The Annals of Statistics
2013, Vol. 41, No. 2, 693–721
DOI: 10.1214/13-AOS1101

THE MULTI-ARMED BANDIT PROBLEM WITH COVARIATES

BY VIANNEY PERCHET[1] AND PHILIPPE RIGOLLET[2]

Université Paris Diderot and Princeton University

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.

1. Introduction. The seminal paper [19] 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.

When the populations being sampled are homogeneous, that is, when the sequential rewards are independent and identically distributed (i.i.d.) in each arm, the family of upper-confidence-bound (UCB) policies, introduced in [14], incur a regret of order $\log n$, where n is the length of the time horizon, and no other

Received June 2012; revised January 2013.
[1]Supported in part by the ANR Grant ANR-10-BLAN-0112.
[2]Supported in part by the NSF Grants DMS-09-06424, DMS-10-53987.
MSC2010 subject classifications. Primary 62G08; secondary 62L12.
Key words and phrases. Nonparametric bandit, contextual bandit, multi-armed bandit, adaptive partition, successive elimination, sequential allocation, regret bounds.

"good" policy can (asymptotically) achieve a smaller regret; see also [4]. The elegance of the theory and sharp results developed in [14] 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 [24] under parametric assumptions and in a somewhat restricted setting; see [9, 10] and [23] 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 [25]. 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., [4]) and show that the average regret tends to zero as the time horizon n grows to infinity. However, it is unclear whether this policy satisfies 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 [18] 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 [18] 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 [18] suffered two 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 other policies that deal with covariates. During a first phase, this policy explores the different arms, builds estimates and eliminates sequentially suboptimal arms; when only one arm remains, it is pulled until the horizon is reached. A variant of SE was originally introduced in [8]. 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 [14] and [4].

(2) Binned Successive Elimination (BSE) follows a simple principle to solve the problem with covariates. It consists of grouping similar covariates into bins and then looks 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 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 covariates. In the particular case where $K = 2$ and the problem is *difficult*, these bounds improve upon the results of [18] 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 [18], 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 dominated 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.

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 [15] and also [13]. A related problem online convex optimization with side information was studied in [11], 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 [16] and [20], 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 [8] in the traditional setup of multi-armed bandit problems. As opposed to the more popular UCB policy (see, e.g., [4, 14]), it allows 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 [4]. Such an improvement was recently attempted in [5] 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, for example, [4]. 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 rewards $Y_1^{(i)}, Y_2^{(i)}, \dots$ that are i.i.d. 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 $f^{(1)} \leq \cdots \leq f^{(K+1)}$ so that $K+1$ is one of the best arms. For simplicity, we further assume that the best arm is *unique* since for the SE policy, having multiple optimal arms only improves the regret bound. In the analysis, it is convenient to denote this optimal arm by $* := K+1$ and to define the *gaps* traditionally denoted by $\Delta_1 \geq \cdots \geq \Delta_* = 0$, by $\Delta_i = f^{(*)} - f^{(i)} \geq 0$.

A *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

$$R_n(\pi) := \sum_{t=1}^{n} \bigl(f^{(*)} - f^{(\pi_t)}\bigr).$$

Note that for a data-driven policy $\hat{\pi}$, this quantity is random and, in the rest of the paper, we provide upper bounds on $\mathbb{E}R(\hat{\pi})$. Such bounds are referred to as *regret bounds*.

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 $\mathcal{I}_\tau$ remains. During round τ, each arm in $\mathcal{I}_\tau$ is pulled exactly once (EXPLORATION).

Policy 1 Successive Elimination (SE)

Input: Set of arms $\mathcal{I} = \{1, \dots, K\}$; parameters T, γ; horizon n.
Output: $(\hat{\pi}_1, \hat{\tau}_1, \hat{\mathbb{I}}_1), (\hat{\pi}_2, \hat{\tau}_2, \hat{\mathbb{I}}_2), \dots \in \mathcal{I} \times \mathbb{N} \times \mathcal{P}(\mathcal{I})$.

```
τ ← 1, S ← I, t ← 0, Ȳ ← (0, ..., 0) ∈ [0, 1]^K
loop
  Ȳ^max ← max{Ȳ^(i) : i ∈ S}
  for i ∈ S do
    if Ȳ^(i) ≥ Ȳ^max − γ U(τ, T) then
      t ← t + 1
      π̂_t ← i (observe Y^(i))                                   EXPLORATION
      Î_t ← S, τ̂_t ← τ
      Ȳ^(i) ← (1/τ)[(τ − 1)Ȳ^(i) + Y^(i)]
    else
      S ← S \ {i}.                                              ELIMINATION
    end if
  end for
  τ ← τ + 1.
end loop
```

At the end of the round, for each remaining arm in $\mathcal{I}_\tau$, we decide whether to eliminate it 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, which is parameterized by two quantities $T \in \mathbb{N}$ and $\gamma > 0$ and 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 perfect knowledge of the horizon n and in particular, when the horizon is a random variable with expected value n; nevertheless, in the static case, it is manifest from our result that, when the horizon is known to be n, choosing $T = n$ is always the best choice when possible and that other choices may lead to suboptimal results.

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 $i \in \mathcal{I}_\tau$ at round τ that is defined by $\bar{Y}^{(i)}(\tau) = (1/\tau) \sum_{t=1}^{\tau} Y_\tau^{(i)}$, where here and throughout this paper, we use the convention $1/0 = \infty$. In the rest of the paper, log denotes the natural logarithm and $\overline{\log}(x) = \log(x) \vee 1$. For any positive integer T, define also

$$U(\tau, T) = 2\sqrt{\frac{2\overline{\log}(T/\tau)}{\tau}}, \tag{2.1}$$

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 pseudo-code 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 $\hat{\mathbb{I}}_t \in \mathcal{P}(\mathcal{I})$ of arms that are active at time t, where $\mathcal{P}(\mathcal{I})$ denotes the power set of $\mathcal{I}$.

The following theorem gives a first upper bound on the expected regret of the SE policy.

THEOREM 2.1. *Consider a $(K+1)$-armed bandit problem where horizon is a random variable N of expectation n that is independent of the random rewards. When implemented with parameters $T, \gamma \geq 1$, the* SE *policy $\hat{\pi}$ exhibits an expected regret bounded, for any $\Delta \geq 0$, as*

$$\mathbb{E}\big[R_N(\hat{\pi})\big] \leq 392\gamma^2\bigg(1 + \frac{n}{T}\bigg)\frac{K}{\Delta}\overline{\log}\bigg(\frac{T\Delta^2}{18\gamma^2}\bigg) + n\Delta^-,$$

where Δ^- is the largest Δ_j such that $\Delta_j < \Delta$ if it exists, otherwise $\Delta^- = 0$.

PROOF. Assume without loss of generality that $\Delta_j > 0$, for $j \geq 1$ since arms j such that $j = 0$ do not contribute to the regret. Define $\varepsilon_\tau = U(\tau, T)$. Moreover, for any i in the set $\mathcal{I}_\tau$ of arms that remain active at the beginning of round τ, define $\hat{\Delta}_i(\tau) := \bar{Y}^{(*)}(\tau) - \bar{Y}^{(i)}(\tau)$. Recall that, at round τ, if arms $i, * \in \mathcal{I}_\tau$, then (i) the optimal arm $*$ eliminates arm i if $\hat{\Delta}_i(\tau) \geq \gamma\varepsilon_\tau$, and (ii) arm i eliminates arm $*$ if $\hat{\Delta}_i(\tau) \leq -\gamma\varepsilon_\tau$.

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

$$\Delta_i = \frac{3}{2}\gamma\varepsilon_{\tau_i^*} = 3\gamma\sqrt{\frac{2}{\tau_i^*}\overline{\log}\bigg(\frac{T}{\tau_i^*}\bigg)} \qquad \text{so that } \tau_i \leq \tau_i^* + 1 \leq \frac{18\gamma^2}{\Delta_i^2}\overline{\log}\bigg(\frac{T\Delta_i^2}{18\gamma^2}\bigg) + 1.$$

Note that $1 \leq \tau_1 \leq \cdots \leq \tau_K$ as well as the bound

$$\tau_i \leq \frac{19\gamma^2}{\Delta_i^2}\overline{\log}\bigg(\frac{T\Delta_i^2}{18\gamma^2}\bigg). \tag{2.2}$$

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 $*$ 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 $*$).

