\frac{\pi^*(a)}{\pi_b(a)}$$

Finally, we set $\delta_1 = 0\,,\ \delta_t = (t-1)^{-2}$ so that we have $\sum_{t=1}^n\delta_t = \pi^2/6 \le 2$ and we can write

$$\sum_{t=1}^n\varepsilon_t = 1 + \sum_{t=1}^{n-1}\sqrt{\frac{\log(Kt^2)}{2t}} \le 2\sqrt{n\log(Kn)},$$

where we also used the standard upper bound $\sum_{t=1}^{n} 1/\sqrt{t} \leq 2\sqrt{n}$. Putting everything together, we finally get

$$\mathbb{E}\left[\sum_{t=1}^{n}\sum_{a}\left(\pi^*(a)r_t(a) - \pi_t(a)r_t(a)\right)\right] \leq \frac{16+\log K}{\eta} + \left(\frac{\eta n}{2} + 2\sqrt{n\log(Kn)}\right) C(\pi^*;\pi_B) + 2.$$

Setting $\eta = \sqrt{\frac{\log K}{n}}$ concludes the proof.

Appendix C. The proof of Theorem 3

The proof combines ideas from the previous two proofs with ideas from Neu and Olkhovskaya (2020) to deal with the contextual aspect of the problem setting. In the following, let $\widehat{r}_t(x,a) = \langle \widehat{\theta}_t, \varphi(x,a)\rangle$. As a starting point, we fix a context $x \in \mathcal{X}$ and define the estimated regret in context x against comparator π^* as

$$\widehat{\mathcal{R}}(\pi^*, x) = \sum_{t=1}^{n}\sum_{a}\left(\pi^*(a|x) - \pi_t(a|x)\right)\widehat{r}_t(x,a). \tag{21}$$

The following lemma gives a bound on the above quantity:

Lemma 5 *Suppose that $\eta\widehat{r}_t(x,a) \geq -1/2$ holds for all x, a. Then, for any fixed x and π^*,*

$$\widehat{\mathcal{R}}(\pi^*, x) \leq \frac{\log K}{\eta} + \eta\sum_{t=1}^{n}\sum_{a}\pi^*(a|x)\left(\widehat{r}_t(x,a)\right)^2. \tag{22}$$

The proof follows from a careful combination of techniques by Cesa-Bianchi et al. (2007) and Neu and Olkhovskaya (2020), and is deferred to Appendix C.2. We proceed by noting that for any fixed x, the second term in the bound can be bounded as follows:

$$\begin{aligned}
\mathbb{E}_t\left[\left(\widehat{r}_t(x,a)\right)^2\right] &= \mathbb{E}_t\left[\left(R_t^B\right)^2\varphi(x,a)^\intercal\left(\overline{V}(\pi_B)\right)^{-1}\varphi(X_t,A_t^B)\varphi(X_t,A_t^B)^\top\left(\overline{V}(\pi_B)\right)^{-1}\varphi(x,a)\right] \\
&\leq \varphi(x,a)^\top\left(\overline{V}(\pi_B)\right)^{-1}\overline{V}(\pi_B)\left(\overline{V}(\pi_B)\right)^{-1}\varphi(x,a) = \varphi(x,a)^\intercal\left(\overline{V}(\pi_B)\right)^{-1}\varphi(x,a) \\
&= \mathrm{Tr}\left(\left(\overline{V}(\pi_B)\right)^{-1}\varphi(x,a)\varphi(x,a)^\intercal\right),
\end{aligned} \tag{23}$$

where we have used $R_t^B \leq 1$ in the inequality. Furthermore, in order to use the lemma, we first need to verify that its precondition is satisfied. To this end, notice that

$$|\widehat{r}_t(x,a)| = \left|R_t\phi(x,a)^\intercal\left(\overline{V}(\pi_B)\right)^{-1}\phi(X_t,A_t^b)\right| \leq \frac{\sup_{x,a}\|\phi(x,a)\|_2^2}{\lambda_{\min}(\overline{V}(\pi_B))},$$

which follows from a straightforward applicaiton of the Cauchy–Schwarz inequality. Thus, the condition on η we impose in the theorem guarantees that $\eta|\widehat{r}_t(x,a)| \leq 1/2$. Now we are in position to invoke Lemma 5, albeit with a specific choice for the context x. Specifically, we let X_0 be a

