Nonparametric Bandits with Covariates

Philippe Rigollet [*] Assaf Zeevi [†]

March 9, 2010

Abstract

We consider a bandit problem which involves sequential sampling from two populations (arms). Each arm produces a noisy reward realization which depends on an observable random *covariate*. The goal is to maximize cumulative expected reward. We derive general lower bounds on the performance of any admissible policy, and develop an algorithm whose performance achieves the order of said lower bound up to logarithmic terms. This is done by decomposing the global problem into suitably "localized" bandit problems. Proofs blend ideas from nonparametric statistics and traditional methods used in the bandit literature.

Mathematics Subject Classification: Primary 62G08, Secondary 62L12, 62L05, 62C20.
Key Words: Bandit, regression, regret, inferior sampling rate, minimax rate.

1 Introduction

The seminal paper of Robbins (1952) 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 two–armed bandit problem can be described as follows. Consider two statistical populations (arms), where at each point in time it is possible to sample from only one of the two 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 (or discounted) rewards over a finite (or infinite) time horizon. The difference between the performance of said sampling policy and that of an oracle, that repeatedly samples from the population with the higher mean reward, is called the *regret*. Thus, one can re-phrase the objective as minimizing the regret.

The original motivation for bandit-type problems originates from treatment allocation in clinical trials; see, e.g., Lai and Robbins (1985) for further discussion and references therein. Here patients enter sequentially and receive one of several treatments. The efficacy of each treatment is unknown, and for each patient a noisy measurement of it is recorded. The goal is to assign as many patients as possible to the best treatment. An example of more recent work can be found in the area of web-based advertising, and more generally customized marketing.

[*]Princeton University. Partially supported by the National Science Foundation (DMS-0906424).
[†]Columbia University.

An on-line publisher needs to choose one of several ads to present to consumers, where the efficacy of these ads is unknown. The publisher observes click-through-rates (CTRs) for each ad, which provide a noisy measurement of the efficacy, and based on that needs to assign ads that maximize CTR.

When the populations being sampled are homogenous, i.e., when the sequential rewards are independent and identically distributed (iid) in each arm, Lai and Robbins (1985) proposed a family of policies that at each step compute the empirical mean reward in each arm, and adds to that a confidence bound that accounts for uncertainty in these estimates. These so-called upper-confidence-bound (UCB) policies were shown to be asymptotically optimal. In particular, it is proven in Lai and Robbins (1985) that such a policy incurs 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 Auer et al. (2002). The elegance of the theory and sharp results developed in Lai and Robbins (1985) 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 Woodroofe (1979) under parametric assumptions and in a somewhat restricted setting; see Goldenshluger and Zeevi (2009) and Wang et al. (2005) for two 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 was that of Yang and Zhu (2002). In particular, they assumed the mean response in each arm, conditional on the covariate value, follows a general functional form, hence one can view their setting as as *nonparametric* bandit problem. They proposed a policy that is based on estimating each response function, and then, rather than greedily choosing the arm with the highest estimated mean response given the covariate, allows with some small probability of selecting a potentially inferior arm. (This is a variant of ε-greedy policies; see Auer et al. (2002).) If the nonparametric estimators of the arms' functional response are consistent, and the randomization is chosen in a suitable manner, then the above policies ensure that the average regret tends to zero as the time horizon n grows to infinity. In the typical bandit terminology, such policies are said to be *consistent*. However, it is unclear whether they 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. Moreover, the study by Yang and Zhu (2002) does not spell out the connection between the characteristics of the class of response functions, and the resulting complexity of the nonparametric bandit problem.

The purpose of the present paper is to further understanding of nonparametric bandit problems, deriving regret-optimal policies and shedding light on some of the elements that dictate the complexity of such problems. We make only two assumptions on the underlying functional form that governs the arms' responses. The first is a mild smoothness condition. Smoothness assumptions can be exploited using "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 Audibert and Tsybakov (2007). Minimum contrast type policies have

already received some attention in the bandit literature with side information, aka *contextual bandits*, in the papers of Langford and Zhang (2008) and also Kakade et al. (2008). 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 by Hazan and Megiddo (2007), where the authors use discretization technique similar to the one employed in this paper. It isi 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 nsame suggests there are uncountably many arms. Notable entries in that stream of work are Slivkins (2009) and Lu et al. (2009), who impose a smoothness condition both on the space of arms and the space of covariates, obtaining optimal regret bounds up to logarithmic terms.

The second key assumption in our paper is a so-called *margin condition*, as it has been come be known in the full information setup; cf. Tsybakov (2004). In that setting, it has been shown to critically affect the complexity classification problems Tsybakov (2004); Boucheron et al. (2005); Audibert and Tsybakov (2007). In the bandit setup, this condition encodes the "separation" between the functions that describe the arms' responses and was originally studied by Goldenshluger and Zeevi (2009) in the one armed bandit problem; see further discussion in section 2. We will see later that the margin condition is a natural measure of complexity in the nonparametric bandit problem.

