

The multi-armed bandit problem with covariates

VIANNEY PERCHET*

PHILIPPE RIGOLLET†

October 27, 2011

Abstract

We consider a multi-armed bandit problem in a setting where each arm produces a noisy reward realization which depends on an observable random *covariate*. As opposed to the traditional *static* multi-armed bandit problem, this setting allows for dynamically changing rewards that better describe applications where side information is available. We adopt a nonparametric model where the expected rewards are smooth functions of the covariate and where the hardness of the problem is captured by a *margin* parameter. To maximize the expected cumulative reward, we introduce a policy called Adaptively Binned Successive Elimination (ABSE) that adaptively decomposes the global problem into suitably “localized” static bandit problems. This policy constructs an adaptive partition using a variant of the Successive Elimination (SE) policy. Our results include sharper regret bounds for the SE policy in a static bandit problem and minimax optimal regret bounds for the ABSE policy in the dynamic problem.

Mathematics Subject Classifications: Primary 62G08, Secondary 62L12.

Key Words: Nonparametric bandit, contextual bandit, multi-armed bandit, adaptive partition, successive elimination, sequential allocation, regret bounds

1 Introduction

The seminal paper [16] introduced an important class of sequential optimization problems, otherwise known as multi-armed bandits. These models have since been used extensively in such fields as statistics, operations research, engineering, computer science and economics. The traditional multi-armed bandit problem can be described as follows. Consider $K \geq 2$ statistical populations (arms), where at each point in time it is possible to sample from (pull) only one of them and receive a random reward dictated by the properties of the sampled population. The objective is to devise a sampling policy that maximizes expected cumulative rewards over a finite time horizon. The difference between the performance of a given sampling policy and that of an oracle, that repeatedly samples from the population with the highest mean reward, is called the *regret*. Thus, one can re-phrase the objective as minimizing the regret.

*ENS Cachan (CMLA) and Université Paris 7 (LPMA). Email: vianney.perchet@normalesup.org

†Princeton University. Partially supported by the NSF (DMS-0906424, DMS-1053987). Email: rigollet@princeton.edu

When the populations being sampled are homogenous, i.e., when the sequential rewards are independent and identically distributed (iid) in each arm, the family of upper-confidence-bound (UCB) policies, introduced in [11], incur a regret of order $\log n$, where n is the length of the time horizon, and no other “good” policy can (asymptotically) achieve a smaller regret; see also [3]. The elegance of the theory and sharp results developed in [11] hinge to a large extent on the assumption of homogenous populations and hence identically distributed rewards. This, however, is clearly too restrictive for many applications of interest. Often, the decision maker observes further information and based on that, a more *customized* allocation can be made. In such settings, rewards may still be assumed to be independent, but no longer identically distributed in each arm. A particular way to encode this is to allow for an exogenous variable (a covariate) that affects the rewards generated by each arm at each point in time when this arm is pulled.

Such a formulation was first introduced in [20] under parametric assumptions and in a somewhat restricted setting; see [7, 8] and [19] for very different recent approaches to the study of such bandit problems, as well as references therein for further links to antecedent literature. The first work to venture outside the realm of parametric modeling assumptions appeared in [21]. In particular, the mean response in each arm, conditionally on the covariate value, was assumed to follow a general functional form, hence one can view their setting as a *nonparametric* bandit problem. They propose a variant of the ε -greedy policy, see, e.g., [3] and show that the average regret tends to zero as the time horizon n grows to infinity. However, it is unclear whether this policy satisfy a more refined notion of optimality, insofar as the magnitude of the regret is concerned, as is the case for UCB-type policies in traditional bandit problems. Such questions were partially addressed in [15] where near-optimal bounds on the regret are proved in the case of a two-armed bandit problem under only two assumptions on the underlying functional form that governs the arms’ responses. The first is a mild smoothness condition and the second is a so-called *margin condition* that involves a *margin parameter* which encodes the “separation” between the functions that describe the arms’ responses.

The purpose of the present paper is to extend the setup of [15] to the K -armed bandit problem with covariates when K may be large. This involves a customized definition of the margin assumption. Moreover, the bounds proved in [15] suffered to deficiencies. First, they hold only for a limited range of values of the margin parameter and second, the upper bounds and the lower bounds mismatch by a logarithmic factor. Improving upon these results requires radically new ideas. To that end, we introduce three policies:

1. Successive Elimination (SE) is dedicated to the static bandit case. It is the cornerstone of the others policies that deal with covariates. During a first phase, this policy explores the different arms, builds estimates and eliminates sequentially suboptimal arms; once there only remains one arm, it is pulled until the horizon is reached. A variant of SE was originally introduced in [6]. However, it was not tuned to minimize the regret as other measures of performance were investigated in this paper. We prove new regret bounds for this policy that improve upon the canonical papers [11] and [3].
2. Binned Successive Elimination (BSE) follows a simple principle to solve the problem with covariates. It consists in grouping similar covariates into bins and then look only at the average reward over each bin. These bins are viewed as indexing “local” bandit problems,

solved by the aforementioned SE policy. We prove optimal regret bounds, polynomial in the horizon but only for a restricted class of difficult problems. For the remaining class of easy problems, the BSE policy is suboptimal.

3. Adaptively Binned Successive Elimination (ABSE) overcomes a severe limitation of the simple but naive BSE. Indeed, if the problem is globally easy (this is characterized by the margin condition), the BSE policy employs a fixed and too fine discretization of the covariate space. Instead, the ABSE policy partitions the space of covariates in a fashion that adapts to the local difficulty of the problem: cells are smaller when different arms are hard to distinguish and bigger when one arm dominates the other. This adaptive partitioning allows us to prove optimal regrets bounds for the whole class of problems.

The optimal polynomial regret bounds that we prove are much larger than the logarithmic bounds proved in the static case. Nevertheless, it is important to keep in mind that they are valid for a much more flexible model that incorporates the covariates. In the particular case where $K = 2$ and the problem is *difficult*, these bounds improve upon the results of [15] by removing a logarithmic factor that is idiosyncratic to the *exploration vs. exploitation* dilemma encountered in bandit problems. Moreover, it follows immediately from the previous minimax lower bounds of [2] and [15], that these bounds are optimal in a minimax sense and thus cannot be further improved. It reveals an interesting and somewhat surprising phenomenon: the price to pay for the partial information in the bandit problem is washed away by the price to pay for nonparametric estimation. Indeed the bound on the regret that we obtain in the bandit setup for $K = 2$ is of the same order as the best attainable bound in the *full information* case, where at each round, the operator receives the reward from only one arm but observes the rewards of both arms. An important example of the full information case is sequential binary classification.

Finally, beyond these analytical results, one of the contributions of the present paper is in pointing to some possible synergies and potentially interesting connections between the traditional bandit literature and nonparametric statistics.

Our policies for the problem with covariates fall into the family of “plug-in” policies as opposed “minimum contrast” policies; a detailed account of the differences and similarities between these two setups in the full information case can be found in [2]. Minimum contrast type policies have already received some attention in the bandit literature with side information, aka *contextual bandits*, in the papers [12] and also [10]. In these studies, admissible policies are restricted to a more limited set than the general class of non-anticipating policies. A related problem online convex optimization with side information was studied in [9], where the authors use a discretization technique similar to the one employed in this paper. It is worth noting that the cumulative regret in these papers is defined in a weaker form compared to the traditional bandit literature, since the cumulative reward of a proposed policy is compared to that of the best policy in a certain restricted class of policies. Therefore, bounds on the regret depend, among other things, on the complexity of said class of policies. Plug-in type policies have received attention in the context of the continuum armed bandit problem, where as the name suggests there are uncountably many arms. Notable entries in that stream of work are [13] and [17], who impose a smoothness condition both on the space of arms and the space of covariates, obtaining optimal regret bounds up to logarithmic terms.

2 Improved regret bounds for the static problem

In this section, it will be convenient for notational purposes, to consider a multi-armed bandit problem with $K + 1$ arms.

We revisit the Successive Elimination (SE) policy introduced in [6] in the traditional setup of multi-armed bandit problems. As opposed to the more popular UCB policy (see, e.g., [11, 3]), it will allow us in the next section to construct an adaptive partition that is crucial to attain optimal rates on the regret for the dynamic case with covariates. In this section, we prove refined regret bounds for the SE policy that exhibit a better dependence on the expected rewards of the arms compared to the bounds for UCB that were derived in [3]. Such an improvement was recently attempted in [4] and also in [1] for modified UCB policies and we compare these results to ours below.

Let us recall the traditional setup for the static multi-armed bandit problem (see, e.g., [3]). Let $\mathcal{I} = \{1, \dots, K + 1\}$ be a given set of $K + 1 \geq 2$ arms. Successive pulls of arm $i \in \mathcal{I}$ yield reward $Y_1^{(i)}, Y_2^{(i)}, \dots$ that are iid random variables in $[0, 1]$ with expectation given by $\mathbb{E}[Y_t^{(i)}] = f^{(i)} \in [0, 1]$. Assume without loss of generality that $f^{(1)} \leq \dots \leq f^{(K+1)}$ so that $K + 1$ is one of the best arms. For simplicity, we further assume that the best arm is *unique* for simplicity since for the SE policy, having multiple optimal arms only improves the regret bound. In the analysis, it will be convenient to denote this optimal arm by $\star := K + 1$ and to define the *gaps* traditionally denoted by $\Delta_1 \geq \dots \geq \Delta_\star = 0$, by $\Delta_i = f^{(\star)} - f^{(i)} \geq 0$.

A *non-anticipating policy* $\pi = \{\pi_t\}$ is a sequence of random variables $\pi_t \in \{1, \dots, K + 1\}$ indicating which arm to pull at each time $t = 1, \dots, n$, and such that π_t depends only on observations strictly anterior to t .

The performance of a policy π is measured by its (*cumulative*) *regret* at time n defined by

$$R_n(\pi) := \sum_{t=1}^n \left(f^{(\star)} - f^{(\pi_t)} \right).$$

We begin with a high-level description of the SE policy denoted by $\hat{\pi}$. It operates in rounds that are different from the decision times $t = 1, \dots, n$. At the beginning of each round τ , a subset of the arms has been eliminated and only a subset \mathcal{I}_τ remains. During round τ , each arm in \mathcal{I}_τ is pulled exactly once (EXPLORATION). At the end of the round, for each remaining arm, we decide whether to eliminate an arm in \mathcal{I}_τ using a simple statistical hypothesis test: if we conclude that its mean is significantly smaller than the mean of any remaining arm, then we eliminate this arm and we keep it otherwise (ELIMINATION). We repeat this procedure until n pulls have been made. The number of rounds is random but obviously smaller than n .

The SE policy described in Policy 1 outputs an infinite sequence of arms $\hat{\pi}_1, \hat{\pi}_2, \dots$ without a prescribed horizon. Of course, it can be truncated at any horizon n . This description emphasizes the fact that the policy can be implemented without knowledge of the horizon n .

Note that after the exploration phase of each round $\tau = 1, 2, \dots$, each remaining arm $i \in \mathcal{I}_\tau$ has been pulled exactly τ times, generating rewards $Y_1^{(i)}, \dots, Y_\tau^{(i)}$. Denote by $\bar{Y}^{(i)}(\tau)$ the average reward collected from arm $i \in \mathcal{I}_\tau$ before round τ that is defined by $\bar{Y}^{(i)}(\tau) = (1/\tau) \sum_{t=1}^{\tau} Y_t^{(i)}$, where and throughout this paper, we use the convention $1/0 = \infty$. For any positive integer T ,

define also

$$U(\tau, T) = 2\sqrt{\frac{2\log\left(\frac{T}{\tau}\right)}{\tau}}, \quad (2.1)$$

which is essentially a high probability upper bound on the magnitude of deviations of $\bar{Y}^{(j)}(\tau) - \bar{Y}^{(i)}(\tau)$ from its mean $f^{(j)} - f^{(i)}$.