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

$$\mathcal{A}_i = \{\text{the arm } * \text{ has not been eliminated before round } \tau_i\};$$
$$\mathcal{B}_i = \big\{\text{every arm } j \in \{1,\dots,i\} \text{ has been eliminated before round } \tau_j\big\}.$$

Moreover, define $\mathcal{C}_i = \mathcal{A}_i \cap \mathcal{B}_i$ and observe that $\mathcal{C}_1 \supseteq \mathcal{C}_2 \supseteq \cdots \supseteq C_K$. 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 sample space denoted by $\mathcal{C}_0$ into the disjoint union $(\mathcal{C}_0 \setminus \mathcal{C}_1) \cup \cdots \cup (\mathcal{C}_{K_0-1} \setminus \mathcal{C}_{K_0}) \cup \mathcal{C}_{K_0}$ where $K_0 \in \{1,\dots,K\}$ is chosen later. It implies the following decomposition of the expected regret:

$$\mathbb{E}R_N(\hat{\pi}) \leq \sum_{i=1}^{K_0} n\Delta_i \mathbb{P}(\mathcal{C}_{i-1} \setminus \mathcal{C}_i) + \sum_{i=1}^{K_0} \tau_i \Delta_i + n\Delta_{K_0+1}. \tag{2.3}$$

Define by A^c the complement of an event A. Note that the first term on the right-hand side of the above inequality can be decomposed as follows:

$$\begin{aligned}\sum_{i=1}^{K_0} n\Delta_i \mathbb{P}(\mathcal{C}_{i-1} \setminus \mathcal{C}_i) &= n\sum_{i=1}^{K_0} \Delta_i \mathbb{P}\big(\mathcal{A}_i^c \cap \mathcal{C}_{i-1}\big) \\ &\quad + n\sum_{i=1}^{K_0} \Delta_i \mathbb{P}\big(\mathcal{B}_i^c \cap \mathcal{A}_i \cap \mathcal{B}_{i-1}\big),\end{aligned} \tag{2.4}$$

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 $\mathcal{A}_i \subseteq \mathcal{A}_{i-1}$.

From Hoeffding's inequality, we have that for every $\tau \geq 1$,

$$\begin{aligned}\mathbb{P}\big(\hat{\Delta}_i(\tau) < \gamma\varepsilon_\tau\big) &= \mathbb{P}\big(\hat{\Delta}_i(\tau) - \Delta_i < \gamma\varepsilon_\tau - \Delta_i\big) \\ &\leq \exp\bigg(-\frac{\tau(\Delta_i - \gamma\varepsilon_\tau)^2}{2}\bigg).\end{aligned} \tag{2.5}$$

On the event $\mathcal{B}_i^c \cap \mathcal{A}_i \cap \mathcal{B}_{i-1}$, arm $*$ has not eliminated arm i at τ_i. Therefore $\mathbb{P}(\mathcal{B}_i^c \cap \mathcal{A}_i \cap \mathcal{B}_{i-1}) \leq \mathbb{P}(\hat{\Delta}_i(\tau_i) < \gamma\varepsilon_{\tau_i})$. Together with the above display with $\tau = \tau_i$, it yields

$$\mathbb{P}\big(\mathcal{B}_i^c \cap \mathcal{A}_i \cap \mathcal{B}_{i-1}\big) \leq \exp\bigg(-\frac{\tau_i\gamma^2\varepsilon_{\tau_i}^2}{8}\bigg) \leq \bigg(\frac{1}{e} \wedge \frac{\tau_i}{T}\bigg)^{\gamma^2} \leq \frac{\tau_i}{T}, \tag{2.6}$$

where we used the fact that $\Delta_i \geq (3/2)\gamma\varepsilon_{\tau_i}$.

It remains to bound the first term in the right-hand side of (2.4). On the event $\mathcal{C}_{i-1}$, the optimal arm $*$ has not been eliminated before round τ_{i-1}, but every suboptimal arm $j \le i-1$ has. So the probability that there exists an arm $j \ge i$ that eliminates $*$ between τ_{i-1} and τ_i can be bounded as

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

where $\Phi_j(\tau) = \mathbb{P}(\exists s \le \tau; \hat{\Delta}_j(s) \le -\gamma\varepsilon_s)$. Using Lemma A.1, we get $\Phi_j(\tau) \le 4\tau/T$. This bound implies that

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

Using (2.2) and $\Delta_{i+1} \le \Delta_i$, the first sum can be bounded as

$$\begin{aligned}\sum_{j=1}^{K} \sum_{i=1}^{j\wedge K_0 - 1} \tau_i(\Delta_i - \Delta_{i+1}) &\le 19\gamma^2 \sum_{j=1}^{K} \sum_{i=1}^{j\wedge K_0 - 1} \overline{\log}\Big(\frac{T\Delta_i^2}{18\gamma^2}\Big)\frac{\Delta_i - \Delta_{i+1}}{\Delta_i^2}\\ &\le 19\gamma^2 \sum_{j=1}^{K} \int_{\Delta_{j\wedge K_0}}^{\Delta_1} \overline{\log}\Big(\frac{Tx^2}{18\gamma^2}\Big)\frac{\mathrm{d}x}{x^2}\\ &\le 19\gamma^2 \sum_{j=1}^{K} \frac{1}{\Delta_{j\wedge K_0}}\Big[\overline{\log}\Big(\frac{T\Delta_{j\wedge K_0}^2}{18\gamma^2}\Big) + 2\Big].\end{aligned}$$

The previous two displays together with (2.2) yield

$$\sum_{i=1}^{K_0} \Delta_i \mathbb{P}\big(\mathcal{A}_i^c \cap \mathcal{C}_{i-1}\big) \le \frac{304\gamma^2}{T} \sum_{j=1}^{K} \frac{1}{\Delta_{j\wedge K_0}} \overline{\log}\Big(\frac{T\Delta_{j\wedge K_0}^2}{18\gamma^2}\Big).$$

Putting together (2.3), (2.4), (2.6) and the above display yield that the expected regret $\mathbb{E}R_N(\hat{\pi})$of the SE policy is bounded above by

$$
\begin{aligned}
&323\gamma^2\Big(1+\frac{n}{T}\Big)\sum_{i=1}^{K_0}\frac{1}{\Delta_i}\overline{\log}\Big(\frac{n\Delta_i^2}{18\gamma^2}\Big)\\
&\quad+304\frac{\gamma^2 n}{T}\frac{K-K_0}{\Delta_{K_0}}\overline{\log}\Big(\frac{n\Delta_{K_0}^2}{18\gamma^2}\Big)+n\Delta_{K_0+1}.
\end{aligned}
\tag{2.7}
$$

Fix $\Delta\geq 0$ and let K_0 be such that $\Delta_{K_0+1}=\Delta^-$. An easy study of the variations of the function

$$
x\mapsto\phi(x)=\frac{1}{x}\overline{\log}\Big(\frac{nx^2}{18\gamma^2}\Big),\qquad x>0,
$$

reveals that $\phi(x)\leq(2e^{-1/2})\phi(x')$ for any $x\geq x'\geq 0$. Using this bound equation (2.7) with $x'=\Delta_i, i\leq K_0$ and $x=\Delta$ completes the proof. □

The following corollary is obtained from a slight variations on the proof of Theorem 2.1. It allows us to better compare our results to the extant literature.

COROLLARY 2.1. *Under the setup of Theorem 2.1, the* SE *policy* $\hat{\pi}$ *run with parameter* $T=n$ *and* $\gamma=1$ *satisfies for any* $K_0\leq K$,

$$
\mathbb{E}R_N(\hat{\pi})\leq 646\sum_{i=1}^{K_0}\frac{\overline{\log}(n\Delta_i^2)}{\Delta_i}+304\frac{K-K_0}{\Delta_{K_0}}\overline{\log}(n\Delta_{K_0}^2)+n\Delta_{K_0+1}.
\tag{2.8}
$$

In particular,

$$
\mathbb{E}R_N(\hat{\pi})\leq\min\Bigg\{646\sum_{i=1}^{K}\frac{\overline{\log}(n\Delta_i^2)}{\Delta_i},166\sqrt{nK\log(K)}\Bigg\}.
\tag{2.9}
$$

PROOF. Note that (2.8) follows from (2.7). To prove (2.9), take $K_0=K$ in (2.8) and $\Delta=28\sqrt{K\log(784K/18)/n}$ in Theorem 2.1, respectively. □

This corollary is actually closer to the result of [5]. The additional second term in our bound comes from the fact that we had to take into account the probability that an optimal arm $*$ can be eliminated by any arm, not just by some *suboptimal arm* with index lower than K_0; see [5], page 8. It is unclear why it is enough to look at the elimination by those arms, since if $*$ is eliminated—no matter the arm that eliminated it—the Hoeffding bound (2.5) no longer holds.

The right-hand side of (2.9) is the minimum of two terms. The first term is distribution-dependent and shows that the SE policy adapts to the unknown distribution of the rewards. It is very much in the spirit of the original bound of [14] and

of the more recent finite sample result of [4]. Our bound for the SE policy is smaller than the aforementioned bounds for the UCB policy by a logarithmic factor. Reference [14] did not provide the first bounds on the expected regret. Indeed, [22] and [6] had previously derived what is often called *gap-free* bound as they hold uniformly over the Δ_i's. The second term in our bound is such a gap-free bound. It is of secondary interest in this paper and arise as a byproduct of refined distribution dependent bound. Nevertheless, it allows us to recover near optimal bounds of the same order as [12]. They depart from optimal rates by a factor $\sqrt{\log K}$ as proved in [1]. Actually, the result of [1] is much stronger than our gap-free bound since it holds for any sequence of bounded rewards, not necessarily drawn independently.