In this paper, we introduce a family of policies called UCBograms. The term is indicative of two salient ingredients of said policies: they build on *regressogram* estimators; and augment the resulting mean response estimates with upper-confidence-bound terms. The idea of the regressogram is quite natural and easy to implement. It groups the covariate vectors into bins and then estimates, by means of simple averaging, a constant which is a proxy for the mean response of each arm over each such bin. One then views these bins as indexing "local" bandit problems, which are solved by applying a suitable UCB-type modification, following the logic of Lai and Robbins (1985) and Auer et al. (2002). In other words, this family of policies decomposes the non-parametric bandit problem into a sequence of localized standard bandit problems; see section 3 for a complete description. The idea of binning covariates lends itself to natural implementation in the two motivating examples described earlier: patients and consumers are segmented into groups with "similar" characteristics; and then the treatment or ad is allocated based on the characteristic response over that group.

In terms of performance, we prove that the UCBogram policies achieve a regret that is fairly large compared to typical orders of regret observed in the literature. In particular, as opposed to a bounded or logarithmic growth, in our setting the order of the regret is *polynomial* in the time horizon n; see Theorem 3.1. One may question, especially given the simple structure and logic underlying the UCBogram policy, whether this is the best that can be achieved in such problems. To that end, we prove a lower bound which demonstrates that for any admissible policy there exist arm response functions satisfying our assumptions for which one cannot improve on the polynomial order of the upper bound established in Theorem 3.1; see Theorem 4.1. Finally, beyond these analytical results, in our view 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.

2 Description of the problem

2.1 Machine and game

A *bandit machine with covariates* is characterized by a sequence

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

of independent random vectors, where $\big(X_t\big)$, $t = 1, 2, \dots$ is a sequence of iid covariates in $\mathcal{X} \subset \mathbb{R}^d$ with probability distribution P_X, and $Y_t^{(i)}$ denotes the random reward yielded by arm i at time t. We assume that, for each $i = 1, 2$, conditionally on $\{X_t = j\}$, the rewards $Y_t^{(i)}, t = 1, \dots, n$ are i.i.d random variables in $[0, 1]$ with conditional expectation given by

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

where $f^{(i)}, i = 1, 2$, are unknown functions such that $0 \le f^{(i)}(x) \le 1$, for any $i = 1, 2,\ x \in \mathcal{X}$. A natural example arises when $Y_t^{(i)}$ 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 two arms at each time $t = 1, \dots, n$. A *non-anticipating policy* $\pi = \{\pi_t\}$ is a sequence of random functions $\pi_t : \mathcal{X} \to \{1, 2\}$ 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 rule* $\pi^\star$, refers to the strategy that would be played by an omniscient operator with complete knowledge of the functions $f^{(i)}, i = 1, 2$. Given side information X_t, the oracle policy $\pi^\star$ prescribes the arm with the largest expected reward, i.e.,

$$\pi^\star(X_t) := \arg\max_{i=1,2} f^{(i)}(X_t)\,.$$

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

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

Also, let $S_n(\pi)$ denote the *inferior sampling rate* at time n defined by

$$S_n(\pi) := \mathbb{E}\sum_{t=1}^{n} \mathbb{1}(\pi_t(X_t) \neq \pi_t^\star(X_t), f^{(1)}(X_t) \neq f^{(2)}(X_t))\,, \tag{1}$$

where $\mathbb{1}(A)$ is the indicator function that takes value 1 if event A is realized and 0 otherwise. The quantity $S_n(\pi)$ measures the expected number of times at which a *strictly* suboptimal arm has been pulled, and note that in our setting the suboptimal arm varies as a function of the covariate value x.

Without further assumptions on the machine, the game can be arbitrarily difficult and, as a result, the regret and inferior sampling rate can be arbitrarily close to n. In the following subsection, we describe natural assumptions on the regularity of the machine that allow to control its complexity.

2.2 Smoothness and margin conditions

As usual in nonparametric estimation we first impose some regularity on the functions $f^{(i)}, i = 1, 2$. Here and in what follows we use $\|\cdot\|$ to denote the Euclidean norm.

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

$$|f^{(i)}(x) - f^{(i)}(x')| \leq L\|x - x'\|^{\beta}, \quad \forall\, x, x' \in \mathcal{X}, i = 1, 2 \tag{2}$$

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

Notice that a direct consequence of the smoothness condition with parameters (β, L) is that the function $\Delta := |f^{(1)} - f^{(2)}|$ also satisfies the smoothness condition with parameters $(\beta, 2L)$. The behavior of function Δ critically controls the complexity of the problem and the smoothness condition gives a local upper bound on this function. The second condition imposed gives a lower bound on this function though in a weaker global sense. It is closely related to the margin condition employed in classification Tsybakov (2004); Mammen and Tsybakov (1999), which drives the terminology employed here.