"ghost sample" drawn independently from the context distribution for the sake analysis, and apply Lemma 5 to obtain

$$\widehat{\mathcal{R}}(\pi^*, X_0) \le \frac{\log K}{\eta} + \eta \sum_{t=1}^{n} \sum_{a} \pi^*(a|X_0) \operatorname{Tr}\left(\left(\overline{V}(\pi_B)\right)^{-1} \varphi(X_0, a)\varphi(X_0, a)^\intercal \right). \tag{24}$$

Then, a straightforward calculation inspired by the analysis of Neu and Olkhovskaya (2020) shows that the left-hand side is related to the expected regret as

$$\mathbb{E}\left[\widehat{\mathcal{R}}(\pi^*, X_0)\right] = \mathbb{E}\left[\sum_{t=1}^{n} \sum_{a} \left(\pi^*(a|X_t) - \pi_t(a|X_t)\right) r_t(X_t, a)\right]. \tag{25}$$

For completeness, we include this calculation in Appendix C.1. The same technique can be used to deal with the term on the right-hand side as follows:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{n} \sum_{a} \pi^*(a|X_0) \operatorname{Tr}\left(\left(\overline{V}(\pi_B)\right)^{-1} \phi(X_0, a)\phi(X_0, a)^\intercal \right)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{n} \operatorname{Tr}\left(\left(\overline{V}(\pi_B)\right)^{-1} \sum_{a} \pi^*(a|X_0)\phi(X_0, a)\phi(X_0, a)^\intercal \right)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{n} \operatorname{Tr}\left(\left(\overline{V}(\pi_B)\right)^{-1} \overline{V}(\pi^*) \right)\right] = C_\varphi(\pi^*; \pi_B).
\end{aligned}$$

Thus, taking expectations of both sides of Equation (24) and using the above two results concludes the proof. □

C.1. The proof of the regret decomposition of Equation (25)

We start by fixing an arbitrary x and defining the following notion of pseudo-regret in context x:

$$\mathcal{R}(\pi^*, x) = \sum_{t=1}^{n} \sum_{a} \left(\pi^*(a|x) - \pi_t(a|x)\right) r_t(x, a).$$

We first note that $\mathbb{E}[\widehat{\mathcal{R}}(\pi^*, x)] = \mathbb{E}[\mathcal{R}(\pi^*, x)]$ holds thanks to the unbiasedness of $\widehat{r}_t$ and the independence of π_t and $\widehat{r}_t$. In particular, this follows from the following derivation:

$$\begin{aligned}
\mathbb{E}[\widehat{\mathcal{R}}(\pi^*, x)] &= \mathbb{E}\left[\sum_{t=1}^{n} \mathbb{E}_t\left[\sum_{a} \left(\pi^*(a|x) - \pi_t(a|x)\right) \widehat{r}_t(x, a)\right]\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{n} \sum_{a} \left(\pi^*(a|x) - \pi_t(a|x)\right) \mathbb{E}_t\left[\widehat{r}_t(x, a)\right]\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{n} \sum_{a} \left(\pi^*(a|x) - \pi_t(a|x)\right) r_t(x, a)\right] = \mathbb{E}[\mathcal{R}(\pi^*, x)],
\end{aligned}$$

where we used the tower rule of expectation in the first step, the fact that π_t is $\mathcal{F}_{t-1}$-measurable in the second step, and the unbiasedness of the reward estimator in the last step. To relate $\mathbb{E}[\mathcal{R}(\pi^*, x)]$ and the true expected Regret $\mathbb{E}[\mathcal{R}(\pi^*)]$, we consider the random variable $\mathcal{R}(\pi^*, X_0)$ with X_0 being a ghost sample drawn from the context distribution independently from the history of contexts $(X_t)_{t=1}^n$. Then, we can write the expectation of this random variable as

$$\begin{aligned}
\mathbb{E}[\mathcal{R}(\pi^*, X_0)] &= \mathbb{E}\left[\sum_{t=1}^{n}\sum_{a}\left(\pi^*(a|X_0) - \pi_t(a|X_0)\right) r_t(X_0, a)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{n}\mathbb{E}_t\left[\sum_{a}\left(\pi^*(a|X_0) - \pi_t(a|X_0)\right) r_t(X_0, a)\right]\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{n}\mathbb{E}_t\left[\sum_{a}\left(\pi^*(a|X_t) - \pi_t(a|X_t)\right) r_t(X_t, a)\right]\right] = \mathbb{E}[\mathcal{R}(\pi^*)],
\end{aligned}$$

where the second line uses the tower rule of expectation and the third one the fact that X_0 is distributed identically with X_t given $\mathcal{F}_{t-1}$. This concludes the proof. □