None of the distribution-dependent bounds in Corollary 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 (e.g., if all 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.

3. Bandit with covariates. This section is dedicated to a detailed description of the 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

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

of independent random vectors, where $(X_t)_{t\geq 1}$, is a sequence of i.i.d. 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. Throughout the paper, we assume that P_X has a density, with respect to the Lebesgue measure, bounded above and below by some $\bar{c} > 0$ and $\underline{c} > 0$, respectively. We denote by E_X the expectation with respect to P_X. We assume that, for each $i \in \mathcal{I} = \{1, \dots, K\}$, rewards $Y_t^{(i)}$, $t = 1, \dots, n$, are random variables in $[0, 1]$ with conditional expectation given by

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

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 *policy* $\pi = \{\pi_t\}$ is a sequence of random functions $\pi_t : \mathcal{X} \to \{1, \dots, K\}$ 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* $\pi^\star$, 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 $\pi^\star$ prescribes to pull any arm with the largest expected reward, that is,

$$\pi^\star(X_t) \in \arg\max_{i=1,\dots,K} f^{(i)}(X_t)$$

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

$$f^\star(x) = \max\{f^{(i)}(x); i = 1, \dots, K\}.$$

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

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

Without further assumptions on the machine, the game can be arbitrarily difficult and, as a result, expected 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 expected 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 $f^{(i)}$ is (β, L)-Hölder, that is, if

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

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 $\min_i f^{(i)}(x) \ne \max_i f^{(i)}(x)$ it is defined by

$$f^\sharp(x) = \max_i\big\{f^{(i)}(x); f^{(i)}(x) < f^\star(x)\big\}$$

and by $f^\sharp(x) = f^\star(x) = f^{(1)}(x)$ otherwise. Notice that a direct consequence of the smoothness condition is that the function $f^\star$ is (β, L)-Hölder; however, $f^\sharp$ might not even be continuous.

The behavior of the function $\Delta := f^\star - f^\sharp$ critically controls the complexity of the problem and the Hölder regularity 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

[17, 21], which drives the terminology employed here. It was originally imported to the bandit setup by [9].

MARGIN CONDITION. We say that the machine satisfies the margin condition with parameter $\alpha > 0$ if there exists $\delta_0 \in (0, 1)$, $C_0 > 0$ such that

$$P_X\big[0 < f^\star(X) - f^\sharp(X) \le \delta\big] \le C_0\delta^\alpha \qquad \forall \delta \in [0, \delta_0].$$

If the marginal P_X has a density bounded above and below, the margin condition contains only information about the behavior of the function Δ and not the marginal P_X itself. This is in contrast with [9] 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 reflects in the bounds on the expected 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 hold*:

- *if* $\alpha\beta > d$, *then a given arm is either always or never optimal*; *in the latter case, it is bounded away from* $f^\star$ *and one can take* $\alpha = \infty$;
- *if* $\alpha\beta \le d$, *then there exist machines with nontrivial oracle policies*.

PROOF. This proposition is a straightforward consequences of, respectively, the first two points of Proposition 3.4 in [3].

For completeness, we provide an example with $d = 1$, $\mathcal{X} = [0, 1]$, $f^{(2)} = \cdots = f^{(K)} \equiv 0$ and $f^{(1)}(x) = L\,\mathrm{sign}(x - 0.5)|x - 0.5|^{1/\alpha}$. Notice that $f^{(1)}$ is (β, L)-Hölder if and only if $\alpha\beta \le 1$. Any oracle policy is nontrivial, and, for example, one can define $\pi^\star(x) = 2$ if $x \le 0.5$ and $\pi^\star(x) = 1$ if $x > 0.5$. Moreover, it can be easily shown that the machine satisfies the margin condition with parameter α and with $\delta_0 = C_0 = 1$. □

We denote by $\mathcal{M}^K_{\mathcal{X}}(\alpha, \beta, L)$ 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 $\underline{c} > 0$, respectively.

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 $B_j \cap B_k = \varnothing$, $j, k \geq 1$, up to sets of null P_X probability. For any $i \in \{\star, 1, \dots, K\}$ and any bin B, define

$$\bar{f}_B^{(i)} = \mathbb{E}\big[f^{(i)}(X_t) | X_t \in B\big] = \frac{1}{P_X(B)} \int_B f^{(i)}(x) \, \mathrm{d}P_X(x). \tag{3.1}$$

To define and analyze our policies, it is convenient to reindex the random vectors $(X_t, Y_t^{(1)}, \dots, Y_t^{(K)})_{t \geq 1}$ as follows. Given a bin B, let $t_B(s)$ denote the sth time at which the sequence $(X_t)_{t \geq 1}$ is in B and observe that it is a stopping time. It is a standard exercise to show that, for any bin B and any arm i, the random variables $Y_{t_B(s)}^{(i)}, s \geq 1$ are i.i.d. with expectation given by $\bar{f}_B^{(i)} \in [0, 1]$. As a result, the random variables $Y_{B,1}^{(i)}, Y_{B,2}^{(i)}, \dots$ obtained by successive pulls of arm i when $X_t \in B$ form an i.i.d. sequence in $[0, 1]$ with expectation given by $\bar{f}_B^{(i)} \in [0, 1]$. 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. This observation leads to the notion of *policy on* B. More precisely, fix a subset $B \subset \mathcal{X}$, an integer $t_0 \geq 1$ and recall that $\{t_B(s) : s \geq 1, t_B(s) \geq t_0\}$ is the set of chronological times t posterior to t_0 at which $X_t \in B$. Fix $\mathcal{I}' \subset \mathcal{I}$ and consider the static bandit problem with arms $\mathcal{I}'$ defined in Section 2 where successive pulls of arm $i \in \mathcal{I}'$, at times posterior to t_0, yield rewards $Y_{B,1}^{(i)}, Y_{B,2}^{(i)}, \dots$, that are i.i.d. in $[0, 1]$ with mean $\bar{f}_B^{(i)} \in [0, 1]$. The SE policy with parameters T, γ on this static problem is called SE *policy on* B *initialized at time* t_0 *with initial set of arms* $\mathcal{I}'$ *and parameters* T, γ.

4. Binned Successive Elimination. We first outline a naive policy to operate the bandit machine described in Section 3. It consists of fixing a partition of $\mathcal{X}$ and for each set B in this partition, to run the SE policy on B initialized at time $t_0 = 1$ with initial set of arms $\mathcal{I}$ and parameters T, γ to be defined below.

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 $\mathcal{X} = [0, 1]^d$: the collection of hypercubes defined for $\mathsf{k} = (k_1, \dots, k_d) \in \{1, \dots, M\}^d$ by

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

Policy 2 Binned Successive Elimination (BSE)

Input: Set of arms $\mathcal{I} = \{1, \dots, K\}$. Parameters n, M.
Output: $\bar{\pi}_1, \dots, \bar{\pi}_n \in \mathcal{I}$.

```
B ← B_M
for B ∈ B_M do
    Initialize a SE policy π̂_B with parameters T = nM^{-d}, γ = 1.
    N_B ← 0.
end for

for t = 1, ..., n do
    B ← B(X_t).
    N_B ← N_B + 1.
    π̄_t ← π̂_{B,N_B} (observe Y_t^{(π̄_t)}).
end for
```

In this paper, all sets are defined up to sets of null Lebesgue measure. As mentioned in Section 3.3, the problem can be decomposed into M^d independent static bandit problems, one for each $B \in \mathcal{B}_M$.

Denote by $\hat{\pi}_B$ the SE policy on bin B with initial set of arms $\mathcal{I}$ and parameters $T = nM^{-d}, \gamma = 1$. For any $x \in \mathcal{X}$, let $\mathcal{B}(x) \in \mathcal{B}_M$ denote the bin such that $x \in \mathcal{B}(x)$. Moreover, for any time $t \geq 1$, define

$$N_B(t) = \sum_{l=1}^{t} \mathbb{1}(X_l \in B) \tag{4.1}$$

to be the number of times before t when the covariate fell in bin B. The BSE policy $\bar{\pi}$ is a sequence of functions $\bar{\pi}_t : \mathcal{X} \to \mathcal{I}$ defined by $\bar{\pi}_t(x) = \hat{\pi}_{B,N_B(t)}$, where $B = \mathcal{B}(x)$. It can be implemented according to the pseudo-code of Policy 2.

The following theorem gives an upper bound on the expected 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 $\mathcal{M}_{\mathcal{X}}^{K}(\alpha, \beta, L)$. Then the* BSE *policy $\bar{\pi}$ with $M = \lfloor (\frac{n}{K \log(K)})^{1/(2\beta+d)} \rfloor$ has an expected regret at time n bounded as follows*:

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

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

The case $K = 2$ was studied in [18] using a similar policy called UCBogram. Unlike in [18] where suboptimal bounds for the UCB policy are used, we use here the sharper regret bounds of Theorem 2.1 and the SE policy as a running horse for

our policy, thus leading to a better bound than [18]. Optimality for the two-armed case is discussed after Theorem 5.1.