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

$$P_X\big[\, 0 < |f^{(1)}(X) - f^{(2)}(X)| \leq \delta \big] \leq C_\delta \delta^{\alpha}\,, \quad \forall\, \delta \in [0, \delta_0]$$

for some $\alpha > 0$.

In what follows, we will focus our attention on marginals P_X that are equivalent to the Lebesgue measure on a compact subset of $\mathbb{R}^d$. In that way, the margin condition will only contain information about the behavior of the function Δ and not the marginal P_X itself. 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 difficulty to distinguish among them. This will be reflected in the bounds on the regret which are 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 accurately quantifies the extent to which the conditions are conflicting.

Proposition 2.1 *Under the smoothness condition with parameters (β, L), any machine that satisfies the margin condition with parameter α such that $\alpha\beta > 1$ exhibits an oracle policy $\pi^\star$ which dictates pulling only one of the two arms all the time, P_X-almost surely. Conversely, if $\alpha\beta \leq 1$ there exists machines with nontrivial oracle policies.*

Proof. The first part of the proof is a straightforward consequence of Proposition 3.4 in Audibert and Tsybakov (2007). To prove the second part, consider the following example. Assume that $d = 1$, $\mathcal{X} = [0, 2]$, $f^{(2)} \equiv 0$ and $f^{(1)}(x) = L\text{sign}(x-1)|x-1|^{1/\alpha}$. Notice that $f^{(1)}$ satisfies the smoothness condition with parameters (β, L) if and only if $\alpha\beta \leq 1$. The oracle policy is not trivial and defined by $\pi^\star(x) = 2$ if $x \leq 1$ and $\pi^\star(x) = 1$ if $x > 1$. Moreover, it can be easily shown that the machine satisfies the margin condition with parameter α and with $\delta_0 = C_\delta = 1$. ∎

3 Policy and main result

We first outline a policy to operate the bandit machine described in the previous section. Then we state the main result which is an upper bound on the regret for this policy. Finally, we state a proposition which allows us to translate the bound on the regret into a bound on the inferior sampling rate.

3.1 Binning and regressograms

To design a policy that solves the bandit problem described in the previous section, one has to inevitably find an estimate of the functions $f^{(i)}, i = 1, 2$ at the current point X_t. There exists a wide variety of nonparametric regression estimators ranging from local polynomials to wavelet estimators. However, a very simple piecewise constant estimator, commonly referred to as *regressogram* will be particularly suitable for our purposes.

Assume now that $\mathcal{X} = [0, 1]^d$ and let $\{B_j, j = 1, \ldots, M^d\}$ be the regular partition of $\mathcal{X}$, i.e., the reindexed collection of hypercubes defined for $\mathsf{k} = (k_1, \ldots, k_d) \in \{1, \ldots, M\}^d$,

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

For each arm $i = 1, 2$, consider the average reward for each bin $B_j, j = 1, \ldots, M^d$ defined by

$$\bar{f}_j^{(i)} = \frac{1}{p_j} \int_{B_j} f^{(i)}(x) \mathrm{d}x \,,$$

where $p_j = P_X(B_j)$. By analogy with histograms, the empirical counterpart of the piecewise constant function $x \mapsto \sum_{j=1}^{M^d} \bar{f}_j^{(i)} \mathbb{1}(x \in B_j)$, is often called *regressogram*. To define it, we need the following quantities. Let $N_t^{(i)}(j, \pi)$ denote the number of times π prescribed to pull arm i at times anterior to t when the covariate was in bin B_j,

$$N_t^{(i)}(j, \pi) = \sum_{s=1}^{t} \mathbb{1}(X_s \in B_j, \pi_s(X_s) = i) \,,$$

and let $\overline{Y}_t^{(i)}(j, \pi)$ denote the average reward collected at those times,

$$\overline{Y}_t^{(i)}(j, \pi) = \frac{1}{N_t^{(i)}(j, \pi)} \sum_{s=1}^{t} Y_s^{(i)} \mathbb{1}(X_s \in B_j, \pi_s(X_s) = i) \,,$$

where here and throughout this paper, we use the convention $1/0 = \infty$. For any arm $i = 1, 2$ and any time $t \geq 1$ the regressograms obtained from a policy π at time t are defined by the following piecewise constant estimators

$$\hat{f}^{(i)}_{t,\pi}(x) = \sum_{j=1}^{M^d} \overline{Y}^{(i)}_t(j, \pi) \mathbb{1}(x \in B_j) \, .$$

While regressograms are rather rudimentary nonparametric estimators of the functions $f^{(i)}$, they allow us to decompose the original problem into a collection of M^d traditional bandit machines without covariates, each one corresponding to a different bin.