The SE policy for a K -armed bandit problem can be implemented according to the pseudocode of Policy 1. Note that, to ease the presentation of Sections 4 and 5, the SE policy also returns at each time t , the number of rounds $\hat{\tau}_t$ completed at time t and a subset $\hat{S}_t \in \mathcal{P}(\mathcal{I})$ of arms that are active at time t , where $\mathcal{P}(\mathcal{I})$ denotes the power set of \mathcal{I} .

Policy 1 Successive Elimination (SE)

Input: Set of arms $\mathcal{I} = \{1, \dots, K\}$; Parameters T, γ ; Horizon n .

Output: $(\hat{\pi}_1, \hat{\tau}_1, \hat{S}_1), (\hat{\pi}_2, \hat{\tau}_2, \hat{S}_2), \dots \in \mathcal{I} \times \mathbb{N} \times \mathcal{P}(\mathcal{I})$.

$\tau \leftarrow 1, S \leftarrow \mathcal{I}, t \leftarrow 0, \bar{Y} \leftarrow (0, \dots, 0) \in [0, 1]^K$

loop

$\bar{Y}^{\max} \leftarrow \max\{\bar{Y}^{(i)} : i \in S\}$

for $i \in S$ **do**

if $\bar{Y}^{(i)} \geq \bar{Y}^{\max} - \gamma U(\tau, T)$ **then**

$t \leftarrow t + 1$

$\hat{\pi}_t \leftarrow i$ (observe $Y^{(i)}$). EXPLORATION

$\hat{S}_t \leftarrow S, \hat{\tau}_t \leftarrow \tau$

$\bar{Y}^{(i)} \leftarrow \frac{1}{\tau}[(\tau - 1)\bar{Y}^{(i)} + Y^{(i)}]$.

else

$S \leftarrow S \setminus \{i\}$. ELIMINATION

end if

end for

$\tau \leftarrow \tau + 1$.

end loop

The following theorem gives a first upper bound on the regret of the SE policy. In the rest of the paper, \log denotes the natural logarithm and $\overline{\log}(x) = \log(x) \vee 1$.

Theorem 2.1 Consider a $(K+1)$ -armed bandit problem and let K_0 be an integer between 1 and K . When implemented with parameters $T = n, \gamma = 1$, the SE policy $\hat{\pi}$ exhibits a regret bounded as

$$\mathbb{E}R_n(\hat{\pi}) \leq 324 \sum_{i=1}^{K_0} \frac{1}{\Delta_i} \overline{\log}\left(\frac{n\Delta_i^2}{18}\right) + 288 \frac{K - K_0}{\Delta_{K_0}} \overline{\log}\left(\frac{n\Delta_{K_0}^2}{18}\right) + n\Delta_{K_0+1}.$$

In particular, choosing $K_0 = K$ and $K_0 = \max\{i \leq K; \Delta_i \geq \sqrt{K \log(K)/n}\}$ respectively give

$$\mathbb{E}R_n(\hat{\pi}) \leq 324 \sum_{i=1}^K \frac{1}{\Delta_i} \overline{\log}\left(\frac{n\Delta_i^2}{18}\right) \quad \text{and} \quad \mathbb{E}R_n(\hat{\pi}) \leq 324\sqrt{nK \log(K)}.$$

PROOF. Define $\varepsilon_\tau = U(\tau, n)$. Moreover, for any i in the set \mathcal{I}_τ of arms that remain at the beginning of round τ , define $\widehat{\Delta}_i(\tau) := \bar{Y}^{(\star)}(\tau) - \bar{Y}^{(i)}(\tau)$. Recall that, at round τ , if arms $i, \star \in \mathcal{I}_\tau$, then (i) the optimal arm \star eliminates arm i if $\widehat{\Delta}_i(\tau) \geq \varepsilon_\tau$ and (ii) arm i eliminates arm \star if $\widehat{\Delta}_i(\tau) \leq -\varepsilon_\tau$.

Since $\widehat{\Delta}_i(\tau)$ estimates Δ_i , the event in (i) happens approximately, when $\varepsilon_\tau \simeq \Delta_i$, so we introduce the deterministic, but unknown, quantity τ_i^* (and its approximation $\tau_i = \lceil \tau_i^* \rceil$) defined as the solution of:

$$\Delta_i = \frac{3}{2} \varepsilon_{\tau_i^*} = 3 \sqrt{2 \frac{\log\left(\frac{n}{\tau_i^*}\right)}{\tau_i^*}}, \text{ so that } \tau_i \leq 18 \frac{\overline{\log}(n\Delta_i^2/18)}{\Delta_i^2}.$$

Moreover, it holds that $1 \leq \tau_1 \leq \dots \leq \tau_K$.

We are going to decompose the regret accumulated by a suboptimal arm i into three quantities:

- the regret accumulated by pulling this arm at most until round τ_i : this regret is smaller than $\tau_i \Delta_i$;
- the regret accumulated by eliminating the optimal arm \star between round $\tau_{i-1} + 1$ and τ_i ,
- the regret induced if arm i is still present at round τ_i (and in particular, if it has not been eliminated by the optimal arm \star).

We will prove that the second and third events happen with small probability, because of the choice of τ_i . Formally, define the following *good* events:

$$\begin{aligned} \mathcal{A}_i &= \{\text{The arm } \star \text{ has not been eliminated before round } \tau_i\}, \\ \mathcal{B}_i &= \{\text{Every arm } j \in \{1, \dots, i\} \text{ has been eliminated before round } \tau_j\}. \end{aligned}$$

Moreover, define $\mathcal{C}_i = \mathcal{A}_i \cap \mathcal{B}_i$ and observe that $\mathcal{C}_1 \supseteq \mathcal{C}_2 \supseteq \dots$. For any $i = 1, \dots, K$, the contribution to the regret incurred after time τ_i on \mathcal{C}_i is at most $n\Delta_{i+1}$ since each pull of arm $j \geq i + 1$ contributes to the regret by $\Delta_j \leq \Delta_{i+1}$. We decompose the underlying probability space denoted by \mathcal{C}_0 into the disjoint union $(\mathcal{C}_0 \setminus \mathcal{C}_1) \cup \dots \cup (\mathcal{C}_{K_0-1} \setminus \mathcal{C}_{K_0}) \cup \mathcal{C}_{K_0}$.

It implies the following decomposition of the regret:

$$\mathbb{E}R_n(\widehat{\pi}) \leq \sum_{i=1}^{K_0} n\Delta_i \mathbb{P}(\mathcal{C}_{i-1} \setminus \mathcal{C}_i) + \sum_{i=1}^{K_0} \tau_i \Delta_i + n\Delta_{K_0+1}. \quad (2.2)$$

Note that the first term on the right-hand side of the above inequality can be bounded as follows

$$\sum_{i=1}^{K_0} n\Delta_i \mathbb{P}(\mathcal{C}_{i-1} \setminus \mathcal{C}_i) \leq n \sum_{i=1}^{K_0} \Delta_i \mathbb{P}(\mathcal{A}_i^c | \mathcal{C}_{i-1}) + n \sum_{i=1}^{K_0} \Delta_i \mathbb{P}(\mathcal{B}_i^c | \mathcal{A}_i \cap \mathcal{B}_{i-1}), \quad (2.3)$$

where the right-hand side was obtained using the decomposition $\mathcal{C}_i^c = \mathcal{A}_i^c \cup (\mathcal{B}_i^c \cap \mathcal{A}_i)$ and the fact that $\mathcal{A}_i \subseteq \mathcal{A}_{i-1}$.

On the event $\mathcal{A}_i \cap \mathcal{B}_{i-1}$, every suboptimal arm $j \leq i-1$ has been eliminated before round τ_{i-1} and the optimal arm \star is present at round τ_i ; so the probability $\mathbb{P}(\mathcal{B}_i^c | \mathcal{A}_i \cap \mathcal{B}_{i-1})$ is smaller than $\mathbb{P}(\widehat{\Delta}_i(\tau_i) < \varepsilon_{\tau_i})$. From Hoeffding's inequality, we have that for every $\varepsilon \in (0, \Delta)$ and every $\tau \geq 1$:

$$\mathbb{P}\left(\widehat{\Delta}_i(\tau) < \varepsilon\right) = \mathbb{P}\left(\widehat{\Delta}_i(\tau) - \Delta_i < \varepsilon - \Delta_i\right) \leq \exp\left(-\frac{\tau(\Delta_i - \varepsilon)^2}{2}\right). \quad (2.4)$$

The choice of τ_i implies that $\Delta_i \geq \frac{3}{2}\varepsilon_{\tau_i}$, so that

$$\mathbb{P}(\mathcal{B}_i^c | \mathcal{A}_i \cap \mathcal{B}_{i-1}) \leq \mathbb{P}(\widehat{\Delta}_i(t) < \varepsilon_{\tau_i}) \leq \exp\left(-\frac{\tau_i \varepsilon_i^2}{8}\right) \leq \frac{\tau_i}{n}. \quad (2.5)$$

It remains to bound the first term in the rhs of (2.3). On the event \mathcal{C}_{i-1} , the optimal arm \star has not been eliminated before the round τ_{i-1} but every suboptimal arm $j \leq i-1$ has. So the probability that there exists an arm $j \geq i$ that eliminates \star between τ_{i-1} and τ_i can be bounded as

$$\begin{aligned} \mathbb{P}(\mathcal{A}_i^c | \mathcal{C}_{i-1}) &\leq \mathbb{P}(\exists(j, s), i \leq j \leq K, \tau_{i-1} + 1 \leq s \leq \tau_i; \widehat{\Delta}_j(s) \leq -\varepsilon_s) \\ &\leq \sum_{j=i}^K \mathbb{P}(\exists s, \tau_{i-1} + 1 \leq s \leq \tau_i; \widehat{\Delta}_j(s) \leq -\varepsilon_s) \\ &= \sum_{j=i}^K [\Phi_j(\tau_i) - \Phi_j(\tau_{i-1})] \end{aligned}$$

where, using Lemma A.1, we get $\Phi_j(\tau) := \mathbb{P}(\exists s \leq \tau; \widehat{\Delta}_j(s) \leq -\varepsilon_s) \leq 4\frac{\tau}{n}$. Moreover, the above inequality implies that

$$\begin{aligned} \sum_{i=1}^{K_0} \Delta_i \mathbb{P}(\mathcal{A}_i^c | \mathcal{C}_{i-1}) &\leq \sum_{i=1}^{K_0} \Delta_i \sum_{j=i}^K [\Phi_j(\tau_i) - \Phi_j(\tau_{i-1})] \\ &\leq \sum_{j=1}^K \sum_{i=1}^{j \wedge K_0 - 1} \Phi_j(\tau_i) (\Delta_i - \Delta_{i+1}) + \sum_{j=1}^K \Phi_{j \wedge K_0}(\tau_{j \wedge K_0}) \Delta_{j \wedge K_0} \\ &\leq \frac{4}{n} \sum_{j=1}^K \sum_{i=1}^{j \wedge K_0 - 1} \tau_i (\Delta_i - \Delta_{i+1}) + \frac{4}{n} \sum_{j=1}^K \tau_{j \wedge K_0} \Delta_{j \wedge K_0}. \end{aligned}$$

Using the facts that $\tau_i \leq 18 \frac{\overline{\log}(n\Delta_i^2/18)}{\Delta_i^2}$ and $\Delta_{i+1} \leq \Delta_i$, the first sum can be bounded as :

$$\begin{aligned} \sum_{j=1}^K \sum_{i=1}^{j \wedge K_0 - 1} \tau_i (\Delta_i - \Delta_{i+1}) &\leq 18 \sum_{j=1}^K \sum_{i=1}^{j \wedge K_0 - 1} \overline{\log}\left(\frac{n\Delta_i^2}{18}\right) \frac{\Delta_i - \Delta_{i+1}}{\Delta_i^2} \\ &\leq 18 \sum_{j=1}^K \int_{\Delta_{j \wedge K_0}}^{\Delta_1} \overline{\log}\left(\frac{nx^2}{18}\right) \frac{dx}{x^2} \\ &\leq 18 \sum_{j=1}^K \frac{1}{\Delta_{j \wedge K_0}} \left[\overline{\log}\left(\frac{n\Delta_{j \wedge K_0}^2}{18}\right) + 2 \right]. \end{aligned}$$

The previous two displays yield

$$\sum_{i=1}^{K_0} \Delta_i \mathbb{P}(\mathcal{A}_i^c | \mathcal{C}_{i-1}) \leq \frac{288}{n} \sum_{j=1}^K \frac{1}{\Delta_j \wedge K_0} \overline{\log} \left(\frac{n \Delta_j^2}{18} \right).$$

Putting together (2.2), (2.3), (2.5) and the above display completes the proof of the theorem. ■

None of the bounds in Theorem 2.1 or the one provided in [1] is stronger than the other. The superiority of one over the other depends on the set $\{\Delta_1, \dots, \Delta_K\}$: in some cases (for example if every suboptimal arms have the same expectation) the latter is the best while in other cases (if the Δ_i are spread) our bounds are better.