PROOF OF THEOREM 4.1. 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 $P_X(f \in A) = P_X(f(X) \in A)$. Moreover, for any $i \in \{\star, 1, \dots, K\}$, we use the notation $\bar{f}_j^{(i)} = \bar{f}_{B_j}^{(i)}$.

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 $c_1 M^{-\beta} < \delta_0$.

Recall that the BSE policy $\bar{\pi}$ is a collection of functions $\bar{\pi}_t(x) = \hat{\pi}_{\mathcal{B}(x), N_{\mathcal{B}(x)}(t)}$ that are constant on each B_j. Therefore, the regret of $\bar{\pi}$ can be decomposed as $R_n(\bar{\pi}) = \sum_{j=1}^{M^d} \mathsf{R}_j(\bar{\pi})$, where

$$\mathsf{R}_j(\bar{\pi}) = \sum_{t=1}^{n} \bigl(f^\star(X_t) - f^{(\hat{\pi}_{B, N_B(t)})}(X_t)\bigr) \mathbb{1}(X_t \in B_j).$$

Conditioning on the event $\{X_t \in B_j\}$, it follows from (3.1) that

$$\mathbb{E}\mathsf{R}_j(\bar{\pi}) = \mathbb{E}\left[\sum_{t=1}^{n} \bigl(\bar{f}_j^\star - \bar{f}_j^{(\bar{\pi}_t)}\bigr) \mathbb{1}(X_t \in B_j)\right] = \mathbb{E}\left[\sum_{s=1}^{N_j(n)} \bigl(\bar{f}_j^\star - \bar{f}_j^{(\hat{\pi}_{B_j, s})}\bigr)\right],$$

where $N_j(n) = N_{B_j}(t)$ is defined in (4.1); it satisfies, by assumption, $\underline{c} n M^{-d} \leq \mathbb{E}[N_j(n)] \leq \bar{c} n M^{-d}$.

Let $\tilde{\mathsf{R}}_j(\bar{\pi}) = \sum_{s=1}^{N_j(n)} f_j^* - \bar{f}_j^{(\hat{\pi}_{B_j, s})}$ be the regret associated to a static bandit problem with arm i yielding reward $\bar{f}_j^{(i)}$ and where $f_j^* = \max_i \bar{f}_j^{(i)} \leq \bar{f}_j^\star$ is the largest average reward. It follows from the smoothness condition that $\bar{f}_j^\star \leq f_j^* + c_1 M^{-\beta}$ so that

$$\mathbb{E}\mathsf{R}_j(\bar{\pi}) \leq \mathbb{E}\tilde{\mathsf{R}}_j(\bar{\pi}) + \bar{c} n M^{-d}\bigl(\bar{f}_j^\star - f_j^*\bigr) \leq \mathbb{E}\tilde{\mathsf{R}}_j(\bar{\pi}) + c_1 \bar{c} n M^{-\beta - d}. \tag{4.2}$$

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

$$\mathcal{J} := \bigl\{j \in \{1, \dots, M^d\} \text{ s.t. } \exists x \in B_j,\ f^\star(x) - f^\sharp(x) > c_1 M^{-\beta}\bigr\}.$$

A bin B that is not well behaved is called *strongly ill behaved* if there is some $x \in B$ such that $f^\star(x) = f^\sharp(x) = f^{(i)}(x)$, for all $i \in \mathcal{I}$, and *weakly ill behaved* otherwise. Strongly and weakly ill behaved bins have indices in

$$\mathcal{J}_s^c := \bigl\{j \in \{1, \dots, M^d\} \text{ s.t. } \exists x \in B_j,\ f^\star(x) = f^\sharp(x)\bigr\}$$

and

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

respectively. Note that for any $i \in \mathcal{I}$, the function $f^\star - f^{(i)}$ is $(\beta, 2L)$-Hölder. Thus for any $j \in \mathcal{J}_s^c$ and any $i = 1, \dots, K$, we have $f^\star(x) - f^{(i)}(x) \le c_1 M^{-\beta}$ for all $x \in B_j$ so that the inclusion $\mathcal{J}_s^c \subset \{1, \dots, M^d\} \setminus \mathcal{J}$ 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 $x \in B_j$, $f^\star(x) - f^{(i)}(x) \le c_1 M^{-\beta}$. Therefore,

$$\begin{aligned}\sum_{j \in \mathcal{J}_s^c} \mathbb{E}\mathsf{R}_j(\bar{\pi}) &\le c_1 n M^{-\beta} P_X\{0 < f^\star(X) - f^\sharp(X) \le c_1 M^{-\beta}\} \\ &\le c_1^{1+\alpha} n M^{-\beta(1+\alpha)},\end{aligned} \tag{4.3}$$

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

Second part: *Weakly ill behaved bins in* $\mathcal{J}_w^c$. The numbers of weakly ill behaved bins can be bounded using $f^\star(x) - f^\sharp(x) > 0$ on such a bin; indeed, the margin condition implies that

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

It yields $|\mathcal{J}_w^c| \le \frac{c_1^\alpha}{\underline{c}} M^{d-\beta\alpha}$. Moreover, we bound the expected regret on weakly ill behaved bins using Theorem 2.1 with specific values

$$\Delta^- < \Delta := \sqrt{K \log(K) M^d / n}, \qquad \gamma = 1 \quad \text{and} \quad T = n M^{-d}.$$

Together with (4.2), it yields

$$\sum_{j \in \mathcal{J}_w^c} \mathbb{E}\mathsf{R}_j(\bar{\pi}) \le c_2\big[\sqrt{K \log(K)} M^{d/2-\beta\alpha} \sqrt{n} + n M^{-\beta(1+\alpha)}\big]. \tag{4.4}$$

Third part: *Well-behaved bins in* $\mathcal{J}$. This part is decomposed into two steps. In the first step, we bound the expected 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 these expected regrets.

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

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

$$f^\star(x) - f^{(i)}(x) > c_1 M^{-\beta} - 2L\|x - x_j\|^\beta \ge M^{-\beta}. \tag{4.5}$$

Therefore, $f^\star - f^\sharp > 0$ on B_j. Moreover, for any arm $i \in \mathcal{I}_j^\star$ that is not the best arm at some $x \neq x_j$, then necessarily $0 < f^\star(x) - f^\sharp(x) \leq f^\star(x) - f^{(i)}(x) \leq c_1 M^{-\beta}$. So for any $x \in B_j$ and any $i \in \mathcal{I}_j^\star$, it holds that either $f^\star(x) = f^{(i)}(x)$ or $f^\star(x) - f^{(i)}(x) \leq c_1 M^{-\beta}$. It yields

$$f^\star(x) - f^{(i)}(x) \leq c_1 M^{-\beta} \mathbb{1}\{0 < f^\star(x) - f^\sharp(x) \leq c_1 M^{-\beta}\}. \tag{4.6}$$

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

$$q_j := P_X\{0 < f^\star - f^\sharp \leq c_1 M^{-\beta} | X \in B_j\}.$$

Together with (4.2), it yields $\mathbb{E}\tilde{\mathsf{R}}_j(\bar{\pi}) \leq \mathbb{E}\mathsf{R}_j(\bar{\pi}) + \bar{c} c_1 n M^{-d-\beta} q_j$. For any suboptimal arms $i \in \mathcal{I}_j^0$, (4.5) implies that $\underline{\Delta}_j^{(i)} := \bar{f}_j^\star - \bar{f}_j^{(i)} > M^{-\beta}$.

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

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

and observe that if $i \in \mathcal{J}$ is such that $\underline{\Delta}_j^{(i)} < \underline{\Delta}_j$, then $i \in \mathcal{I}_j^\star$. Therefore, it follows from (4.6) that $\underline{\Delta}_j^{(i)} \leq c_1 M^{-\beta} q_j$ for such i. Applying Theorem 2.1 with $\underline{\Delta}_j$ as above and $\gamma = 1$, we find that there exists a constant $c_3 > 0$ such that, for any $j \in \mathcal{J}$,

$$\mathbb{E}\tilde{\mathsf{R}}_j(\bar{\pi}) \leq 392(1+\bar{c}) \frac{K}{\underline{\Delta}_j} \overline{\log}(n M^{-d} \underline{\Delta}_j^2) + \bar{c} c_1 n M^{-d-\beta} q_j.$$

Hence,

$$\mathbb{E}\mathsf{R}_j(\bar{\pi}) \leq c_3 \left(\frac{K}{\underline{\Delta}_j} \overline{\log}(n M^{-d} \underline{\Delta}_j^2) + n M^{-d-\beta} q_j \right). \tag{4.7}$$

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 $\mathcal{J} = \{1, \ldots, j_1\}$ and that the gaps are ordered $0 < \underline{\Delta}_1 \leq \underline{\Delta}_2 \leq \cdots \leq \underline{\Delta}_{j_1}$. 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 $\underline{\Delta}_j = \bar{f}_j^\star - \bar{f}_j^{(i)}$. But since the function $f^\star - f^{(i)}$ satisfies the smoothness condition with parameters $(\beta, 2L)$, we find that if $\underline{\Delta}_j \leq \delta$ for some $\delta > 0$, then

$$0 < f^\star(x) - f^{(i)}(x) \leq \delta + 2L d^{\beta/2} M^{-\beta} \qquad \forall x \in B_j.$$