3.2 The UCBogram

The "UCBogram" is an index type policy based on upper confidence bounds for the regressogram defined above. Upper confidence bounds (UCB) policies are known to perform optimally in the traditional two armed bandit problem, i.e., without covariates Lai and Robbins (1985); Auer et al. (2002). The index of each arm is computed as the sum of the average past reward and a stochastic term accounting for the deviations of the observed average reward from the true average reward. In the UCBogram, the average reward is simply replaced by the value of the regressogram at the current covariate X_t.

For any $s \geq 1$ the upper confidence bound at time t bound is of the form

$$U_t(s) = \sqrt{\frac{2 \log t}{s}} \, .$$

The UCBogram $\hat{\pi}$ is defined as follows. For any $x \in [0, 1]^d$, define

$$N^{(i)}_t(x) = \sum_{j=1}^{M^d} N^{(i)}_t(j, \hat{\pi}) \mathbb{1}(x \in B_j) \, ,$$

the number of times the UCBogram prescribed to pull arm i at times anterior to t when the covariate was in the same bin as x. Then $\hat{\pi} = (\hat{\pi}_1, \hat{\pi}_2, \ldots)$ is defined recursively by

$$\hat{\pi}_t(x) = \arg\max_{i=1,2} \left\{ \hat{f}^{(i)}_{t,\hat{\pi}}(x) + U_t(N^{(i)}_t(x)) \right\} .$$

Notice that the UCBogram is indeed a UCB-type policy. Indeed, for each arm $i = 1, 2$ and at each point x, it computes an estimator $\hat{f}^{(i)}_{t,\pi}(x)$ of the expected reward and adds an upper confidence bound $U_t(N^{(i)}_t(x))$ to account for stochastic variability in this estimator. The most attractive feature of the regressogram is that it allows to decompose the nonparametric bandit problem into independently operated local machines as detailed in the proof of the following theorem.

Theorem 3.1 *Fix $\beta \in (0,1]$, $L > 0$ and $\alpha \in (0,1]$. Let $\mathcal{X} = [0,1]^d$ and assume that the covariates X_t have a distribution which is equivalent*[1] *to the Lebesgue measure on the unit hypercube $\mathcal{X}$. Let the machine satisfy both the smoothness condition with parameter (β, L) and the margin condition with parameter $0 < \alpha \leq 1$. Then the UCBogram policy $\hat{\pi}$ with $M = \lfloor (n/\log n)^{1/(2\beta+d)} \rfloor$ has an expected cumulative regret at time n bounded by*

$$R_n(\hat{\pi}) \leq Cn \max\left\{ \left(\frac{n}{\log n}\right)^{-\frac{\beta(\alpha+1)}{2\beta+d}}, \left(\frac{n}{(\log n)^2}\right)^{-\frac{2\beta}{2\beta+d}} \right\},$$

where $C > 0$ is a positive constant.

Proof. To keep track of positive constants, we number them $c_1, c_2, \ldots$. Define $c_1 = 2Ld^{\beta/2} + 1$, and let $n_0 \geq 2$ be the largest integer such that

$$\left(\frac{n_0}{\log n_0}\right)^{\beta/(2\beta+d)} \leq \frac{2c_1}{\delta_0},$$

where δ_0 is the constant appearing in the margin condition. If $n \leq n_0$, we have $R_n \leq n_0$ 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 UCBogram policy $\hat{\pi}$ is a collection of functions $\hat{\pi}_t$ that are constant on each B_j, equal to $\hat{\pi}_t(j)$. Define the regret $\mathsf{R}_j(\hat{\pi})$ on bin B_j by

$$\mathsf{R}_j(\hat{\pi}) = \sum_{t=1}^{n} \left(f^{(\pi^\star(X_t))}(X_t) - f^{(\hat{\pi}_t(j))}(X_t)\right) \mathbb{1}(X_t \in B_j),$$

and observe that the overall regret of $\hat{\pi}$ can be written as

$$R_n(\hat{\pi}) = \sum_{j=1}^{M^d} \mathbb{E}\mathsf{R}_j(\hat{\pi}).$$

Consider the set of "well behaved" bins on which the expected reward functions of the two arms are well separated:

$$\mathcal{J} = \{j \,:\, \exists\, x \in B_j\,,\, |f^{(1)}(x) - f^{(2)}(x)| > c_1 M^{-\beta}\}.$$