Theorem 2.1 is actually closer to the result of [4]. The additional second term in our bound comes from the fact that we had to take into account the probability that an optimal arm \star can be eliminated by any arm, not just by some *suboptimal arm* with index lower than K_0 (see [4], page 8). It is unclear why it is enough to look at the elimination by those arms, since if \star is eliminated – no matter the arm that eliminated it –, the Hoeffding bound (2.4) will no longer hold.

The next easy corollaries are induced by slight variations of the policy or the setting; proofs are almost straightforward, so we just give quick insights.

Corollary 2.1 *When implemented with any parameter T and $\gamma = 1$ and run at horizon n , the SE policy $\hat{\pi}$ exhibits a regret bounded as*

$$\mathbb{E}R_n(\hat{\pi}) \leq 324 \max \left\{ 1, \frac{n}{T} \right\} \sum_{j=1}^K \frac{1}{\Delta_j} \overline{\log} \left(\frac{T \Delta_j^2}{18} \right).$$

Thus if the horizon is a random variable N of expectation n that is independent of the random rewards, the SE policy $\hat{\pi}$ implemented with T exhibits a regret bounded as

$$\mathbb{E}[R_N(\hat{\pi})] \leq 324 \left(1 + \frac{n}{T} \right) \sum_{j=1}^K \frac{1}{\Delta_j} \overline{\log} \left(\frac{T \Delta_j^2}{18} \right).$$

PROOF. In this setting, the regret induced by eliminating \star (or not eliminating a suboptimal arm i before round τ_i) is still bounded above by $n \Delta_i$. However, one has to substitute T to n in the probability of making such mistakes. ■

The following corollary is used in Sections 4 and 5.

Corollary 2.2 *If the horizon is a random variable N of expectation n that is independent of the random rewards, the SE policy $\hat{\pi}$ implemented with parameters $T, \gamma \geq 1$ exhibits a regret bounded as*

$$\mathbb{E}[R_N(\hat{\pi})] \leq 612 \gamma^2 \left(1 + \frac{n}{T} \right) \frac{K}{\Delta_{K_0}} \overline{\log} \left(\frac{T \Delta_{K_0}^2}{18} \right) + n \gamma^2 \Delta_{K_0}^-,$$

for any $K_0 \in \{1, \dots, K\}$ and where $\Delta_{K_0}^-$ is the largest Δ_j such that $\Delta_j < \Delta_{K_0}$.

PROOF. If $U'(\tau, T) = \gamma U(\tau, T) = 2\gamma\sqrt{\frac{2\log(\frac{T}{\tau})}{\tau}}$, then $\Delta_i = \frac{3}{2}U'(\tau'_i, T)$ implies that $\tau'_i \leq \gamma^2\tau_i$. With a larger term $U(\tau, T)$, the probability of eliminating \star is smaller and the probability of not eliminating arm i before τ'_i is (at most) multiplied by γ^2 . Therefore, the exact same proof as for Theorem 2.1 gives the same regret bound, only multiplied by γ^2 . ■

3 Bandit with covariates

This section is dedicated to a detailed description of nonparametric bandit with covariates.

3.1 Machine and game

A K -armed bandit machine with covariates (with K an integer greater than 2) is characterized by a sequence

$$(X_t, Y_t^{(1)}, \dots, Y_t^{(K)}), \quad t = 1, 2, \dots$$

of independent random vectors, where $(X_t)_{t \geq 1}$, is a sequence of iid covariates in $\mathcal{X} = [0, 1]^d$ with probability distribution P_X , and $Y_t^{(i)}$ denotes the random reward yielded by arm i at time t . We denote by E_X the expectation with respect to P_X . We assume that, for each $i = 1, \dots, K$, rewards $Y_t^{(i)}$, $t = 1, \dots, n$ are random variables in $[0, 1]$ with conditional expectation given by

$$\mathbb{E}[Y_t^{(i)} | X_t] = f^{(i)}(X_t), \quad i = 1, \dots, K, \quad t = 1, 2, \dots$$

where $f^{(i)}$, $i = 1, \dots, K$, are unknown functions such that $0 \leq f^{(i)}(x) \leq 1$, for any $i = 1, \dots, K$, $x \in \mathcal{X}$. A natural example is where $Y_t^{(i)}$ takes values in $\{0, 1\}$ so that the conditional distribution of $Y_t^{(i)}$ given X_t is Bernoulli with parameter $f^{(i)}(X_t)$.

The *game* takes place sequentially on this machine, pulling one of the arms at each time $t = 1, \dots, n$. A *non-anticipating policy* $\pi = \{\pi_t\}$ is a sequence of random functions $\pi_t : \mathcal{X} \rightarrow \{1, \dots, K\}$ indicating to the operator which arm to pull at each time t , and such that π_t depends only on observations strictly anterior to t . The *oracle policy* π^* , refers to the strategy that would be run by an omniscient operator with complete knowledge of the functions $f^{(i)}$, $i = 1, \dots, K$. Given side information X_t , the oracle policy π^* prescribes to pull any arm with the largest expected reward, i.e.,

$$\pi^*(X_t) \in \operatorname{argmax}_{i=1, \dots, K} f^{(i)}(X_t),$$

with ties broken arbitrarily. Note that the function $f^{(\pi^*(x))}(x)$ is equal to the pointwise maximum of the functions $f^{(i)}$, $i = 1, \dots, K$ and is defined by

$$f^*(x) = \max \left\{ f^{(i)}(x); \quad i = 1, \dots, K \right\}.$$

The oracle rule will be used to benchmark any proposed policy π and to measure the performance of the latter via its (*cumulative*) *regret* at time n defined by

$$R_n(\pi) := \mathbb{E} \sum_{t=1}^n (Y_t^{(\pi^*(X_t))} - Y_t^{(\pi_t(X_t))}) = \sum_{t=1}^n E_X (f^*(X) - f^{(\pi_t(X))}(X)).$$

Without further assumptions on the machine, the game can be arbitrarily difficult and, as a result, the regret can be arbitrarily close to n . In the following subsection, we describe natural regularity conditions under which it is possible to achieve sublinear growth rate of the regret, and characterize policies that perform in a near-optimal manner.

3.2 Smoothness and margin conditions

As usual in nonparametric estimation we first impose some regularity on the functions $f^{(i)}, i = 1, \dots, K$. Here and in what follows we use $\|\cdot\|$ to denote the Euclidean norm on \mathbb{R}^d .

SMOOTHNESS CONDITION. We say that the machine satisfies the smoothness condition with parameters (β, L) if

$$|f^{(i)}(x) - f^{(i)}(x')| \leq L\|x - x'\|^\beta, \quad \forall x, x' \in \mathcal{X}, i = 1, \dots, K$$

for some $\beta \in (0, 1]$ and $L > 0$.

Now denote the second pointwise maximum of the functions $f^{(i)}, i = 1, \dots, K$ by f^\sharp ; formally for every $x \in \mathcal{X}$ such that $\min f^{(i)}(x) \neq \max f^{(i)}(x)$ it is defined by:

$$f^\sharp(x) = \max \left\{ f^{(i)}(x); f^{(i)}(x) < f^*(x) \right\}$$

and by $f^\sharp(x) = f^*(x) = f^{(1)}(x)$ otherwise. The behavior of the function $\Delta := f^* - f^\sharp$ critically controls the complexity of the problem and the smoothness condition gives a local upper bound on this quantity. The second condition gives a lower bound on this function though in a weaker global sense. It is closely related to the margin condition employed in classification [18, 14], which drives the terminology employed here. It was originally imported to the bandit setup by [7].

MARGIN CONDITION. We say that the machine satisfies the margin condition with parameter $\alpha > 0$ if there exists $\delta_0 \in (0, 1), C_\delta > 0$ such that

$$P_X [0 < f^*(X) - f^\sharp(X) \leq \delta] \leq C_\delta \delta^\alpha, \quad \forall \delta \in [0, \delta_0]$$

If the marginal P_X has a density, the margin condition will only contain information about the behavior of the function Δ and not the marginal P_X itself. This is in contrast with [7] where the margin assumption is used precisely to control the behavior of the marginal P_X while that of the reward functions is fixed. A large value of the parameter α means that the function Δ either takes value 0 or is bounded away from 0, except over a set of small P_X -probability. Conversely, for values of α close to 0, the margin condition is essentially void and the two functions can be arbitrary close, making it difficult to distinguish them. This will reflect in the bounds on the regret derived in the subsequent section.

Intuitively, the smoothness condition and the margin condition work in opposite directions. Indeed, the former ensures that the function Δ does not “depart from zero” too fast whereas the latter warrants the opposite. The following proposition quantifies the extent of this conflict.

Proposition 3.1 *Under the smoothness condition with parameters (β, L) , and the margin condition with parameter α , the following holds*

- If $\alpha\beta > d$ then a given arm is either always or never optimal; in the latter case, it is bounded away from f^* and one can take $\alpha = \infty$;
- If $\alpha\beta > 1$ then there exists an oracle policy π^* that dictates to pull only one of the arms all the time, P_X -almost surely;
- If $\alpha\beta \leq 1$ then there exist machines with nontrivial oracle policies.

PROOF. The first two parts of the proof are straightforward consequences of Proposition 3.4 in [2]. To prove the last part, consider the example with $d = 1$, $\mathcal{X} = [0, 1]$, $f^{(2)} = \dots = f^{(K)} \equiv 0$ and $f^{(1)}(x) = L\text{sign}(x - .5)|x - .5|^{1/\alpha}$. Notice that $f^{(1)}$ satisfies the smoothness condition with parameters (β, L) if and only if $\alpha\beta \leq 1$. Any oracle policy is non-trivial, and, for example, one can define $\pi^*(x) = 2$ if $x \leq .5$ and $\pi^*(x) = 1$ if $x > .5$. Moreover, it can be easily shown that the machine satisfies the margin condition with parameter α and with $\delta_0 = C_\delta = 1$. ■

We denote by $\mathcal{M}_X^K(\alpha, \beta, L)$ the class of K -armed bandit problems with covariates in $\mathcal{X} = [0, 1]^d$ with a machine satisfying the margin condition with parameter $\alpha > 0$, the smoothness condition with parameters (β, L) and such that P_X has a density, with respect to the Lebesgue measure, bounded above and below by some $\bar{c} > 0$ and $\underline{c} > 0$.

3.3 Binning of the covariate space

To design a policy that solves the bandit problem with covariates described above, one has to inevitably find an estimate of the functions $f^{(i)}, i = 1, \dots, K$ at the current point X_t . There exists a wide variety of nonparametric regression estimators ranging from local polynomials to wavelet estimators. Both of the policies introduced below are based on estimators of $f^{(i)}, i = 1, \dots, K$ that are P_X almost surely piecewise constant over a particular collection of subsets, called *bins* of the covariate space \mathcal{X} .