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

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

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

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

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

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

Therefore, using the fact that $\underline{\Delta}_j \geq \delta_0/c_1$ for $j \geq j_2$, we get from (4.7) that

$$\begin{aligned}\sum_{j\in\mathcal{J}} \mathbb{E}\mathsf{R}_j(\bar\pi) &\\ \leq c_5\bigg[\sum_{j=1}^{j_2} K\frac{\overline{\log}(n\gamma_j^2/M^d)}{\gamma_j} &+ \sum_{j=j_2+1}^{j_1} K\log(n) + \sum_{j\in\mathcal{J}} nM^{-d-\beta}q_j\bigg].\end{aligned} \tag{4.8}$$

Fourth part: *Putting things together.* Combining (4.3), (4.4) and (4.8), we obtain the following bound:

$$\begin{aligned}\mathbb{E}R_n(\bar\pi) \leq c_6\bigg[&nM^{-\beta(1+\alpha)}\\ &+ \sqrt{K\log(K)}M^{d/2-\alpha\beta}\sqrt{n} + K\sum_{j=1}^{j_2}\frac{\overline{\log}(n\gamma_j^2/M^d)}{\gamma_j}\\ &+ KM^d\log n + nM^{-d-\beta}\sum_{j\in\mathcal{J}} q_j\bigg].\end{aligned} \tag{4.9}$$

We now bound from above the first sum in (4.9) by decomposing it into two terms. From the definition of γ_j, there exists an integer j_3 satisfying $c_7M^{d-\alpha\beta} \leq j_3 \leq 2c_7M^{d-\alpha\beta}$ and such that $\gamma_j = M^{-\beta}$ for $j \leq j_3$ and $\gamma_j = c_4(jM^{-d})^{1/\alpha}$ for $j > j_3$. It holds

$$\sum_{j=1}^{j_3}\frac{\overline{\log}(n\gamma_j^2/M^d)}{\gamma_j} \leq c_8M^{d+\beta(1-\alpha)}\overline{\log}\bigg(\frac{n}{M^{2\beta+d}}\bigg) \tag{4.10}$$

and

$$\begin{aligned}\sum_{j=j_3+1}^{j_2}\frac{\overline{\log}(n\gamma_j^2/M^d)}{\gamma_j} &\leq c_9\sum_{j=j_3+1}^{M^d}\bigg(\frac{j}{M^d}\bigg)^{-1/\alpha}\overline{\log}\bigg(\frac{n}{M^d}\bigg[\frac{j}{M^d}\bigg]^{2/\alpha}\bigg)\\ &\leq c_{10}M^d\int_{M^{-\alpha\beta}}^1 \overline{\log}\bigg(\frac{n}{M^d}x^{2/\alpha}\bigg)x^{-1/\alpha}\,\mathrm{d}x.\end{aligned} \tag{4.11}$$

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

The second sum in (4.9) can be bounded as

$$\begin{aligned}\sum_{j\in\mathcal{J}} q_j &= \sum_{j\in\mathcal{J}} \mathbb{P}\{0 < f^\star(X) - f^\sharp(X) \le c_1 M^{-\beta} | X \in B_j\} \\ &\le \frac{M^d}{\underline{c}} \mathbb{P}\{0 < f^\star(X) - f^\sharp(X) \le c_1 M^{-\beta}\} \le \frac{c_1^\alpha}{\underline{c}} M^{d-\beta\alpha}.\end{aligned} \tag{4.12}$$

Putting together (4.9)–(4.12), we obtain

$$\begin{aligned}\mathbb{E}R_n(\bar{\pi}) \le c_{11}\Bigg[& nM^{-\beta(1+\alpha)} + \sqrt{K\log(K)}M^{d/2-\alpha\beta}\sqrt{n} + KM^{d+\beta(1-\alpha)} \\ & + KM^{d+\beta(1-\alpha)}\overline{\log}\left(\frac{n}{M^{2\beta+d}}\right) + KM^d \log n\Bigg],\end{aligned}$$

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* (see, e.g., page 17 in [7]) which consists of "reseting" the game at times $2^k, k = 1, 2, \ldots$.

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 $\alpha \ge 1$, it is possible to minimize the expression in (4.9) 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: for n large enough, it requires the operator to pull each arm at least once in each bin and therefore to incur an expected regret of at least order M^d. In other words, the BSE splits the space $\mathcal{X}$ in "too many" bins when $\alpha \ge 1$. Intuitively this can be understood as follows. When $\alpha \ge 1$, the gap function $f^\star(x) - f^\sharp(x)$ is bounded away from zero on a large subset of $\mathcal{X}$. Hence there is no need to carefully estimate it since the optimal arm is the same across the region. 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.

5. Adaptively Binned Successive Elimination. We need the following definitions. Assume that $n \ge K\log(K)$ and let k_0 be the smallest integer such that

$$2^{-k_0} \le \left(\frac{K\log(K)}{n}\right)^{1/(d+2\beta)}. \tag{5.1}$$

For any bin $B \in \bigcup_{k=0}^{k_0} \mathcal{B}_{2^k}$, let ℓ_B be the smallest integer such that

$$U\big(\ell_B, n|B|^d\big) \le 2c_0|B|^\beta, \tag{5.2}$$

where U is defined in (2.1) and $c_0 = 2Ld^{\beta/2}$. This definition implies that

$$\ell_B \le C_\ell |B|^{-2\beta} \log\bigl(n|B|^{(2\beta+d)}\bigr) \tag{5.3}$$

for $C_\ell > 0$, because $x \mapsto U(x, n|B|^d)$ is decreasing for $x > 0$.

The ABSE policy operates akin to the BSE except that instead of fixing a partition $\mathcal{B}_M$, it relies on an adaptive partition that is refined over time. This partition is better understood using the notion of rooted tree.

Let $\mathcal{T}^*$ be a tree with root $\mathcal{X}$ and maximum depth k_0. A node B of $\mathcal{T}^*$ with depth $k = 0, \dots, k_0 - 1$ is a set from the regular partition $\mathcal{B}_{2^k}$. The children of node $B \in \mathcal{B}_{2^k}$ are given by $\mathsf{burst}(B)$, defined to be the collection of 2^d bins in $\mathcal{B}_{2^{k+1}}$ that forms a partition of B.

Note that the set $\mathcal{L}$ of leaves of each subtree $\mathcal{T}$ of $\mathcal{T}^*$ forms a partition of $\mathcal{X}$. The ABSE policy constructs a sequence of partitions $\mathcal{L}_1, \dots, \mathcal{L}_n$ that are leaves of subtrees of $\mathcal{T}^*$. At a given time $t = 1, \dots, n$, we refer to the elements of the current partition $\mathcal{L}_t$ as *live* bins. The sequence of partitions is nested in the sense that if $B \in \mathcal{L}_t$, then either $B \in \mathcal{L}_{t+1}$ or $\mathsf{burst}(B) \subset \mathcal{L}_{t+1}$. The sequence $\mathcal{L}_1, \dots, \mathcal{L}_n$ is constructed as follows.

In the initialization step, set $\mathcal{L}_0 = \varnothing$, $\mathcal{L}_1 = \mathcal{X}$, and the initial set of arms $\mathcal{I}_\mathcal{X} = \{1, \dots, K\}$. Let $t \le n$ be a time such that $\mathcal{L}_t \ne \mathcal{L}_{t-1}$, and let B_t be the collection of sets B such that $B \in \mathcal{L}_t \setminus \mathcal{L}_{t-1}$. We say that the bins $B \in \mathsf{B}_t$ are *born* at time t. For each set $B \in \mathsf{B}_t$, assume that we are given a set of active arms $\mathcal{I}_B$. Note that $t = 1$ is such a time with $\mathsf{B}_1 = \{\mathcal{X}\}$ and active arms $\mathcal{I}_\mathcal{X}$. For each born bin $B \in \mathsf{B}_t$, we run a SE policy $\hat{\pi}_B$ initialized at time t with initial set of arms $\mathcal{I}_B$ and parameters $T_B = n|B|^{-d}$, $\gamma = 2$. Such a policy is defined in Section 3.3. Let $t(B)$ denote the time at which $\hat{\pi}_B$ has reached ℓ_B rounds and let

$$\tilde{N}_B(t) = \sum_{l=1}^{t} \mathbb{1}(X_t \in B, B \in \mathcal{L}_t) \tag{5.4}$$

denote the number of times covariate X_t fell in bin B while B was a live B. At time $t(B) + 1$, we replace the node B by its children $\mathsf{burst}(B)$ in the current partition. Namely, $\mathcal{L}_{t(B)+1} = (\mathcal{L}_{t(B)} \setminus B) \cup \mathsf{burst}(B)$. Moreover, to each bin $B' \in \mathsf{burst}(B)$, we assign the set $\mathcal{I}_{B'} = \hat{\mathbb{I}}_{B, \tilde{N}_B(t(B))}$ of arms that were left active by policy $\hat{\pi}_B$ on its parent B at the end of the ℓ_B rounds. This procedure is repeated until the horizon n is reached.