For any $j \notin \mathcal{J}$ and any $x \in B_j$, we have $|f^{(1)}(x) - f^{(2)}(x)| \leq c_1 M^{-\beta} < \delta_0$ so that

$$\mathbb{E}\mathsf{R}_j(\hat{\pi}) \leq c_1 M^{-\beta} \sum_{t=1}^{n} \mathbb{P}\left[0 < |f^{(1)}(X_t) - f^{(2)}(X_t)| \leq c_1 M^{-\beta}, X_t \in B_j\right],$$

Summing over $j \notin \mathcal{J}$, we obtain from the margin condition that

$$\sum_{j \notin \mathcal{J}} \mathbb{E}\mathsf{R}_j(\hat{\pi}) \leq C_\delta c_1^{1+\alpha} n M^{-\beta(1+\alpha)}. \tag{3}$$

[1] Two measures μ and ν are said to be *equivalent* if there exist two positive constants $\underline{c}$ and $\bar{c}$ such that $\underline{c}\mu(A) \leq \nu(A) \leq \bar{c}\mu(A)$ for any measurable set A.

We now treat the well behaved bins, i.e., bins B_j such that $j \in \mathcal{J}$. Notice that since each bin is a hypercube with side length $1/M$ and since the reward functions satisfy the smoothness condition with parameters (β, L), we have

$$|f^{(1)}(x) - f^{(2)}(x)| > c_1 M^{-\beta} - 2Ld^{\beta/2}M^{-\beta} = M^{-\beta}\,,$$

for *any* $x \in B_j, j \in \mathcal{J}$. In particular, for such j, since the two functions are continuous, the difference $f^{(1)}(x) - f^{(2)}(x)$ has constant sign over B_j and $|\bar{f}_j^{(1)} - \bar{f}_j^{(2)}| > M^{-\beta}$. As a consequence, the oracle policy $\pi^\star$ is constant on B_j, equal to $\pi^\star(j)$ for any $j \in \mathcal{J}$ and, conditionally on $\{X_t \in B_j\}$, the game can be viewed as a standard bandit problem, i.e., without covariates, where arm i has bounded reward with mean $\bar{f}_j^{(i)}$. Moreover, conditionally on $\{X_t \in B_j\}$, the UCBogram can be seen as a standard UCB policy. Applying for example Theorem 1 in Auer et al. (2002), we find that for $j \in \mathcal{J}$,

$$\mathbb{ER}_j(\hat{\pi}) \leq \Big[\big(1 + \frac{\pi^2}{3}\big)\Delta_j\Big] + \frac{8\log n}{\Delta_j} \leq c_2 \frac{\log n}{\Delta_j}\,, \tag{4}$$

where $\Delta_j = |\bar{f}_j^{(1)} - \bar{f}_j^{(2)}|$ is the average gap in bin B_j. We now use the margin condition to provide lower bounds on Δ_j. Assume without loss of generality that the gaps are ordered $0 < \Delta_1 \leq \Delta_2 \leq \ldots, \leq \Delta_{M^d}$ and define the integers j_1, j_2 such that $\mathcal{J} = \{j_1, \ldots, M^d\}$ and $j_2 \in \{j_1, \ldots, M^d\}$ is the largest integer such that $\Delta_{j_2} \leq \delta_0/c_1$. Therefore, for any $j \in \{j_1, \ldots, j_2\} \subset \mathcal{J}$, we have on the one hand,

$$P_X\big[\,0 < |f^{(1)} - f^{(2)}| \leq \Delta_j + (c_1 - 1)M^{-\beta}\big] \geq \sum_{k=1}^{M^d} p_k \mathbb{1}(0 < \Delta_k \leq \Delta_j) \geq \frac{\underline{c} j}{M^d}\,, \tag{5}$$

where we use the fact that $p_k = P_X(B_k) \geq \underline{c}/M^d$ since P_X is equivalent to the Lebesgue measure on $[0,1]^d$ (see footnote 1). On the other hand, the margin condition yields for any $j \in \{j_1, \ldots, j_2\}$ that,

$$P_X\big[\,0 < |f^{(1)} - f^{(2)}| \leq \Delta_j + (c_1 - 1)M^{-\beta}\big] \leq C_\delta\big(c_1\Delta_j\big)^\alpha\,. \tag{6}$$

where we used the fact that $\Delta_j + (c_1 - 1)M^{-\beta} \leq c_1\Delta_j \leq \delta_0$, for any $j \in \{j_1, \ldots, j_2\}$. The previous two inequalities yield

$$\Delta_j \geq c_3\Big(\frac{j}{M^d}\Big)^{1/\alpha}\,, \quad \forall\, j \in \{j_1, \ldots, j_2\}\,. \tag{7}$$

Combining (3), (4) and (7), we obtain the following bound,

$$R_n(\hat{\pi}) \leq c_4\Big[nM^{-\beta(1+\alpha)} + j_1 M^{-\beta} + (\log n)\sum_{j=j_1}^{j_2}\left(\frac{M^d}{j}\right)^{1/\alpha} + M^d \log n\Big]\,. \tag{8}$$