C.2. Proof of Lemma 5

The proof is inspired by the classic PROD analysis of Cesa-Bianchi et al. (2007), and follows from similar arguments as the proof of Lemma 4. The main adjustment we need to these proofs is that now we have to include contexts in our derivations. To this end, let us fix one context $x \in \mathcal{X}$ and suppose that the condition of the theorem is satisfied: $\eta\,\widehat{r}_t(x, a) \geq -1/2$ for all actions $a \in \mathcal{A}$.

As before, we will study the evolution of the potential function $\frac{1}{\eta}\log\frac{W_{n+1}(x)}{W_1(x)}$. For every action $\bar{a} \in \mathcal{A}$ we have:

$$\begin{aligned}
\frac{1}{\eta}\log W_{n+1}(x) &= \frac{1}{\eta}\log\sum_{a}\prod_{t=1}^{n}(1 + \eta\widehat{r}_t(x, a)) \geq \frac{1}{\eta}\log\prod_{t=1}^{n}(1 + \eta\widehat{r}_t(x, \bar{a})) \\
&= \frac{1}{\eta}\sum_{t=1}^{n}\log(1 + \eta\widehat{r}_t(x, \bar{a})) \geq \sum_{t=1}^{n}\left(\widehat{r}_t(x, \bar{a}) - \eta\left(\widehat{r}_t(x, \bar{a})\right)^2\right),
\end{aligned}$$

where we used our condition on the magnitude of the reward estimates twice: once to use $(1 - \eta\widehat{r}_t(x, a)) \geq 0$ in the first line and once when using the elementary inequality $\log(1 + z) \geq z - z^2$

that holds for all $z \geq -1/2$ in the second line. Moreover, we can upper-bound the potential as

$$\begin{aligned}
\frac{1}{\eta}\log W_{n+1}(x) &= \frac{1}{\eta}\log W_1 + \frac{1}{\eta}\log\prod_{t=1}^{n}\frac{W_{t+1}(x)}{W_t(x)} = \frac{\log K}{\eta} + \frac{1}{\eta}\sum_{t=1}^{n}\log\frac{W_{t+1}(x)}{W_t(x)} \\
&= \frac{\log K}{\eta} + \frac{1}{\eta}\sum_{t=1}^{n}\log\sum_a \frac{w_t(x,a)}{W_t(x)}(1+\eta\,\widehat{r}_t(x,a)) && \big(\text{def. of } W_{t+1}\big) \\
&= \frac{\log K}{\eta} + \frac{1}{\eta}\sum_{t=1}^{n}\log\sum_a \pi_t(a|x)(1+\eta\,\widehat{r}_t(x,a)) && \big(\text{def. of } \pi_t\big) \\
&= \frac{\log K}{\eta} + \frac{1}{\eta}\sum_{t=1}^{n}\log\Big(1+\eta\sum_a \pi_t(a|x)\widehat{r}_t(x,a)\Big) \\
&\leq \frac{\log K}{\eta} + \sum_{t=1}^{n}\sum_a \pi_t(a|x)\widehat{r}_t(x,a),
\end{aligned}$$

where we used the inequality $\log(1+z) \leq z$ that holds for all $z > -1$.

Combining the lower bound and upper bound, we obtain

$$\sum_{t=1}^{n}\Big(\widehat{r}_t(x,\bar{a}) - \sum_a \pi_t(a|x)\widehat{r}_t(x,a)\Big) \leq \frac{\log K}{\eta} + \eta\sum_{t=1}^{n}\left(\widehat{r}_t(x,\bar{a})\right)^2.$$

Multiplying both sides by $\pi^*(\bar{a}|x)$ and summing over all actions $\bar{a} \in \mathcal{A}$ yields the desired result. □