We define a partition of \mathcal{X} in a measure theoretic sense as a collection of measurable sets, hereafter called *bins*, B_1, B_2, \dots such that $P_X(B_j) > 0$, $\bigcup_{j \geq 1} B_j = \mathcal{X}$ and $B_i \cap B_k = \emptyset, i, j \geq 1$, up to sets of null P_X probability. For any reward function f on \mathcal{X} , the average reward on bin B is defined by

$$\bar{f}_B = \mathbb{E}[f(X_t)|X_t \in B] = \frac{1}{p_B} \int_B f(x) dP_X(x), \quad (3.6)$$

where $p_B = P_X(B)$.

To define and analyze our policies, it will be convenient to reindex our observations $(X_t, Y_t^{(1)}, \dots, Y_t^{(K)})$ as follows. Given a partition \mathcal{B} of $[0, 1]^d$ and $x \in [0, 1]^d$, define $\mathcal{B}(x)$ to be the bin $B \in \mathcal{B}$ such that $x \in B$. Moreover, let T_B denote the times at which $X_t \in B, t = 1, \dots, n$. Conditionally on T_B , successive pulls of arm $i \in \{1, \dots, K\}$ at times in T_B yield rewards $Y_{B,1}^{(i)}, Y_{B,2}^{(i)}, \dots$ that are iid random variables in $[0, 1]$ with expectation given by $\mathbb{E}[Y_{B,t}^{(i)}] = \bar{f}_B^{(i)} \in [0, 1]$. Therefore, if we restrict our attention to observations in a given bin B , we are in the same setup as the static bandit problem studied in the previous section. The idea behind our first policy is to run the SE policy on each bin independently.

4 Binned Successive Elimination

We first outline a naive policy to operate the bandit machine described in section 3. It consists in fixing a partition of \mathcal{X} and applying the SE policy on each of the bins independently of each others.

The *Binned Successive Elimination* (BSE) policy $\bar{\pi}$ relies on a specific partition of \mathcal{X} . Let $\mathcal{B}_M := \{B_1, \dots, B_{M^d}\}$ be the regular partition of $\mathcal{X} = [0, 1]^d$, i.e., the reindexed collection of hypercubes defined for $\mathbf{k} = (k_1, \dots, k_d) \in \{1, \dots, M\}^d$ by

$$B_{\mathbf{k}} = \left\{ x \in \mathcal{X} : \frac{k_\ell - 1}{M} \leq x_\ell \leq \frac{k_\ell}{M}, \ell = 1, \dots, d \right\}.$$

In this paper, all sets are defined up to sets of null Lebesgue measure. As mentioned in subsection 3.3, the problem can be decomposed into M^d independent static bandit problems, one for each $B \in \mathcal{B}_M$. For the bandit problem corresponding to B , recall that arm i has expected reward $\bar{f}_B^{(i)}$ defined in (3.6) for any $i = 1, \dots, K$. The horizon $N_n(B)$ of the problem on B is random and equal to the number of covariates $X_t, t = 1, \dots, n$ that fall in B . Therefore, a policy π_j corresponding to the static problem on B is a sequence $\pi_{B,1}, \dots, \pi_{B,N_n(B)}$ of indices in $\{1, \dots, K\}$. Even though the horizon $N_n(B)$ is random, there exists positive constants \underline{c} and \bar{c} such that $\underline{c}nM^{-d} \leq \mathbb{E}[N_n(B)] \leq \bar{c}nM^{-d}$ for all $B \in \mathcal{B}_M$ because P_X has a density. As a result the SE policy on bin B is implemented with parameter $T = nM^{-d}$.

The BSE policy simply consists in running the SE policy on each of the B_j s, independently of each other. We denote by $\hat{\pi}_{B,1}, \hat{\pi}_{B,2}, \dots \in \mathcal{I}$, the sequence of actions indicated by the SE policy on bin B that is implemented with parameters $T = nM^{-d}, \gamma = 1$ for all $j = 1, \dots, M^d$. Moreover, $\hat{S}_{B,1}, \hat{S}_{B,2}, \dots \in \mathcal{P}(\mathcal{I})$ denotes the sequence of active arms for each policy $\hat{\pi}_B$. The BSE policy can be implemented according to the pseudo-code of Policy 2.

Policy 2 Binned Successive Elimination (BSE)

Input: Set of arms $\mathcal{I} = \{1, \dots, K\}$. Parameters n, M .

Output: $\bar{\pi}_1, \dots, \bar{\pi}_n \in \mathcal{I}$.

$\mathcal{B} \leftarrow \mathcal{B}_M$

for $B \in \mathcal{B}_M$ **do**

 Initialize a SE policy $\hat{\pi}_B$ with parameters $T = nM^{-d}, \gamma = 1$.

$t_B \leftarrow 0$.

end for

for $t = 1, \dots, n$ **do**

$B \leftarrow \mathcal{B}(X_t)$.

$t_B \leftarrow t_B + 1$.

$\bar{\pi}_t \leftarrow \hat{\pi}_{B,t_B}$ (observe $Y_t^{(\bar{\pi}_t)}$).

end for

The following theorem gives an upper bound on the regret of the BSE policy in the case where the problem is difficult, that is, when the margin parameter α satisfies $0 < \alpha < 1$.

Theorem 4.1 Fix $\beta \in (0, 1]$, $L > 0$ and $\alpha \in (0, 1)$ and consider a problem in $\mathcal{M}_X^K(\alpha, \beta, L)$. Then the BSE policy $\bar{\pi}$ with $M = \lfloor \left(\frac{n}{K \log(K)}\right)^{1/(2\beta+d)} \rfloor$ has a regret at time n bounded as follows,

$$\mathbb{E}R_n(\bar{\pi}) \leq Cn \left(\frac{K \log K}{n} \right)^{\frac{\beta(\alpha+1)}{2\beta+d}},$$

where $C > 0$ is a positive constant that does not depend on K .

The case $K = 2$ was studied in [15] using a similar policy called UCBogram. The bound in Theorem 4.1 improves upon the rate for the UCBogram by a logarithmic term in n . This comes from the fact that, unlike in [15] where suboptimal bounds for the UCB policy are used, we use here the sharper regret bounds of Corollary 2.2 and the SE policy as a running horse for our policy. The absence of logarithmic terms reveals a surprising phenomenon: there is no price to pay for exploration. Indeed, as discussed after Theorem 5.1, in the case of $K = 2$ arms, the rate in Theorem 4.1 is minimax optimal even in the full information case where at each round, the operator receives the reward from only one arm but observes the rewards of both arms.

PROOF. We assume that $\mathcal{B}_M = \{B_1, \dots, B_{M^d}\}$ where the indexing will be made clearer later in the proof. Moreover, to keep track of positive constants, we number them c_1, c_2, \dots . For any real valued function f on \mathcal{X} and any measurable $A \subseteq \mathcal{X}$, we use the notation $P_X(f \in A) = P_X(f(X) \in A)$. We also write $\mathcal{I} = \{1, \dots, K\}$.

Define $c_1 = 2Ld^{\beta/2} + 1$, and let $n_0 \geq 2$ be the largest integer such that $n_0^{\beta/(2\beta+d)} \leq 2c_1/\delta_0$, where δ_0 is the constant appearing in the margin condition. If $n \leq n_0$, we have $R_n(\bar{\pi}) \leq n_0$ so that the result of the theorem holds when C is chosen large enough, depending on the constant n_0 . In the rest of the proof, we assume that $n > n_0$ so that $c_1M^{-\beta} < \delta_0$.

Recall that the BSE policy $\bar{\pi}$ is a collection of functions $\bar{\pi}_t$ that are constant on each B_j . For any such policy π , define the regret $R_j(\pi)$ on bin B_j by

$$R_j(\pi) = \sum_{t=1}^n E_X \left\{ (f^*(X) - f^{\pi_t(X)}(X)) \mathbb{I}(X \in B_j) \right\},$$

where x_j is an arbitrary element of B_j . Observe that the overall regret of π can be written as

$$R_n(\pi) = \sum_{j=1}^{M^d} R_j(\pi).$$

Consider the set of *well behaved* bins on which the expected reward functions of the arms are well separated. These are the bins B_j with indices in \mathcal{J} defined by

$$\mathcal{J} := \{j \in \{1, \dots, M^d\} \text{ s.t. } \exists x \in B_j, f^*(x) - f^\sharp(x) > c_1M^{-\beta}\}.$$

A bin B that is not well behaved is called *strongly ill behaved* if there is some $x \in B$ such that $f^*(x) = f^\sharp(x)$ and *weakly ill behaved* otherwise. Respectively, the sets of strongly or weakly ill behaved bins have indices in

$$\mathcal{J}_s^c := \left\{ j \in \{1, \dots, M^d\} \text{ s.t. } \exists x \in B_j, f^*(x) = f^\sharp(x) \right\} \quad \text{and}$$

$$\mathcal{J}_w^c := \{j \in \{1, \dots, M^d\} \text{ s.t. } \forall x \in B_j, 0 < f^*(x) - f^\sharp(x) \leq c_1 M^{-\beta}\}.$$

Note that for any $i = 1, \dots, K$, the function $f^* - f^{(i)}$ satisfies the smoothness condition with parameters $(\beta, 2L)$. It implies that for any $j \in \mathcal{J}_s^c$ and any $i = 1, \dots, K$, we have $f^*(x) - f^{(i)}(x) \leq c_1 M^{-\beta}$ for all $x \in B_j$ so that the inclusion $\mathcal{J}_s^c \subset \mathcal{J}^c$ indeed holds.

First part: Strongly ill behaved bins in \mathcal{J}_s^c .

Recall that for any $j \in \mathcal{J}_s^c$, any arm $i \in \mathcal{I}$, and any $x \in B_j$, $f^*(x) - f^{(i)}(x) \leq c_1 M^{-\beta}$. Therefore,

$$\begin{aligned} \sum_{j \in \mathcal{J}_s^c} \mathbb{E} R_j(\bar{\pi}) &\leq c_1 n M^{-\beta} P_X \left\{ 0 < f^*(X) - f^\sharp(X) \leq c_1 M^{-\beta} \right\} \\ &\leq c_1^{1+\alpha} n M^{-\beta(1+\alpha)}, \end{aligned} \quad (4.7)$$

where we used the fact that the set $\{x \in \mathcal{X} : f^*(x) = f^\sharp(x)\}$ does not contribute to the regret.

Second part: Weakly ill behaved bins in \mathcal{J}_s^c .

The numbers of weakly ill behaved bins can be bounded using the fact that $f^*(x) - f^\sharp(x) > 0$ on such a bin; the margin condition along with the fact that P_X has a density imply indeed that

$$\sum_{j \in \mathcal{J}_w^c} \frac{c}{M^d} \leq P_X \left\{ 0 < f^*(X) - f^\sharp(X) \leq c_1 M^{-\beta} \right\} \leq c_1^\alpha M^{-\beta\alpha}.$$

It yields $|\mathcal{J}_w^c| \leq \frac{c_1^\alpha}{c} M^{d-\beta\alpha}$.

We bound the regret on each weakly ill behaved bins using Corollary 2.2 and the specific values $\Delta_{K_0} := \sqrt{\frac{K \log K}{nM^{-d}}}$, $\gamma = 1$ and $T = nM^{-d}$. It implies that there exists a positive constant $c_2 > 0$ such that:

$$\sum_{j \in \mathcal{J}_w^c} \mathbb{E} R_j(\bar{\pi}) \leq |\mathcal{J}_w^c| \sup_{j \in \mathcal{J}_w^c} \mathbb{E} R_j(\bar{\pi}) \leq c_2 \sqrt{K \log(K)} M^{\frac{d}{2}-\beta\alpha} \sqrt{n}. \quad (4.8)$$

Third part: Well behaved bins in \mathcal{J} .