The intuition behind this policy is the following. The parameters of the SE policy $\hat{\pi}_B$ run at the birth of bin B are chosen exactly such that arms i with average gap $|\bar{f}^\star_B - \bar{f}^{(i)}_B| \ge C|B|^\beta$ are eliminated by the end of ℓ_B rounds with high probability. The smoothness condition ensures that these eliminated arms satisfy $f^\star(x) > f^{(i)}(x)$ for all $x \in B$ so that such arms are uniformly suboptimal on bin B. Among the kept arms, none is uniformly better than another, so bin B is burst and the process is repeated on the children of B where other arms may be uniformly suboptimal. The formal definition of the ABSE is given in Policy 3; it satisfies the following theorem.

Policy 3 Adaptively Binned Successive Elimination (ABSE)

Input: Set of arms $\mathcal{I}_{\mathcal{X}} = \{1, \ldots, K\}$. Parameters n, $c_0 = 2Ld^{\beta/2}$, k_0.
Output: $\tilde{\pi}_1, \ldots, \tilde{\pi}_n \in \mathcal{I}$.

```
t ← 0, k ← 0, L ← {X}.
Initialize a SE policy π̂_X with parameters T = n, γ = 2 and arms I = I_X.
N_X ← 0.
for t = 1, ..., n do
  B ← L(X_t).
  N_B ← N_B + 1.                                   /count times X_t ∈ B/
  π̃_t ← π̂_{B,N_B} (observe Y_t^{(π̃_t)}).          /choose arm from SE policy π̂_B/
  τ_B ← τ̂_{B,N_B}                                  /update number of rounds for π̂_B/
  I_B ← Î_{B,N_B}                                  /update active arms for π̂_B/
  if τ_B ≥ ℓ_B and |B| ≥ 2^{-k_0+1} and |I_B| ≥ 2   /conditions to burst(B)/
  then
    for B' ∈ burst(B) do
      I_{B'} ← I_B                                 /assign remaining arms as initial arms/
      Initialize SE policy π̂_{B'} with T = n|B'|^d, γ = 2 and arms I = I_{B'}.
      N_{B'} ← 0.                                  /set time to 0 for new SE policy/
    end for
    L ← L \ B                                      /remove B from current partition/
    L ← L ∪ burst(B)                               /add B's children to current partition/
  end if
end for
```

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 an expected regret at time n bounded by*

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

where $C > 0$ 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 [18] imply that the bound on expected regret cannot be improved as a function of n except for a constant multiplicative term. 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 OF THEOREM 5.1. We first consider the case where $\alpha < \infty$, which implies that $\alpha\beta \le d$; see Proposition 3.1.

We keep track of positive constants by numbering them $c_1, c_2, \ldots$, yet they might differ from previous sections. On each newly created bin B, a new SE policy is initialized, and we denote by $Y_{B,1}^{(i)}, Y_{B,2}^{(i)}, \ldots$, 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).

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

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

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

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

Denote by $r_n^{\mathrm{live}}(B)$ the regret incurred by the ABSE policy $\tilde{\pi}$ when covariate X_t fell in a *live* bin $B \in \mathcal{L}_t$, where we recall that $\mathcal{L}_t$ denotes the current partition at time t. It is defined by

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

We also define $\mathcal{B}_t := \bigcup_{s\le t}\mathcal{L}_s$ to be the set of bins that were born at some time $s \le t$. We denote by $r_n^{\mathrm{born}}(B)$ the regret incurred when covariate X_t fell in such a bin. It is defined by

$$r_n^{\mathrm{born}}(B) = \sum_{t=1}^{n}\big[f^\star(X_t) - f^{(\tilde{\pi}_t(X_t))}(X_t)\big]\mathbb{1}(X_t \in B)\mathbb{1}(B \in \mathcal{B}_t).$$

Observe that if we define $\tilde{r}_n := r_n^{\mathrm{born}}(\mathcal{X})$, we have $\mathbb{E}R_n(\tilde{\pi}) = \mathbb{E}\tilde{r}_n$ since $\mathcal{X} \in \mathcal{B}_t$ and $X_t \in \mathcal{X}$ for all t. Note that for any $B \in \mathcal{T}^*$,

$$r_n^{\mathrm{born}}(B) = r_n^{\mathrm{live}}(B) + \sum_{B' \in \mathsf{burst}(B)} r_n^{\mathrm{born}}(B'). \tag{5.5}$$

Denote by $\mathcal{I}_B = \hat{\mathbb{I}}_{B,t_B}$ 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:

$$\underline{\mathcal{I}}_B := \Big\{ i \in \{1, \dots, K\} : \sup_{x \in B} f^\star(x) - f^{(i)}(x) \le c_0 |B|^\beta \Big\},$$

$$\bar{\mathcal{I}}_B := \Big\{ i \in \{1, \dots, K\} : \sup_{x \in B} f^\star(x) - f^{(i)}(x) \le 8 c_0 |B|^\beta \Big\}.$$

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 (5.5) implies that

$$r_n^{\mathrm{born}}(B) = r_n^{\mathrm{born}}(B) \mathbb{1}\big(\mathcal{A}_B^c\big) + r_n^{\mathrm{live}}(B) \mathbb{1}(\mathcal{A}_B) + \sum_{B' \in \mathsf{burst}(B)} r_n^{\mathrm{born}}\big(B'\big) \mathbb{1}(\mathcal{A}_B).$$

Let $\mathcal{L}^*$ denote the set of leaves of $\mathcal{T}^*$, that is the set of bins B such that $|B| = 2^{-k_0}$. In what follows, we adapt the convention that $\prod_{B' \in \mathcal{P}(\mathcal{X})} \mathbb{1}(\mathcal{A}_{B'}) = 1$.

We are going to treat regret incurred on live nonterminal nodes and live leaves separately and differently. As a result, the quantity we are interested in is decomposed as $\tilde{r}_n = \tilde{r}_n(\mathcal{T}^* \setminus \mathcal{L}^*) + \tilde{r}_n(\mathcal{L}^*)$ where

$$\tilde{r}_n\big(\mathcal{T}^* \setminus \mathcal{L}^*\big) := \sum_{B \in \mathcal{T}^* \setminus \mathcal{L}^*} \big(r_n^{\mathrm{born}}(B) \mathbb{1}\big(\mathcal{A}_B^c\big) + r_n^{\mathrm{live}}(B) \mathbb{1}(\mathcal{A}_B) \big) \prod_{B' \in \mathcal{P}(B)} \mathbb{1}(\mathcal{A}_{B'})$$

is the regret accumulated on live nonterminal nodes, and

$$\tilde{r}_n\big(\mathcal{L}^*\big) := \sum_{B \in \mathcal{L}^*} r_n^{\mathrm{born}}(B) \prod_{B' \in \mathcal{P}(B)} \mathbb{1}(\mathcal{A}_{B'}) = \sum_{B \in \mathcal{L}^*} r_n^{\mathrm{live}}(B) \prod_{B' \in \mathcal{P}(B)} \mathbb{1}(\mathcal{A}_{B'})$$

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

First part: *Control of the regret on the nonterminal nodes.* Fix $B \in \mathcal{T}^* \setminus \mathcal{L}^*$. On $\mathcal{G}_B$, we have $\mathcal{I}_{\mathsf{p}(B)} \subseteq \bar{\mathcal{I}}_{\mathsf{p}(B)}$ so that any active arm $i \in \mathcal{I}_{\mathsf{p}(B)}$ satisfies $\sup_{x \in \mathsf{p}(B)} |f^\star(x) - f^{(i)}(x)| \le 8 c_0 |\mathsf{p}(B)|^\beta$. Moreover, regret is only incurred at points where $f^* - f^\sharp > 0$, so defining $c_1 := 2^{3+\beta} c_0$ and conditioning on events $\{X_t \in B\}$ yields

$$\mathbb{E}\big[r_n^{\mathrm{live}}(B) \mathbb{1}(\mathcal{G}_B \cap \mathcal{A}_B) \big] \le \mathbb{E}\big[\tilde{N}_B(n)\big] c_1 |B|^\beta q_B \le c_1 K \ell_B |B|^\beta q_B,$$

where $q_B = P_X(0 < f^\star - f^\sharp \le c_1 |B|^\beta | X \in B)$ and $\tilde{N}_B(n)$ is defined in (5.4).

We can always assume that n is greater than $n_0 \in \mathbb{N}$, defined by

$$n_0 = \left\lceil K \log(K) \left(\frac{c_1}{\delta_0} \right)^{(d+2\beta)/\beta} \right\rceil \qquad \text{so that } c_1 2^{-k_0 \beta} \le \delta_0,$$

and let $k_1 \le k_0$ be the smallest integer such that $c_1 2^{-k_1 \beta} \le \delta_0$. Indeed, if $n \le n_0$, the result is true with a constant large enough.