Note that applying the same arguments as in (5) and (6), we find that j_1 satisfies

$$\frac{\underline{c} j_1}{M^d} \leq P_X\big[\,0 < |f^{(1)} - f^{(2)}| \leq c_1 M^{-\beta}\big] \leq C_\delta\big(c_1 M^{-\beta}\big)^\alpha\,,$$

so that $j_1 \leq c_5 M^{d-\alpha\beta}$. We now bound from above the sum in (8) using the following integral approximation:

$$\sum_{j=j_1}^{j_2} \left(\frac{M^d}{j}\right)^{1/\alpha} \leq \sum_{j=j_1}^{M^d} \left(\frac{M^d}{j}\right)^{1/\alpha} \leq c_7 M^d \int_{M^{-\alpha\beta}}^{1} x^{-1/\alpha} \mathrm{d}x \,. \tag{9}$$

If $\alpha < 1$, this integral is bounded by $c_6 M^{\beta(1-\alpha)}$ and if $\alpha = 1$, it is bounded by $c_7 \log M$. As a result, the integral in (9) is of order $M^d(M^{\beta(1-\alpha)} \vee \log M)$ and we obtain from (8) that

$$R_n(\hat{\pi}) \leq c_8 \left[nM^{-\beta(1+\alpha)} + M^d(M^{\beta(1-\alpha)} \vee \log M) \log n \right], \tag{10}$$

and the result follows by choosing M as prescribed. ∎

We should point out that the version of the UCBogram described above specifies the number of bins M as a function of the horizon n, while in practice one does not have foreknowledge of this value. This limitation can be easily circumvented by using the so-called *doubling argument* Cesa-Bianchi and Lugosi (2006) which consists of "reseting" the game at times $2^k, k = 1, 2, \ldots$

The reader will note that when $\alpha = 1$ there is an additional $\log n$ factor appearing in the upper bound given in the statement of the theorem. More generally, for any $\alpha > 1$, it is possible to minimize the expression on the right hand side of (10) 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 UCBogram 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 UCBogram splits the space $\mathcal{X}$ in "too many" bins when $\alpha \geq 1$. Intuitively this can be understood as follows. When $\alpha = 1$, the gap function $\Delta(x)$ is bounded away from zero for most $x \in \mathcal{X}$. For such x, there is no need to carefully estimate the gap function since it has constant sign for "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. Of course, such limitations are intrinsic to the UCBogram and may not appear with other policies but it is beyond the scope of this paper.

3.3 The inferior sampling rate

Unlike traditional bandit problems, the connection between the inferior sampling rate defined in (1) and the regret is more intricate here. The following lemma establishes a connections between the two.

Lemma 3.1 *For any $\alpha > 0$, under the margin condition we have*

$$S_n(\pi) \leq C n^{\frac{1}{1+\alpha}} R_n(\pi)^{\frac{\alpha}{1+\alpha}} \,,$$

for any policy π and for some positive constant $C > 0$.

Proof. The idea of the proof is quite standard and originally appeared in Tsybakov (2004). It has been used in Rigollet and Vert (2009) and Goldenshluger and Zeevi (2009). Define the two random quantities:

$$r_n(\pi) = \sum_{t=1}^{n} |f^{(1)}(X_t) - f^{(2)}(X_t)| \mathbb{1}(\pi_t(X_t) \neq \pi^\star(X_t)) \,,$$

and

$$s_n(\pi) = \sum_{t=1}^{n} \mathbb{1}(f^{(1)}(X_t) \neq f^{(2)}(X_t), \pi_t \neq \pi^\star(X_t)) \,.$$

We have

$$\begin{aligned} r_n(\pi) &\geq \delta \sum_{t=1}^{n} \mathbb{1}(\pi_t(X_t) \neq \pi^\star(X_t)) \mathbb{1}(|f^{(1)}(X_t) - f^{(2)}(X_t)| > \delta) \\ &\geq \delta \big[s_n(\pi) - \sum_{t=1}^{n} \mathbb{1}(\pi_t(X_t) \neq \pi^\star(X_t), 0 < |f^{(1)}(X_t) - f^{(2)}(X_t)| \leq \delta) \big] \\ &\geq \delta \big[s_n(\pi) - \sum_{t=1}^{n} \mathbb{1}(0 < |f^{(1)}(X_t) - f^{(2)}(X_t)| \leq \delta) \big] \,. \end{aligned} \tag{11}$$

Taking expectations on both sides of (11), we obtain that $R_n(\pi) \geq \delta \big[S_n(\pi) - n\delta^\alpha \big]$, where we used the margin condition. The proof follows by choosing $\delta = (S_n(\pi)/cn)^{1/\alpha}$ for $c \geq 2$ large enough to ensure that $\delta < \delta_0$ ∎