This part is decomposed into two steps. In the first step, we bound the regret in a given bin $B_j, j \in \mathcal{J}$; in the second step we use the margin condition to control the sum of all those regrets.

Step 1. Fix $j \in \mathcal{J}$ and recall that there exists $x_j \in B_j$ such that $f^*(x_j) - f^\sharp(x_j) > c_1 M^{-\beta}$. Define $\mathcal{I}_j^* = \{i \in \mathcal{I} : f^{(i)}(x_j) = f^*(x_j)\}$ and $\mathcal{I}_j^0 = \mathcal{I} \setminus \mathcal{I}_j^* = \{i \in \mathcal{I} : f^*(x_j) - f^{(i)}(x_j) > c_1 M^{-\beta}\}$. We call \mathcal{I}_j^* the set of (almost) *optimal* arms over B_j and \mathcal{I}_j^0 the set of *suboptimal* arms over B_j . Note that $\mathcal{I}_j^0 \neq \emptyset$ for any $j \in \mathcal{J}$.

The smoothness condition implies that for any $i \in \mathcal{I}_j^0, x \in B_j$,

$$f^*(x) - f^{(i)}(x) > c_1 M^{-\beta} - 2L \|x - x_j\|^\beta \geq M^{-\beta}. \quad (4.9)$$

Therefore, $f^* - f^\sharp > 0$ on B_j . Moreover, for $x_0 \in B_j$ such that $f^*(x_0) - f^\sharp(x_0) > c_1 M^{-\beta}$, then $f^*(x_0) = f^{(i)}(x_0)$ for all $i \in \mathcal{I}_j^*$. Therefore for any $x \in B_j$ and any $i \in \mathcal{I}_j^*$, it holds

$$f^*(x) - f^{(i)}(x) \leq c_1 M^{-\beta} \mathbb{1} \left\{ 0 < f^*(x) - f^\sharp(x) \leq c_1 M^{-\beta} \right\}. \quad (4.10)$$

It yields that for any optimal arm $i \in \mathcal{I}_j^*$, the reward functions averaged over B_j satisfy $\bar{f}_j^* - \bar{f}_j^{(i)} \leq c_1 M^{-\beta} q_j$, where

$$q_j := P_X \left\{ 0 < f^* - f^\# \leq c_1 M^{-\beta} \mid X \in B_j \right\}.$$

For any suboptimal arms $i \in \mathcal{I}_j^0$, (4.9) implies that $\underline{\Delta}_j^{(i)} := \bar{f}_j^* - \bar{f}_j^{(i)} > M^{-\beta}$.

Assume now without loss of generality that the average gaps $\underline{\Delta}_j^{(i)}$ are ordered in such a way that $\underline{\Delta}_j^{(1)} \geq \dots \geq \underline{\Delta}_j^{(K)}$. Define

$$K_0 := \operatorname{argmin}_{i \in \mathcal{I}_j^0} \underline{\Delta}_j^{(i)} \quad \text{and} \quad \underline{\Delta}_j := \underline{\Delta}_j^{(K_0)}$$

and observe that if $i \in \mathcal{J}$ is such that $\underline{\Delta}_j^{(i)} < \underline{\Delta}_j$, then $i \in \mathcal{I}_j^*$. Therefore, it follows from (4.10) that $\underline{\Delta}_j^{(i)} \leq c_1 M^{-\beta} q_j$ for such i . Recall that on each bin B_j , the BSE can be seen as a static SE policy with random horizon $N_j(n) = \sum_{t=1}^n \mathbb{I}(X_t \in B_j)$ so that $\mathbb{E}[N_j(n)] \leq \bar{c} n M^{-d}$. Applying Corollary 2.2 with K_0 as above and $\gamma = 1$, we find that there exists a constant $c_3 > 0$ such that, for any $j \in \mathcal{J}$,

$$\begin{aligned} \mathbb{E} R_j(\bar{\pi}) &\leq 612(1 + \bar{c}) \frac{K}{\underline{\Delta}_j} \overline{\log} \left(\frac{n M^{-d} \underline{\Delta}_j^2}{18} \right) + \bar{c} c_1 n M^{-d-\beta} q_j \\ &\leq c_3 \left(\frac{K}{\underline{\Delta}_j} \overline{\log} \left(n M^{-d} \underline{\Delta}_j^2 \right) + n M^{-d-\beta} q_j \right). \end{aligned} \quad (4.11)$$

Step 2. We now use the margin condition to provide lower bounds on $\underline{\Delta}_j$ for each $j \in \mathcal{J}$. Assume without loss of generality that the indexing of the bins is such that $\mathcal{J} = \{1, \dots, j_1\}$ and that the gaps are ordered $0 < \underline{\Delta}_1 \leq \underline{\Delta}_2 \leq \dots \leq \underline{\Delta}_{j_1}$. For any $j \in \mathcal{J}$, from the definition of $\underline{\Delta}_j$, there exists a suboptimal arm $i \in \mathcal{I}_j^0$ such that $\underline{\Delta}_j = \bar{f}_j^* - \bar{f}_j^{(i)}$. But since the function $f^* - f^{(i)}$ satisfies the smoothness condition with parameters $(\beta, 2L)$, we find that if $\underline{\Delta}_j \leq \delta$ for some $\delta > 0$, then

$$0 < f^*(x) - f^{(i)}(x) \leq \delta + 2Ld^{\beta/2} M^{-\beta}, \quad \forall x \in B_j.$$

Together with the fact that $f^* - f^\# > 0$ over B_j for any $j \in \mathcal{J}$ (see Step 1 above), it yields

$$P_X [0 < f^* - f^\# \leq \underline{\Delta}_j + 2Ld^{\beta/2} M^{-\beta}] \geq \sum_{k=1}^{j_1} p_k \mathbb{I}(0 < \underline{\Delta}_k \leq \underline{\Delta}_j) \geq \frac{\underline{c} j}{M^d},$$

where we used the fact that $p_k = P_X(B_k) \geq \underline{c}/M^d$. Define $j_2 \in \mathcal{J}$ to be the largest integer such that $\underline{\Delta}_{j_2} \leq \delta_0/c_1$. Since for any $j \in \mathcal{J}$, we have $\underline{\Delta}_j > M^{-\beta}$, the margin condition yields for any $j \in \{1, \dots, j_2\}$ that,

$$P_X [0 < f^* - f^\# \leq \underline{\Delta}_j + 2Ld^{\beta/2} M^{-\beta}] \leq C_\delta (c_1 \underline{\Delta}_j)^\alpha,$$

where we have used the fact that $\underline{\Delta}_j + 2Ld^{\beta/2} M^{-\beta} \leq c_1 \underline{\Delta}_j \leq \delta_0$, for any $j \in \{1, \dots, j_2\}$. The previous two inequalities, together with the fact that $\underline{\Delta}_j > M^{-\beta}$ for any $j \in \mathcal{J}$, yield

$$\underline{\Delta}_j \geq c_4 \left(\frac{j}{M^d} \right)^{1/\alpha} \vee M^{-\beta} =: \gamma_j, \quad \forall j \in \{1, \dots, j_2\}.$$

Therefore, using the fact that $\underline{\Delta}_j \geq \delta_0/c_1$ for $j \geq j_2$, we get from (4.11) that

$$\sum_{j \in \mathcal{J}} \mathbb{E} R_j(\bar{\pi}) \leq c_5 \left[\sum_{j=1}^{j_2} K \frac{\overline{\log} \left(\frac{n}{M^d} \gamma_j^2 \right)}{\gamma_j} + \sum_{j=j_2+1}^{j_1} K \log(n) + \sum_{j \in \mathcal{J}} n M^{-d-\beta} q_j \right]. \quad (4.12)$$

Fourth part: Putting things together.

Combining (4.7), (4.8) and (4.12), we obtain the following bound,

$$\begin{aligned} \mathbb{E} R_n(\bar{\pi}) \leq c_6 \left[n M^{-\beta(1+\alpha)} + \sqrt{K \log(K)} M^{\frac{d}{2}-\alpha\beta} \sqrt{n} + K \sum_{j=1}^{j_2} \frac{\overline{\log} \left(\frac{n}{M^d} \gamma_j^2 \right)}{\gamma_j} \right. \\ \left. + K M^d \log n + n M^{-d-\beta} \sum_{j \in \mathcal{J}} q_j \right]. \end{aligned} \quad (4.13)$$

We now bound from above the first sum in (4.13) by decomposing it into two terms. From the definition of γ_j , there exists an integer j_3 satisfying $c_7 M^{d-\alpha\beta} \leq j_3 \leq 2c_7 M^{d-\alpha\beta}$ and such that $\gamma_j = M^{-\beta}$ for $j \leq j_3$ and $\gamma_j = c_4 (j M^{-d})^{1/\alpha}$ for $j > j_3$. It holds

$$\sum_{j=1}^{j_3} \frac{\log \left(\frac{n}{M^d} \gamma_j^2 \right)}{\gamma_j} \leq c_8 M^{d+\beta(1-\alpha)} \overline{\log} \left(\frac{n}{M^{2\beta+d}} \right), \quad (4.14)$$

and

$$\begin{aligned} \sum_{j=j_3+1}^{j_2} \frac{\log \left(\frac{n}{M^d} \gamma_j^2 \right)}{\gamma_j} &\leq c_9 \sum_{j=j_3+1}^{M^d} \left(\frac{j}{M^d} \right)^{-\frac{1}{\alpha}} \overline{\log} \left(\frac{n}{M^d} \left[\frac{j}{M^d} \right]^{\frac{2}{\alpha}} \right), \\ &\leq c_{10} M^d \int_{M^{-\alpha\beta}}^1 \log \left(\frac{n}{M^d} x^{\frac{2}{\alpha}} \right) x^{-1/\alpha} dx. \end{aligned} \quad (4.15)$$

Since $\alpha < 1$, this integral is bounded by $c_{10} M^{\beta(1-\alpha)} (1 + \log(n/M^{2\beta+d}))$.

The second sum in (4.13) can be bound as:

$$\begin{aligned} \sum_{j \in \mathcal{J}} q_j &= \sum_{j \in \mathcal{J}} \mathbb{P} \left\{ 0 < f^*(X) - f^\sharp(X) \leq c_1 M^{-\beta} \mid X \in B_j \right\}, \\ &\leq \frac{M^d}{\underline{c}} \mathbb{P} \left\{ 0 < f^*(X) - f^\sharp(X) \leq c_1 M^{-\beta} \right\} \leq \frac{c_1^\alpha}{\underline{c}} M^{d-\beta\alpha}. \end{aligned} \quad (4.16)$$

Putting together (4.13)–(4.16), we obtain:

$$\begin{aligned} \mathbb{E} R_n(\bar{\pi}) \leq c_{11} \left[n M^{-\beta(1+\alpha)} + \sqrt{K \log(K)} M^{\frac{d}{2}-\alpha\beta} \sqrt{n} + K M^{d+\beta(1-\alpha)} \right. \\ \left. + K M^{d+\beta(1-\alpha)} \log \left(\frac{n}{M^{2\beta+d}} \right) + K M^d \log n \right], \end{aligned}$$

and the result follows by choosing M as prescribed. ■

We should point out that the version of the BSE described above specifies the number of bins M as a function of the horizon n , while in practice one may not have foreknowledge of this value. This limitation can be easily circumvented by using the so-called *doubling argument* [5] which consists of “resetting” the game at times $2^k, k = 1, 2, \dots$.