Applying the same argument as in (4.12) yields the existence of $c_2 > 0$ such that, for any $k \in \{0, \dots, k_0\}$,

$$\sum_{|B|=2^{-k}} q_B \le c_2 2^{k(d-\beta\alpha)}.$$

Indeed, for $k \ge k_1$ one can define $c_2 = c_1^{\alpha}/\underline{c}$, and the same equation holds with $c_2 = 2^{dk_1}$ if $k \le k_1$. Summing over all depths $k \le k_0 - 1$, we obtain

$$\begin{aligned}\mathbb{E}\Bigg[\sum_{B\in\mathcal{T}^*\setminus\mathcal{L}^*} r_n^{\mathrm{live}}(B)\mathbb{1}(\mathcal{G}_B \cap \mathcal{A}_B)\Bigg] \\ \le c_1 c_2 C_\ell K \sum_{k=0}^{k_0-1} 2^{k(d+\beta-\alpha\beta)} \log\bigl(n2^{-k(2\beta+d)}\bigr).\end{aligned} \tag{5.6}$$

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

$$\mathbb{E}\bigl[r_n^{\mathrm{born}}(B)\mathbb{1}\bigl(\mathcal{G}_B \cap \mathcal{A}_B^c\bigr)\bigr] \le c_1 n|B|^{\beta} q_B P_X(B)\mathbb{P}\bigl(\mathcal{G}_B \cap \mathcal{A}_B^c\bigr). \tag{5.7}$$

It remains to control the probability of $\mathcal{G}_B \cap \mathcal{A}_B^c$; we define $\mathbb{P}^{\mathcal{G}_B}(\cdot) := \mathbb{P}(\cdot \cap \mathcal{G}_B)$. 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 $\sup_{x\in B} f^\star(x) - f^{(i)}(x) < c_0|B|^{\beta}$; this event is denoted by $\mathcal{D}_B^1$.

(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 $\sup_{x\in B} f^\star(x) - f^{(i)}(x) \ge 8c_0|B|^{\beta}$; this event is denoted by $\mathcal{D}_B^2$.

We use the following decomposition:

$$\mathbb{P}^{\mathcal{G}_B}\bigl(\mathcal{A}_B^c\bigr) = \mathbb{P}^{\mathcal{G}_B}\bigl(\mathcal{D}_B^1\bigr) + \mathbb{P}^{\mathcal{G}_B}\bigl(\mathcal{D}_B^2 \cap \bigl(\mathcal{D}_B^1\bigr)^c\bigr). \tag{5.8}$$

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

$$\bar{f}_B^{(i')} - \bar{f}_B^{(i)} \le \bar{f}_B^{\star} - \bar{f}_B^{(i)} < c_0|B|^{\beta} \le \frac{\varepsilon_{B,\ell_B}}{2}.$$

Therefore, if an arm $i \in \underline{\mathcal{I}}_B$ is eliminated, that is, if there exists $i' \in \mathcal{I}_{\mathsf{p}(B)}$ such that $\bar{Y}_{B,s}^{(i')} - \bar{Y}_{B,s}^{(i)} > \varepsilon_{B,s}$ for some $s \le \ell_B$, then either $\bar{f}_B^{(i)}$ or $\bar{f}_B^{(i')}$ does not belong to its respective confidence interval $[\bar{Y}_{B,s}^{(i)} \pm \varepsilon_{B,s}/4]$ or $[\bar{Y}_{B,s}^{(i')} \pm \varepsilon_{B,s}/4]$ for some $s \le \ell_B$. Therefore, since $-\bar{f}_B^{(i)} \le Y_s - \bar{f}_B^{(i)} \le 1 - \bar{f}_B^{(i)}$,

$$\mathbb{P}^{\mathcal{G}_B}\bigl(\mathcal{D}_B^1\bigr) \le \mathbb{P}\Biggl\{\exists s \le \ell_B; \exists i \in \mathcal{I}_{\mathsf{p}(B)}; \bigl|\bar{Y}_s^{(i)} - \bar{f}_B^{(i)}\bigr| \ge \frac{\varepsilon_{B,s}}{4}\Biggr\} \le 2K\frac{\ell_B}{n|B|^d}, \tag{5.9}$$

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 $f^\star(x^{(i)}) - f^{(i)}(x^{(i)}) > 8c_0|B|^\beta$. Let $\check{\imath} = \check{\imath}(i) \in \mathcal{I}$ be any arm such that $f^\star(x^{(i)}) = f^{(\check{\imath})}(x^{(i)})$; the smoothness condition implies that

$$\begin{aligned} \bar{f}_B^{(\check{\imath})} &\geq f^{(\check{\imath})}\big(x^{(i)}\big) - c_0|B|^\beta > f^{(i)}\big(x^{(i)}\big) + 7c_0|B|^\beta \\ &\geq \bar{f}_B^{(i)} + 6c_0|B|^\beta \geq \bar{f}_B^{(i)} + \tfrac{3}{2}\varepsilon_{B,\ell_B}. \end{aligned} \tag{5.10}$$

On the event $(\mathcal{D}_B^1)^c$, no arm in $\mathcal{I}_B$, and in particular any of the arms $\check{\imath}(i), i \in \mathcal{I}_{\mathsf{p}(B)} \setminus \bar{\mathcal{I}}_B$, has been eliminated until round ℓ_B. Therefore, the event $\mathcal{D}_B^2 \cap (\mathcal{D}_B^1)^c$ occurs if there exists $i \notin \bar{\mathcal{I}}_B$ such that $\bar{Y}_{B,\ell_B}^{(\check{\imath})} - \bar{Y}_{B,\ell_B}^{(i)} \leq \varepsilon_{B,\ell_B}$. In view of (5.10) and (5.2), it implies that there exists $i \in \mathcal{I}_{\mathsf{p}(B)}$ such that

$$|\bar{Y}_{B,\ell_B}^{(i)} - \bar{f}_B^{(i)}| \geq \frac{\varepsilon_{B,\ell_B}}{4}.$$

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

$$\begin{aligned} \mathbb{P}^{\mathcal{G}_B}\big(\mathcal{D}_B^2 \cap \big(\mathcal{D}_B^1\big)^c\big) &\leq \mathbb{P}\left\{\exists i \in \mathcal{I}_{\mathsf{p}(B)} : |\bar{Y}_{B,\ell_B}^{(i)} - \bar{f}_B^{(i)}| \geq \frac{\varepsilon_{B,\ell_B}}{4}\right\} \\ &\leq 2K\frac{\ell_B}{n|B|^d}, \end{aligned} \tag{5.11}$$

where the second inequality follows from (A.1).

Putting together (5.8), (5.9), (5.11) and (5.3), we get

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

Together with (5.7), it yields for $B \in \mathcal{T}^* \setminus \mathcal{L}^*$ that

$$\mathbb{E}\big[r_n^{\mathrm{born}}(B)\mathbb{1}\big(\mathcal{G}_B \cap \mathcal{A}_B^c\big)\big] \leq c_3K|B|^{-(\beta+d)}\log\big(n|B|^{(2\beta+d)}\big)q_B P_X(B).$$

If k is such that $c_1 2^{-k\beta} > \delta_0$, then any bin B such that $|B| = 2^{-k}$ satisfies $\mathbb{E}[r_n^{\mathrm{born}}(B)\mathbb{1}(\mathcal{G}_B \cap \mathcal{A}_B^c)] \leq c_4K\log n$. If k is such that $c_1 2^{-k\beta} \leq \delta_0$, then the above display together with the margin condition yield

$$\mathbb{E}\Bigg[\sum_{|B|=2^{-k}} r_n^{\mathrm{born}}(B)\mathbb{1}\big(\mathcal{G}_B \cap \mathcal{A}_B^c\big)\Bigg] \leq c_5K2^{k(\beta+d-\alpha\beta)}\log\big(n2^{-k(2\beta+d)}\big).$$

Summing over all depths $k = 0, \ldots, k_0 - 1$ and using (5.6), we obtain

$$\mathbb{E}\big[\tilde{r}_n\big(\mathcal{T}^* \setminus \mathcal{L}^*\big)\big] \leq c_6K\sum_{k=0}^{k_0-1} 2^{k(\beta+d-\alpha\beta)}\log\big(n2^{-k(2\beta+d)}\big). \tag{5.12}$$

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

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

Observe that

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

where $c_7 := 2^{d+\beta-\beta\alpha} - 1$. Therefore, (5.12) can be rewritten as

$$\begin{aligned}
&\mathbb{E}\bigl[\tilde{r}_n(\mathcal{T}^* \setminus \mathcal{L}^*)\bigr] \\
&\quad\le c_6 K\Biggl[\sum_{k=1}^{k_0-1}(S_k - S_{k-1})\log\bigl(n[c_7 S_k + 1]^{-(d+2\beta)/(d+\beta-\beta\alpha)}\bigr) + \log n\Biggr] \\
&\quad\le c_6 K\Biggl[\int_0^{S_{k_0-1}}\log\bigl(n[c_7 x + 1]^{-(d+2\beta)/(d+\beta-\beta\alpha)}\bigr)\,\mathrm{d}x + \log n\Biggr] \qquad (5.13)\\
&\quad\le c_8 K\bigl[2^{k_0(d+\beta-\beta\alpha)}\log\bigl(n2^{-k_0(d+2\beta)}\bigr) + \log n\bigr] \\
&\quad\le c_9 n\biggl(\frac{n}{K\log(K)}\biggr)^{-\beta(1+\alpha)/(d+2\beta)},
\end{aligned}$$

where we used (5.1) in the last inequality and the fact that $\log(n)$ is dominated by $n^{1-\beta(1+\alpha)/(d+2\beta)}$ since $\alpha\beta \le d$.