Using Lemma 3.1, we obtain the following corollary of Theorem 3.1

Corollary 3.1 *Fix $\beta \in (0, 1]$, $L > 0$ and $\alpha \in (0, 1]$. Under the conditions of Theorem 3.1, the UCBogram policy $\hat{\pi}$ with $M = \lfloor (n/\log n)^{1/(2\beta+d)} \rfloor$ has an inferior sampling rate at time n bounded by*

$$S_n(\hat{\pi}) \leq Cn \Big(\frac{n}{\log n} \Big)^{-\frac{\beta\alpha}{2\beta+d}} \,.$$

where $C > 0$ is a positive constant.

4 Lower bound

While the UCBogram is a very simple policy, it still provides good insights as to how to construct a lower bound on the regret for incurred by any admissible policy. Indeed, the main result of this section demonstrates the polynomial rate of the upper bounds in Theorem 3.1 and Corollary 3.1 is optimal in a minimax sense, for a large class of conditional reward distributions. Define the Kullback-Leibler (KL) divergence between P and Q, where P and Q are two probability distributions by

$$\mathcal{K}(P, Q) = \begin{cases} \int \log \left(\frac{\mathrm{d}P}{\mathrm{d}Q} \right) \mathrm{d}P & \text{if } P \ll Q \,, \\ \infty & \text{otherwise.} \end{cases}$$

Denote by $P^{(i)}_{f(X)}$ the conditional distribution of $Y^{(i)}$ given X for any $i = 1, 2$ and assume that there exists $\kappa^2 > 0$ such that for any $\theta, \theta' \in \Theta$ the KL divergence between $P^{(i)}_\theta$ and $P^{(i)}_{\theta'}$ satisfies

$$\mathcal{K}(P^{(i)}_\theta, P^{(i)}_{\theta'}) \le \frac{1}{\kappa^2}(\theta - \theta')^2 \,. \tag{12}$$

Assumption (12) is similar to Assumption (B) employed in Tsybakov (2009, Section 2.5) but does not require absolute continuity with respect to the Lebesgue measure. A direct consequence of the following lemma is that Assumption (12) is satisfied when P_θ is a Bernoulli distribution with parameter $\theta \in (0, 1)$.

Lemma 4.1 *For any $a \in [0, 1]$ and $b \in (0, 1)$ let P_a and P_b denote two Bernoulli distributions with parameters a and b respectively. Then*

$$\mathcal{K}(P_a, P_b) \le \frac{(a-b)^2}{b(1-b)} \,.$$

In particular, if $b_0 \in [0, 1/2)$, Assumption (12) is satisfied with $\kappa^2 = 1/4 - b_0^2$, for any $a \in [0, 1], b \in [1/2 - b_0, 1/2 + b_0]$.

Proof. From the definition of the KL divergence, we have

$$\mathcal{K}(P_a, P_b) = a \log\left(\frac{a}{b}\right) + (1-a)\log\left(\frac{1-a}{1-b}\right) \le a\Big(\frac{a-b}{b}\Big) - (1-a)\Big(\frac{a-b}{1-b}\Big) = \frac{(a-b)^2}{b(1-b)}$$

where in the second line we used the inequality $\log(1 + u) \le u$.

■

Theorem 4.1 *Fix $\alpha, \beta, L > 0$ such that $\alpha\beta < 1$ and let $\mathcal{X} = [0, 1]^d$. Assume that the covariates X_t are uniformly distributed on the unit hypercube $\mathcal{X}$ and that there exists $\tau \in (0, 1/2)$ such that $\{P^{(i)}_\theta\,,\ \theta \in [1/2 - \tau, 1/2 + \tau]\}$ satisfies equation (12) for $i = 1, 2$. Then, there exists a pair of reward functions $f^{(i)}, i = 1, 2$ that satisfy both the smoothness condition with parameters (β, L) and the margin condition with parameter α, such that for* any *non-anticipating policy π the regret is bounded as follows*

$$R_n(\pi) \ge C n^{1 - \frac{\beta(\alpha+1)}{2\beta+d}} \,, \tag{13}$$

and the inferior sampling rate is bounded as follows

$$S_n(\pi) \ge C n^{1 - \frac{\beta\alpha}{2\beta+d}} \,, \tag{14}$$

for some positive constant C.

Proof. To simplify the arguments below, it will be useful to denote arm 2 by -1. Finally, with slight abuse of notation, we use $S_n(\pi, f^{(1)}, f^{(-1)})$ to denote the inferior sampling rate at time n that is defined in (1), making the dependence on the mean reward functions explicit.