The reader will note that when $\alpha = 1$ there is a potentially superfluous $\log n$ factor appearing in the upper bound using the same proof. More generally, for any $\alpha \geq 1$, it is possible to minimize the expression in (4.13) with respect to M , but the optimal value of M would then depend on the value of α . This sheds some light on a significant limitation of the BSE which surfaces in this parameter regime: it requires the operator to pull each arm at least once in each bin and therefore to incur a regret of at least order M^d . In other words, the BSE splits the space \mathcal{X} in “too many” bins when $\alpha \geq 1$. Intuitively this can be understood as follows. When $\alpha \geq 1$, the gap function $f^*(x) - f^\#(x)$ is bounded away from zero on a subset of \mathcal{X} that has large P_X measure. Hence there is no need to carefully estimate it since the optimal arm is the same on “large” contiguous regions. As a result, one could use larger bins in such regions reducing the overall number of bins and therefore removing the extra logarithmic term alluded to above. These limitations are obviously intrinsic to BSE-type policies.

5 Adaptively Binned Successive Elimination

In this section, we introduce a new algorithm that adaptively constructs a finer and finer partition of $\mathcal{X} = [0, 1]^d$ into smaller hypercubes. We call it Adaptively Binned Successive Elimination (ABSE). As we will see, it improves upon the BSE policy and therefore upon the UCBogram defined in [15], both of which rely on a fixed, uniform partition of \mathcal{X} .

The ABSE policy is denoted by $\tilde{\pi}$. It operates in a manner similar to the BSE except that instead of fixing a partition \mathcal{B}_M , it relies on an adaptive partition that is refined over time. To define the adaptive partition, it will be useful to define the following operation on sets. Recall that \mathcal{B}_{2^k} denotes the regular partition of \mathcal{X} into 2^{kd} hypercubes of side length 2^{-k} and for any bin $B \in \mathcal{B}_{2^k}$, let $\text{burst}(B)$ be the collection of 2^d bins in $\mathcal{B}_{2^{k+1}}$ that form a partition of B . Finally, recall that given a partition of the space into bins, the rewards over each bin can be seen as constant, equal to the average reward and we can run a SE policy on each bin. Assume that $n \geq K \log(K)$ and let k_0 be the smallest integer such that

$$2^{-k_0} \leq \left(\frac{K \log(K)}{n} \right)^{\frac{1}{d+2\beta}}. \quad (5.17)$$

For any bin $B \in \bigcup_{k \geq 0} \mathcal{B}_{2^k}$, let ℓ_B be the smallest integer such that

$$U(\ell_B, n|B|^{-d}) \geq 2c_0|B|^\beta, \quad (5.18)$$

where U is defined in (2.1) and c_0 is a parameter to be chosen later. This definition implies that

$$\ell_B \leq C_\ell |B|^{-2\beta} \log(n|B|^{(2\beta+d)}), \quad (5.19)$$

for some positive constant C_ℓ .

The ABSE policy $\tilde{\pi}$ is better understood using the notion of rooted tree. Indeed, the sequence partitions generated by $\tilde{\pi}$ induces a piecewise constant sequence of nested rooted trees $(\mathcal{T}_t)_{t \geq 1}$ whose nodes are sets in $\bigcup_{k \geq 0} \mathcal{B}_{2^k}$ and such that $\mathcal{T}_t \subseteq \mathcal{T}_{t+1}$. The leaves \mathcal{L}_t of a tree \mathcal{T}_t in this family form the partition of \mathcal{X} at time t and we refer to its elements as *live* bins at time t .

The sequence of trees is constructed as follows. The first tree \mathcal{T}_1 is reduced to its root \mathcal{X} and thus, the original partition is $\{\mathcal{X}\}$. A SE policy $\hat{\pi}_{\mathcal{X}}$ is run on \mathcal{X} for $\ell_{\mathcal{X}}$ rounds with initial set of arms $\mathcal{I}_{\mathcal{X}} = \{1, \dots, K\}$ and parameters $T_{\mathcal{X}} = n, \gamma = 2$. Let $t(\mathcal{X})$ denote the time at which $\hat{\pi}_{\mathcal{X}}$ has reached $\ell_{\mathcal{X}}$ rounds, record the set $\hat{\mathcal{S}}_{\mathcal{X}}$ of arms that are active at this time and split \mathcal{X} into $\text{burst}(\mathcal{X})$. Each set in $\text{burst}(\mathcal{X})$ becomes a child of \mathcal{X} in the tree $\mathcal{T}_{t(\mathcal{X})}$. Between $t = 1$ and $t = t(\mathcal{X}) - 1$, the trees are the same: $\mathcal{T}_1 = \dots = \mathcal{T}_{t(\mathcal{X})-1}$. The sequence $(\mathcal{T}_t)_{t \geq 1}$ is piecewise constant and we simply describe its behavior in-between two change points, that is in-between two bursts. Let $t = t(\bar{B})$ be a time at which a leaf \bar{B} of $\mathcal{T}_{t(\bar{B})-1}$ was burst into $\text{burst}(\bar{B})$. The time $t(\bar{B})$ corresponds to both the *death* of \bar{B} and the *birth* of each bin $B \in \text{burst}(\bar{B})$. Assume further that the SE policy $\hat{\pi}_{\bar{B}}$ had active arms $\hat{\mathcal{S}}_{\bar{B}}$ at time t . For each bin $B \in \text{burst}(\bar{B})$, a SE policy $\hat{\pi}_B$ is run on B for ℓ_B rounds with initial set of arms $\mathcal{I}_B = \hat{\mathcal{S}}_{\bar{B}}$ and parameters $T_B = n|B|^{-d}, \gamma = 2$. Let $t(B)$ denote first time at which the SE policy $\hat{\pi}_B$ on a bin $B \in \text{burst}(\bar{B})$ has reached ℓ_B rounds, record the set $\hat{\mathcal{S}}_B$ of arms that are still active. If $|\hat{\mathcal{S}}_B| \geq 2$ and $|B| \geq 2^{-k_0+1}$, then split B into $\text{burst}(B)$. If B is split, each set in $\text{burst}(B)$ becomes a child of B to form the tree $\mathcal{T}_{t(B)}$. Note that it may be the case that $\mathcal{T}_{t(B)-1} \neq \mathcal{T}_{t(\bar{B})}$ because other branches of $\mathcal{T}_{t(\bar{B})}$ may have been grown between $t(\bar{B})$ and $t(B) - 1$. The procedure is repeated until the horizon n is reached.

Policy 3 Adaptively Binned Successive Elimination (ABSE)

Input: Set of arms $\mathcal{I} = \{1, \dots, K\}$. Parameters $n, c_0 = 2Ld^{\beta/2}, k_0, \ell_0, \dots, \ell_{k_0}$.

Output: $\tilde{\pi}_1, \dots, \tilde{\pi}_n \in \mathcal{I}$.

$t \leftarrow 0, k \leftarrow 0, \mathcal{B} \leftarrow \{\mathcal{X}\}, \mathcal{S}_{\mathcal{X}} \leftarrow \mathcal{I}$

Initialize a SE policy $\hat{\pi}_{\mathcal{X}}$ with parameters $T = n, \gamma = 2$ and arms $\mathcal{I} = \mathcal{S}_{\mathcal{X}}$.

$t_{\mathcal{X}} \leftarrow 0$.

for $t = 1, \dots, n$ **do**

$B \leftarrow \mathcal{B}(X_t)$.

$t_B \leftarrow t_B + 1$.

$\tilde{\pi}_t \leftarrow \hat{\pi}_{B, t_B}$ (observe $Y_t^{(\tilde{\pi}_t)}$).

$\tau_B \leftarrow \hat{\tau}_{B, t_B}$

if $\tau_B \geq \ell_B$ and $|B| \geq 2^{-k_0+1}$ and $|\hat{\mathcal{S}}_{B, t_B}| \geq 2$ **then**

$\mathcal{S} \leftarrow \hat{\mathcal{S}}_{B, t_B}$

for $B' \in \text{burst}(B)$ **do**

Initialize a SE policy $\hat{\pi}_{B'}$ with parameters $T = n|B'|^d, \gamma = 2$ and arms $\mathcal{I} = \mathcal{S}$.

$t_{B'} \leftarrow 0$.

end for

$\mathcal{B} \leftarrow \mathcal{B} \setminus B$

$\mathcal{B} \leftarrow \mathcal{B} \cup \text{burst}(B)$

end if

end for

The ABSE policy, described in Policy 3, satisfies the following theorem.

Theorem 5.1 Fix $\beta \in (0, 1]$, $L > 0$, $\alpha > 0$, assume that $n \geq K \log(K)$ and consider a problem in $\mathcal{M}_{\mathcal{X}}^K(\alpha, \beta, L)$. If $\alpha < \infty$, then the ABSE policy $\tilde{\pi}$ has a regret at time n bounded by,

$$\mathbb{E}R_n(\tilde{\pi}) \leq Cn \left(\frac{K \log(K)}{n} \right)^{\frac{\beta(\alpha+1)}{2\beta+d}},$$

where $C > 0$ is a positive constant that does not depend on K . If $\alpha = \infty$, then $\mathbb{E}R_n(\tilde{\pi}) \leq CK \log(n)$.

Note that the bounds given in Theorem 5.1 are optimal in a minimax sense when $K = 2$. Indeed, the lower bounds of [2] and [15] imply that the bound on the regret cannot be improved as a function of n expect for a constant multiplicative term. Note that the lower bound proved in [2] implies that any policy that received information from *both* arms at each round has a regret bound at least as large as the one from Theorem 5.1, up to a multiplicative constant. As a result, there is no price to pay for being in a partial information setup and one could say that the problem of nonparametric estimation dominates the problem associated to making decisions sequentially.

Note also that when $\alpha = \infty$, Proposition 3.1 implies that there exists a unique optimal arm over \mathcal{X} and that all other arms have reward bounded away from that of the optimal arm. As a result, given this information, one could operate as if the problem was static by simply discarding the covariates. Theorem 5.1 implies that in this case, one recovers the traditional regret bound of the static case without the knowledge that $\alpha = \infty$.

PROOF. We first consider the case where $\alpha < \infty$, which implies that $\alpha\beta \leq d$; see Proposition 3.1.

As in the previous section, we keep track of positive constants by numbering them c_1, c_2, \dots . On each newly created bin B , a new SE policy is initialized and we denote by $Y_{B,1}^{(i)}, Y_{B,2}^{(i)}, \dots$, the rewards obtained by successive pulls of a remaining arm i . Their average after τ rounds/pulls is denoted by

$$\bar{Y}_{B,\tau}^{(i)} := \frac{1}{\tau} \sum_{s=1}^{\tau} Y_{B,s}^{(i)}.$$

For any integer s , define $\varepsilon_{B,s} = 2U(s, n|B|^d)$, where U) is defined in (2.1).

Recall that for any $t = 1, \dots, n$, \mathcal{T}_t is the tree associated to the ABSE policy at time t and that it is a subtree of the following reference rooted tree \mathcal{T}^* whose nodes form the collection of sets $\bigcup_{k=0}^{k_0} \mathcal{B}_{2^k}$. The root of \mathcal{T}^* is \mathcal{X} and each node $B \in \mathcal{T}^*$ such that $|B| < 2^{-k_0}$ has children given by the elements of $\text{burst}(B)$. Let \mathcal{L}^* denote the set of leaves of \mathcal{T}^* , that is the set of bins B such that $|B| = 2^{-k_0}$. Moreover, recall that the leaves \mathcal{L}_t of \mathcal{T}_t form the current partition at time t .