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 (5.7), we find that for any $B \in \mathcal{L}^*$, it holds

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

Since $n \ge n_0$, then $c_1 2^{-k_0\beta} \le \delta_0$ and using the margin assumption, we find

$$\begin{aligned}
\sum_{B\in\mathcal{L}^*}\mathbb{E}\bigl[r_n^{\mathrm{live}}(B)\mathbb{1}(\mathcal{G}_B)\bigr] &\le c_1 n 2^{-k_0\beta(1+\alpha)} \\
&\le c_1 n\biggl(\frac{n}{K\log(K)}\biggr)^{-\beta(1+\alpha)/(d+2\beta)}, \qquad (5.14)
\end{aligned}$$

where we used (5.1) in the second inequality.

The theorem follows by summing (5.13) and (5.14). If $\alpha = +\infty$, then the same proof holds except that $\log(n)$ dominates $2^{k_0(\beta+d-\alpha\beta)}\log(n2^{-k_0(2\beta+d)})$ in (5.13). □

APPENDIX: 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 $a \le Z_t \le b$ 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 $T \geq 1$,

$$\mathbb{P}\left\{\bar{Z}_T \geq \sqrt{\frac{(b-a)^2}{2T}\log\left(\frac{1}{\delta}\right)}\right\} \leq \delta. \tag{A.1}$$

The following lemma is a generalization of this result:

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

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

PROOF. Define $\varepsilon_t = \sqrt{\frac{2(b-a)^2}{t}\log(\frac{4}{\delta}\frac{T}{t})}$. Recall first the Hoeffding–Azuma maximal concentration inequality. For every $\eta > 0$ and every integer $t \geq 1$,

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

Using a peeling argument, one obtains

$$\begin{aligned}
\mathbb{P}\{\exists t \leq T, \bar{Z}_t \geq \varepsilon_t\} &\leq \sum_{m=1}^{\lfloor \log_2(T)\rfloor} \mathbb{P}\left\{\bigcup_{t=2^m}^{2^{m+1}-1} \{\bar{Z}_t \geq \varepsilon_t\}\right\} \\
&\leq \sum_{m=1}^{\lfloor \log_2(T)\rfloor} \mathbb{P}\left\{\bigcup_{t=2^m}^{2^{m+1}} \{\bar{Z}_t \geq \varepsilon_{2^{m+1}}\}\right\} \\
&\leq \sum_{m=1}^{\lfloor \log_2(T)\rfloor} \mathbb{P}\left\{\bigcup_{t=2^m}^{2^{m+1}} \{t\bar{Z}_t \geq 2^m\varepsilon_{2^{m+1}}\}\right\} \\
&\leq \sum_{m=1}^{\lfloor \log_2(T)\rfloor} \exp\left(-\frac{2(2^m\varepsilon_{2^{m+1}})^2}{2^{m+1}(b-a)^2}\right) \\
&= \sum_{m=1}^{\lfloor \log_2(T)\rfloor} \frac{2^{m+1}}{T}\frac{\delta}{4} \leq \frac{2^{\log_2(T)+2}}{T}\frac{\delta}{4} \leq \delta.
\end{aligned}$$

Hence the result. □

REFERENCES

[1] AUDIBERT, J.-Y. and BUBECK, S. (2010). Regret bounds and minimax policies under partial monitoring. *J. Mach. Learn. Res.* **11** 2785–2836. MR2738783

[2] AUDIBERT, J.-Y. and TSYBAKOV, A. B. (2007). Fast learning rates for plug-in classifiers. *Ann. Statist.* **35** 608–633. MR2336861

[3] AUDIBERT, J. Y. and TSYBAKOV, A. B. B. (2005). Fast learning rates for plug-in classifiers under the margin condition. Preprint, Laboratoire de Probabilités et Modèles Aléatoires, Univ. Paris VI and VII. Available at arXiv:math/0507180.

[4] AUER, P., CESA-BIANCHI, N. and FISCHER, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.* **47** 235–256.

[5] AUER, P. and ORTNER, R. (2010). UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Period. Math. Hungar.* **61** 55–65. MR2728432

[6] BATHER, J. A. (1981). Randomized allocation of treatments in sequential experiments. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **43** 265–292. MR0637940

[7] CESA-BIANCHI, N. and LUGOSI, G. (2006). *Prediction, Learning, and Games*. Cambridge Univ. Press, Cambridge. MR2409394

[8] EVEN-DAR, E., MANNOR, S. and MANSOUR, Y. (2006). Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *J. Mach. Learn. Res.* **7** 1079–1105. MR2274398

[9] GOLDENSHLUGER, A. and ZEEVI, A. (2009). Woodroofe's one-armed bandit problem revisited. *Ann. Appl. Probab.* **19** 1603–1633. MR2538082

[10] GOLDENSHLUGER, A. and ZEEVI, A. (2011). A note on performance limitations in bandit problems with side information. *IEEE Trans. Inform. Theory* **57** 1707–1713. MR2815844

[11] HAZAN, E. and MEGIDDO, N. (2007). Online learning with prior knowledge. In *Learning Theory. Lecture Notes in Computer Science* **4539** 499–513. Springer, Berlin. MR2397608

[12] JUDITSKY, A., NAZIN, A. V., TSYBAKOV, A. B. and VAYATIS, N. (2008). Gap-free bounds for stochastic multi-armed bandit. In *Proceedings of the* 17*th IFAC World Congress*.

[13] KAKADE, S., SHALEV-SHWARTZ, S. and TEWARI, A. (2008). Efficient bandit algorithms for online multiclass prediction. In *Proceedings of the* 25*th Annual International Conference on Machine Learning* (*ICML* 2008) (A. McCallum and S. Roweis, eds.) 440–447. Omnipress, Helsinki, Finland.

[14] LAI, T. L. and ROBBINS, H. (1985). Asymptotically efficient adaptive allocation rules. *Adv. in Appl. Math.* **6** 4–22. MR0776826

[15] LANGFORD, J. and ZHANG, T. (2008). The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in Neural Information Processing Systems* 20 (J. C. Platt, D. Koller, Y. Singer and S. Roweis, eds.) 817–824. MIT Press, Cambridge, MA.

[16] LU, T., PÁL, D. and PÁL, M. (2010). Showing relevant ads via Lipschitz context multi-armed bandits. *JMLR*: *Workshop and Conference Proceedings* **9** 485–492.

[17] MAMMEN, E. and TSYBAKOV, A. B. (1999). Smooth discrimination analysis. *Ann. Statist.* **27** 1808–1829. MR1765618

[18] RIGOLLET, P. and ZEEVI, A. (2010). Nonparametric bandits with covariates. In *COLT* (A. Tauman Kalai and M. Mohri, eds.) 54–66. Omnipress, Haifa, Israel.

[19] ROBBINS, H. (1952). Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.* (*N.S.*) **58** 527–535. MR0050246

[20] SLIVKINS, A. (2011). Contextual bandits with similarity information. *JMLR*: *Workshop and Conference Proceedings* **19** 679–701.

[21] TSYBAKOV, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32** 135–166. MR2051002

[22] VOGEL, W. (1960). An asymptotic minimax theorem for the two armed bandit problem. *Ann. Math. Statist.* **31** 444–451. MR0116443

[23] WANG, C.-C., KULKARNI, S. R. and POOR, H. V. (2005). Bandit problems with side observations. *IEEE Trans. Automat. Control* **50** 338–355. MR2123095

[24] WOODROOFE, M. (1979). A one-armed bandit problem with a concomitant variable. *J. Amer. Statist. Assoc.* **74** 799–806. MR0556471

[25] YANG, Y. and ZHU, D. (2002). Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates. *Ann. Statist.* **30** 100–121. MR1892657

LPMA, UMR 7599
UNIVERSITÉ PARIS DIDEROT
175, RUE DU CHEVALERET
75013 PARIS
FRANCE
E-MAIL: vianney.perchet@normalesup.org

DEPARTMENT OF OPERATIONS RESEARCH
AND FINANCIAL ENGINEERING
PRINCETON UNIVERSITY
PRINCETON, NEW JERSEY 08544
USA
E-MAIL: rigollet@princeton.edu