For any $B \in \mathcal{T}^* \setminus \{\mathcal{X}\}$, define the unique *parent* of B by,

$$\mathfrak{p}(B) := \{B' \in \mathcal{T}^* : B \in \text{burst}(B')\}.$$

and $\mathfrak{p}(\mathcal{X}) = \emptyset$. Moreover, let $\mathfrak{p}^1(B) = \mathfrak{p}(B)$ and for any $k \geq 2$ define recursively $\mathfrak{p}^k(B) = \mathfrak{p}(\mathfrak{p}^{k-1}(B))$. Then the set of *ancestors* of any $B \in \mathcal{T}^* \setminus \{\mathcal{X}\}$ is denoted by $\mathcal{P}(B)$ and defined by

$$\mathcal{P}(B) = \{B' \in \mathcal{T}^* : B' = \mathfrak{p}^k(B) \text{ for some } k \geq 1\}.$$

Moreover, we define $\mathcal{P}(\mathcal{X}) := \emptyset$. Denote by $r_n^{\text{live}}(B)$ the regret generated by the ABSE policy $\tilde{\pi}$ when covariate X_t fell in a *live* bin $B \in \mathcal{L}_t$. It is defined by

$$r_n^{\text{live}}(B) = \sum_{t=1}^n [f^*(X_t) - f^{(\tilde{\pi}_t(X_t))}(X_t)] \mathbb{I}(X_t \in B) \mathbb{I}(B \in \mathcal{L}_t).$$

Denote also by $r_n^{\text{born}}(B)$ the regret generated when covariate X_t fell in a *xbin* B that was born (and possibly died) before time t . In particular live bins are born bins. It is defined by

$$r_n^{\text{born}}(B) = \sum_{t=1}^n [f^*(X_t) - f^{(\tilde{\pi}_t(X_t))}(X_t)] \mathbb{I}(X_t \in B) \mathbb{I}(\mathcal{P}(B) \cap \mathcal{B} = \emptyset).$$

Observe that if we define $\tilde{r}_n := r_n^{\text{born}}(\mathcal{X})$, we have $\mathbb{E}R_n(\tilde{\pi}) = \mathbb{E}\tilde{r}_n$ since $\mathcal{P}(\mathcal{X}) = \emptyset$ and $X_t \in \mathcal{X}$ for all t .

Denote by $\mathcal{I}_B = \hat{\mathcal{S}}_{B,t_B}$ the set of arms left active by the SE policy $\hat{\pi}_B$ on B at the end of ℓ_B rounds. Moreover, define the following reference sets of arms:

$$\begin{aligned} \underline{\mathcal{I}}_B &:= \left\{ i \in \{1, \dots, K\} : \sup_{x \in B} f^*(x) - f^{(i)}(x) \leq c_0 |B|^\beta \right\} \text{ and} \\ \bar{\mathcal{I}}_B &:= \left\{ i \in \{1, \dots, K\} : \sup_{x \in B} f^*(x) - f^{(i)}(x) \leq 8c_0 |B|^\beta \right\}. \end{aligned}$$

Note that for any $B \in \mathcal{T}^*$,

$$r_n^{\text{born}}(B) = r_n^{\text{live}}(B) + \sum_{B' \in \text{burst}(B)} r_n^{\text{born}}(B').$$

Define the event $\mathcal{A}_B := \{\underline{\mathcal{I}}_B \subseteq \mathcal{I}_B \subseteq \bar{\mathcal{I}}_B\}$ on which the remaining arms have a gap of the correct order and observe that the above display implies that

$$r_n^{\text{born}}(B) = r_n^{\text{born}}(B) \mathbb{I}(\mathcal{A}_B^c) + r_n^{\text{live}}(B) \mathbb{I}(\mathcal{A}_B) + \sum_{B' \in \text{burst}(B)} r_n^{\text{born}}(B') \mathbb{I}(\mathcal{A}_B).$$

As a result, the quantity we are interested in can be decomposed as $\tilde{r}_n = \tilde{r}_n(\mathcal{T}^* \setminus \mathcal{L}^*) + \tilde{r}_n(\mathcal{L}^*)$ where

$$\tilde{r}_n(\mathcal{T}^* \setminus \mathcal{L}^*) := \sum_{B \in \mathcal{T}^* \setminus \mathcal{L}^*} \left(r_n^{\text{born}}(B) \mathbb{I}(\mathcal{A}_B^c) + r_n^{\text{live}}(B) \mathbb{I}(\mathcal{A}_B) \right) \prod_{B' \in \mathcal{P}(B)} \mathbb{I}(\mathcal{A}_{B'}),$$

is the regret on the non-terminal nodes and

$$\tilde{r}_n(\mathcal{L}^*) := \sum_{B \in \mathcal{L}^*} r_n^{\text{born}}(B) \prod_{B' \in \mathcal{P}(B)} \mathbb{I}(\mathcal{A}_{B'}) = \sum_{B \in \mathcal{L}^*} r_n^{\text{live}}(B) \prod_{B' \in \mathcal{P}(B)} \mathbb{I}(\mathcal{A}_{B'})$$

is the regret on the leaves. Our proof relies on the events $\mathcal{G}_B := \bigcap_{B' \in \mathcal{P}(B)} \mathcal{A}_{B'}$.

First part: control of the regret on the non-terminal nodes

Fix $B \in \mathcal{T}^* \setminus \mathcal{L}^*$. On \mathcal{G}_B , we have $\mathcal{I}_{\mathbf{p}(B)} \subseteq \bar{\mathcal{I}}_{\mathbf{p}(B)}$ so that any active arm $i \in \mathcal{I}_{\mathbf{p}(B)}$ satisfies $\sup_{x \in \mathcal{p}(B)} |f^*(x) - f^{(i)}(x)| \leq 8c_0 |\mathbf{p}(B)|^\beta$. Defining $c_1 := 2^{3+\beta} c_0$, it yields

$$\mathbb{E} \left[r_n^{\text{live}}(B) \mathbb{I}(\mathcal{G}_B \cap \mathcal{A}_B) \right] \leq c_1 K \ell_B |B|^\beta q_B,$$

where $q_B = P_X(0 < f^* - f^\# \leq c_1 |B|^\beta \mid X \in B)$. Applying the same argument as in the proof of Theorem 4.1 when bounding the second sum in (4.13) yields that, for any $k \in \{0, \dots, k_0\}$,

$$\sum_{|B|=2^{-k}} q_B \leq \frac{c_1^\alpha}{\underline{c}} 2^{k(d-\beta\alpha)}.$$

Therefore, summing over $\mathcal{T}^* \setminus \mathcal{L}^*$, we obtain

$$\mathbb{E} \left[\sum_{B \in \mathcal{T}^* \setminus \mathcal{L}^*} r_n^{\text{live}}(B) \mathbb{I}(\mathcal{G}_B \cap \mathcal{A}_B) \right] \leq \frac{c_1^{1+\alpha}}{\underline{c}} C_\ell K \sum_{k=0}^{k_0} 2^{k(d+\beta-\alpha\beta)} \log \left(n 2^{-(2\beta+d)} \right). \quad (5.20)$$

On the other hand, for every bin $B \in \mathcal{T}^* \setminus \mathcal{L}^*$, one also has

$$\mathbb{E} \left[r_n^{\text{born}}(B) \mathbb{I}(\mathcal{G}_B \cap \mathcal{A}_B^c) \right] \leq c_1 n |B|^\beta q_B P_X(B) \mathbb{P}(\mathcal{G}_B \cap \mathcal{A}_B^c). \quad (5.21)$$

It remains to control the probability of $\mathcal{G}_B \cap \mathcal{A}_B^c$. Note that $\mathbb{P}(\mathcal{G}_B \cap \mathcal{A}_B^c) \leq \mathbb{P}^{\mathcal{G}_B}(\mathcal{A}_B^c)$ where $\mathbb{P}^{\mathcal{G}_B}$ denotes the conditional probability $\mathbb{P}^{\mathcal{G}_B}(\cdot) := \mathbb{P}(\cdot \mid \mathcal{G}_B)$. Conditionally on \mathcal{G}_B , the event \mathcal{A}_B^c can occur in two ways:

- (i) By eliminating an arm $i \in \underline{\mathcal{I}}_B$ at the end of the at most ℓ_B rounds played on bin B . These arms satisfy $\sup_{x \in B} f^*(x) - f^{(i)}(x) < c_0 |B|^\beta$; this event is denoted by \mathcal{B}_B^1 .
- (ii) By not eliminating an arm $i \notin \bar{\mathcal{I}}_B$ within the at most ℓ_B rounds played on bin B . These arms satisfy $\sup_{x \in B} f^*(x) - f^{(i)}(x) \geq 8c_0 |B|^\beta$; this event is denoted by \mathcal{B}_B^2 .

We use the following decomposition

$$\mathbb{P}^{\mathcal{G}_B}(\mathcal{A}_B^c) = \mathbb{P}^{\mathcal{G}_B}(\mathcal{B}_B^1) + \mathbb{P}^{\mathcal{G}_B}(\mathcal{B}_B^2 \cap (\mathcal{B}_B^1)^c). \quad (5.22)$$

We first control the probability of making error (i). Note that for any $s \leq \ell_B$ and any arms $i \in \underline{\mathcal{I}}_B, i' \in \mathcal{I}_{\mathbf{p}(B)}$, it holds

$$\bar{f}_B^{(i')} - \bar{f}_B^{(i)} \leq \bar{f}_B^* - \bar{f}_B^{(i)} < c_0 |B|^\beta \leq \varepsilon_{B, \ell_B}.$$

Therefore, if an arm $i \in \underline{\mathcal{I}}_B$ is eliminated, that is if there exists $i' \in \mathcal{I}_{\mathbf{p}(B)}$ such that $\bar{Y}_{B,s}^{(i')} - \bar{Y}_{B,s}^{(i)} > \varepsilon_{B,s}$ for some $s \leq \ell_B$, then either $\bar{f}_B^{(i)}$ or $\bar{f}_B^{(i')}$ does not belong to its respective confidence interval $\left[\bar{Y}_{B,s}^{(i)} \pm \frac{\varepsilon_{B,s}}{2} \right]$ or $\left[\bar{Y}_{B,s}^{(i')} \pm \frac{\varepsilon_{B,s}}{2} \right]$ for some $s \leq \ell_B$. Therefore,

$$\mathbb{P}^{\mathcal{G}_B}(\mathcal{B}_B^1) \leq \mathbb{P} \left\{ \exists s \leq \ell_B; \exists i \in \mathcal{I}_{\mathbf{p}(B)}; \left| \bar{Y}_s^{(i)} - \bar{f}_B^{(i)} \right| \geq \frac{\varepsilon_{B,s}}{2} \right\} \leq 2K \frac{\ell_B}{n |B|^d}, \quad (5.23)$$

where in the second inequality, we used Lemma A.1.

Next, we treat error (ii). For any $i \notin \bar{\mathcal{I}}_B$, there exists $x^{(i)}$ such that $f^*(x^{(i)}) - f^{(i)}(x^{(i)}) > 8c_0|B|^\beta$. Let $\check{i} = \check{i}(i) \in \{1, \dots, K\}$ be any arm such that $f^*(x^{(i)}) = f^{(\check{i})}(x^{(i)})$; the smoothness condition implies that

$$\bar{f}_B^{(\check{i})} \geq f^{(\check{i})}(x^{(i)}) - c_0|B|^\beta > f^{(i)}(x^{(i)}) + 7c_0|B|^\beta \geq \bar{f}_B^{(i)} + 6c_0|B|^\beta. \quad (5.24)$$

On the event $(\mathcal{B}_B^1)^c$, no arm in $\underline{\mathcal{I}}_B$, and in particular any of the arms $\check{i}(i), i \in \mathcal{I}_{\mathfrak{p}(B)} \setminus \bar{\mathcal{I}}_B$ has been eliminated until round ℓ_B . Therefore, the event $\mathcal{B}_B^2 \cap (\mathcal{B}_B^1)^c$ occurs if there exists $i \notin \bar{\mathcal{I}}_B$ such that $\bar{Y}_{B, \ell_B}^{(\check{i})} - \bar{Y}_{B, \ell_B}^{(i)} \leq 2\varepsilon_{B, \ell_B}$. In view of (5.24) and (5.18), it implies that there exists $i \in \mathcal{I}_{\mathfrak{p}(B)}$ such that

$$|\bar{Y}_{B, \ell_B}^{(i)} - \bar{f}_B^{(i)}| \geq \frac{\varepsilon_{B, s}}{2}.$$

Hence, the probability of error (ii) can be bounded by

$$\mathbb{P}^{\mathcal{G}_B}(\mathcal{B}_B^2 | (\mathcal{B}_B^1)^c) \leq \mathbb{P} \left\{ \exists i \in \mathcal{I}_{\mathfrak{p}(B)} : |\bar{Y}_{B, \ell_B}^{(i)} - \bar{f}_B^{(i)}| \geq \frac{\varepsilon_{B, s}}{2} \right\} \leq 2K \frac{\ell_B}{n|B|^d}, \quad (5.25)$$

where the second inequality is a consequence of the Hoeffding-Azuma inequality (A.1).

Putting together (5.22), (5.23), (5.25) and (5.19), we get

$$\mathbb{P}^{\mathcal{G}_B}(\mathcal{A}_B^c) \leq 4K \frac{\ell_B}{n|B|^d} \leq 4C_\ell \frac{K}{n} |B|^{-(2\beta+d)} \log(n|B|^{(2\beta+d)}).$$

Together with (5.21), it yields that the regret on any $B \in \mathcal{T}^* \setminus \mathcal{L}^*$ is bounded by

$$\mathbb{E} \left[r_n^{\text{born}}(B) \mathbb{I}(\mathcal{G}_B \cap \mathcal{A}_B^c) \right] \leq c_3 K |B|^{-(\beta+d)} \log(n|B|^{(2\beta+d)}) q_B P_X(B)$$

If k is an integer such that $c_1 2^{-k\beta} > \delta_0$, then any bin B such that $|B| = 2^{-k}$ satisfies $\mathbb{E} \left[r_n^{\text{born}}(B) \mathbb{I}(\mathcal{G}_B \cap \mathcal{A}_B^c) \right] \leq c_4 K \log n$. If k is an integer such that $c_1 2^{-k\beta} \leq \delta_0$, then the above display together with the margin condition yield

$$\mathbb{E} \left[\sum_{|B|=2^{-k}} r_n^{\text{born}}(B) \mathbb{I}(\mathcal{G}_B \cap \mathcal{A}_B^c) \right] \leq c_5 K 2^{k(\beta+d-\alpha\beta)} \log(n 2^{-k(2\beta+d)}).$$

Summing over all $B \in \mathcal{T}^* \setminus \mathcal{L}^*$ and using (5.20), we obtain

$$\mathbb{E}[\tilde{r}_n(\mathcal{T}^* \setminus \mathcal{L}^*)] \leq c_6 K \sum_{k=0}^{k_0} 2^{k(\beta+d-\alpha\beta)} \log(n 2^{-k(2\beta+d)}). \quad (5.26)$$

We now compute an upper bound on the right-hand side of the above inequality. Fix $k = 0, \dots, k_0$ and define

$$S_k = \sum_{j=0}^k 2^{j(d+\beta-\beta\alpha)} = \frac{2^{k(d+\beta-\beta\alpha)} - 1}{2^{d+\beta-\beta\alpha} - 1}.$$

Observe that

$$2^{k(d+\beta-\beta\alpha)} \log \left(n2^{-k(d+2\beta)} \right) = (S_k - S_{k-1}) \log \left(n[c_7 S_{k+1} + 1]^{-\frac{d+2\beta}{d+\beta-\beta\alpha}} \right),$$

where $c_7 := 2^{d+\beta-\beta\alpha} - 1$. Therefore, (5.26) can be rewritten as:

$$\begin{aligned} \mathbb{E}[\tilde{r}_n(\mathcal{T}^* \setminus \mathcal{L}^*)] &\leq c_6 K \left[\sum_{k=1}^{k_0} (S_k - S_{k-1}) \log \left(n[c_7 S_{k+1} + 1]^{-\frac{d+2\beta}{d+\beta-\beta\alpha}} \right) + \log n \right] \\ &\leq c_6 K \left[\int_0^{S_{k_0}} \log \left(n[c_7 x + 1]^{-\frac{d+2\beta}{d+\beta-\beta\alpha}} \right) dx + \log n \right] \\ &\leq c_8 K \left[2^{k_0(d+\beta-\beta\alpha)} \log \left(n2^{-k_0(d+2\beta)} \right) + \log n \right] \\ &\leq c_9 n \left(\frac{n}{K \log(K)} \right)^{-\frac{\beta(1+\alpha)}{d+2\beta}}, \end{aligned} \tag{5.27}$$

where we used (5.17) in the last inequality.

Second part: control of the regret on the leaves

Recall that the set of leaves \mathcal{L}^* is composed of bins B such that $|B| = 2^{-k_0}$. Proceeding in the same way as in (5.21), we find that for any $B \in \mathcal{L}^*$, it holds

$$\mathbb{E} \left[r_n^{\text{live}}(B) \mathbb{1}(\mathcal{G}_B) \right] \leq c_1 n |B|^\beta P_X(0 < f^\star - f^\sharp \leq c_1 |B|^\beta, X \in B).$$

If $c_2^{-k_0} \leq \delta_0$, then using the margin assumption, we find

$$\sum_{B \in \mathcal{L}^*} \mathbb{E} \left[r_n^{\text{live}}(B) \mathbb{1}(\mathcal{G}_B) \right] \leq c_1 n 2^{-k_0 \beta (1+\alpha)} \leq c_1 n \left(\frac{n}{K \log(K)} \right)^{-\frac{\beta(1+\alpha)}{d+2\beta}}, \tag{5.28}$$

where we used (5.17) in the second inequality. If $c_2^{-k_0} > \delta_0$, then there exists a constant $c_{10} > 0$ such that

$$\left(\frac{n}{K \log(K)} \right)^{-\frac{1}{d+2\beta}} \leq c_{10},$$

so that (5.28) also holds but with a different constant.

The theorem follows by summing (5.27) and (5.28). If $\alpha = +\infty$, then the same proof holds except that $\log(n)$ dominates $2^{k(\beta+d-\alpha\beta)} \log(n2^{-k(2\beta+d)})$ for every k . \blacksquare

A Technical lemma

The following lemma is central to our proof of Theorem 2.1. We recall that a process Z_t is a martingale difference sequence if $\mathbb{E}[Z_{t+1} | Z_1, \dots, Z_t] = 0$. Moreover, if $-1 \leq Z_t \leq 1$ and if we

denote the sequence of averages by $\bar{Z}_t = \frac{1}{t} \sum_{s=1}^t Z_s$, then Hoeffding-Azuma's inequality yields that, for every integer $T \geq 1$,

$$\mathbb{P} \left\{ \bar{Z}_T \geq \sqrt{\frac{2}{T} \log \left(\frac{1}{\delta} \right)} \right\} \leq \delta. \quad (\text{A.1})$$

The following lemma is a generalization of this result:

Lemma A.1 *Let Z_t be a martingale difference sequence with $-1 \leq Z_t \leq 1$ then, for every $\delta > 0$ and every integer $T \geq 1$,*

$$\mathbb{P} \left\{ \exists t \leq T, \bar{Z}_t \geq 2 \sqrt{\frac{2}{t} \log \left(\frac{4T}{\delta t} \right)} \right\} \leq \delta.$$

PROOF. Define $\varepsilon_t = 2 \sqrt{\frac{2}{t} \log \left(\frac{4T}{\delta t} \right)}$. We first recall Hoeffding-Azuma's maximal concentration inequality (based on Doob's inequality instead of Chernoff's one). For every $\eta > 0$ and every integer $t \geq 1$,

$$\mathbb{P} \left\{ \exists s \leq t, s \bar{Z}_s \geq \eta \right\} \leq \exp \left(-\frac{\eta^2}{2t} \right).$$

Using a peeling argument, one obtains

$$\begin{aligned} \mathbb{P} \left\{ \exists t \leq T, \bar{Z}_t \geq \varepsilon_t \right\} &\leq \sum_{m=1}^{\lfloor \log(T) \rfloor} \mathbb{P} \left\{ \exists 2^m \leq t < 2^{m+1}, \bar{Z}_t \geq \varepsilon_t \right\} \\ &\leq \sum_{m=1}^{\lfloor \log(T) \rfloor} \mathbb{P} \left\{ \exists 2^m \leq t < 2^{m+1}, \bar{Z}_t \geq \varepsilon_{2^{m+1}} \right\} \\ &\leq \sum_{m=1}^{\lfloor \log(T) \rfloor} \mathbb{P} \left\{ \exists 2^m \leq t < 2^{m+1}, t \bar{Z}_t \geq 2^m \varepsilon_{2^{m+1}} \right\} \\ &\leq \sum_{m=1}^{\lfloor \log(T) \rfloor} \exp \left(-\frac{(2^m \varepsilon_{2^{m+1}})^2}{2 \cdot 2^{m+1}} \right) = \sum_{m=1}^{\lfloor \log(T) \rfloor} \frac{2^{m+1} \delta}{T} \frac{1}{4} \\ &\leq \frac{2^{\log(T)+2} \delta}{T} \frac{1}{4} \leq \delta \end{aligned}$$

Hence the result. ■

References

- [1] J.-Y. Audibert and S. Bubeck. Regret bounds and minimax policies under partial monitoring. *The Journal of Machine Learning Research*, 9999:2785–2836, 2010.

- [2] J.-Y. Audibert and A. B. B. Tsybakov. Fast learning rates for plug-in classifiers. *Ann. Statist.*, 35(2):608–633, 2007.
- [3] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2-3):235–256, 2002.
- [4] P. Auer and R. Ortner. Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1):55–65, 2010.
- [5] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, Cambridge, 2006.
- [6] E. Even-Dar, S. Mannor, and Y. Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *The Journal of Machine Learning Research*, 7:1079–1105, 2006.
- [7] A. Goldenshluger and A. Zeevi. Woodroffe’s one-armed bandit problem revisited. *Ann. Appl. Probab.*, 19(4):1603–1633, 2009.
- [8] A. Goldenshluger and A. Zeevi. Linear response two-armed bandits. *Unpublished.*, 2010.
- [9] E. Hazan and N. Megiddo. Online learning with prior knowledge. In *Learning theory*, volume 4539 of *Lecture Notes in Comput. Sci.*, pages 499–513. Springer, Berlin, 2007.
- [10] S. Kakade, S. Shalev-Shwartz, and A. Tewari. Efficient bandit algorithms for online multiclass prediction. In Andrew McCallum and Sam Roweis, editors, *Proceedings of the 25th Annual International Conference on Machine Learning (ICML 2008)*, pages 440–447. Omnipress, 2008.
- [11] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Adv. in Appl. Math.*, 6(1):4–22, 1985.
- [12] J. Langford and T. Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 817–824. MIT Press, Cambridge, MA, 2008.
- [13] T. Lu, D. Pál, and M. Pál. Showing relevant ads via lipschitz context multi-armed bandits. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, 2010.
- [14] E. Mammen and A. B. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27(6):1808–1829, 1999.
- [15] P. Rigollet and A. Zeevi. Nonparametric bandits with covariates. In Adam Tauman Kalai and Mehryar Mohri, editors, *COLT*, pages 54–66. Omnipress, 2010.
- [16] H. Robbins. Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.*, 58:527–535, 1952.

- [17] A. Slivkins. Contextual bandits with similarity information. In S. Kakade and U. von Luxburg, editors, *24th Annual Conference on Learning Theory*. 2011.
- [18] A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32(1):135–166, 2004.
- [19] C.-C. Wang, S.R. Kulkarni, and H.V. Poor. Bandit problems with side observations. *Automatic Control, IEEE Transactions on*, 50(3):338–355, March 2005.
- [20] M. Woodroofe. A one-armed bandit problem with a concomitant variable. *J. Amer. Statist. Assoc.*, 74(368):799–806, 1979.
- [21] Y. Yang and D. Zhu. Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates. *Ann. Statist.*, 30(1):100–121, 2002.