In view of Lemma 3.1, it is sufficient to prove (14). To do so we reduce our problem to a hypothesis testing problems; an approach this is quite standard in the nonparametric literature, cf. (Tsybakov, 2009, Chapter 2). For any policy π, and any $t = 1, \dots, n$, denote by $\mathbb{P}^t_{\pi,f}$ the joint distribution of the collection of pairs

$$(X_1, Y_1^{(\pi_1(X_1))}), \dots, (X_t, Y_t^{(\pi_t(X_t))})$$

where $\mathbb{E}[Y^{(1)}|X] = f(X)$ and $\mathbb{E}[Y^{(-1)}|X] = 1/2$. Let $\mathbb{E}^t_{\pi,f}$ denote the corresponding expectation. It follows that the oracle policy $\pi^\star_f$ is given by $\pi^\star_f(x) = \mathrm{sign}[f(x)]$ with the convention that $\mathrm{sign}(0) = 1$. Fix $\delta_0 \in (0,1)$ as in the definition of the margin condition. We now construct a class $\mathcal{C}$ of functions $f : \mathcal{X} \to [0,1]$ such that f satisfies (2) and

$$P_X\big[\, 0 < |f(X) - 1/2| \le \delta \big] \le C_\delta \delta^\alpha \,, \quad \forall\, \delta \in [0, \delta_0] \,,$$

As a result, the machine characterized by the expected rewards $f^{(1)} = f$ and $f^{(-1)} = 1/2$ satisfies both the smoothness and the margin conditions. Moreover, we construct $\mathcal{C}$ in such a way that for any policy π

$$\sup_{f \in \mathcal{C}} S_n(\pi, f, 1/2) \ge Cn \Big(\frac{n}{\log n} \Big)^{-\frac{\beta\alpha}{2\beta+d}} \,. \tag{15}$$

for some positive constant C. Consider the regular grid $\mathcal{Q} = \{q_1, \dots, q_{M^d}\}$, where q_k denotes the center of bin B_k, $k = 1, \dots, M^d$, for some $M \ge 1$ to be defined. Define $C_\phi = \min(L, \tau, 1/4)$ and let $\phi_\beta : \mathbb{R}^d \to \mathbb{R}_+$ be a smooth function defined as follows:

$$\phi_\beta(x) = \begin{cases} (1 - \|x\|_\infty)^\beta & \text{if } 0 \le \|x\|_\infty \le 1, \\ 0 & \text{if } \|x\|_\infty > 1 \,. \end{cases}$$

Clearly, we have $|C_\phi \phi_\beta(x) - C_\phi \phi_\beta(x')| \le L\|x - x'\|^\beta_\infty \le L\|x - x'\|^\beta$ for any $x, x' \in \mathbb{R}^d$.

Define the integer $m = \lceil \mu M^{d - \alpha\beta} \rceil$, i.e., the smallest integer that is larger than or equal to $\mu M^{d-\alpha\beta}$, where $\mu \in (0,1)$ is chosen small enough to ensure that $m \le M^d$. Define $\Omega_m = \{-1, 1\}^m$ and for any $\omega \in \Omega_m$, define the function f_ω on $[0,1]^d$ by

$$f_\omega(x) = 1/2 + \sum_{j=1}^m \omega_j \varphi_j(x) \,,$$

where $\varphi_j(x) = M^{-\beta} C_\phi \phi(M[x - q_j]) \mathbb{I}(x \in B_j)$. Notice in particular that $f_\omega(x) = 1/2$ if and only if $x \in \mathcal{X} \setminus \bigcup_{j=1}^m B_j$ up to a set of zero Lebesgue measure. We are now in position to define the family $\mathcal{C}$ as

$$\mathcal{C} = \{ f_\omega \ : \ \omega \in \Omega_m \} \,.$$

Note first that any function $f_\omega \in \mathcal{C}$ satisfies the smoothness condition (2). We now check that the margin condition is satisfied with parameter α. For any $\omega \in \Omega_m$, we have

$$\begin{aligned} P_X(0 < |f_\omega(X) - 1/2| \le C_\phi \delta) &= \sum_{j=1}^{m} P_X(0 < |f_\omega(X) - 1/2| \le C_\phi \delta, X \in B_j) \\ &= m P_X(0 < \phi(M[X - q_1]) \le \delta M^\beta, X \in B_1) \\ &= m \int_{B_1} \mathbb{1}(\phi(Mx) \le \delta M^\beta) \mathrm{d}x \\ &= m M^{-d} \int_{[0,1]^d} \mathbb{1}(\phi(x) \le \delta M^\beta) \mathrm{d}x \,, \end{aligned}$$

where in the third equality, we used the fact that P_X denotes the uniform distribution on $[0,1]^d$. Now, since ϕ is non negative and uniformly bounded by 1, we have on the one hand that for $\delta M^\beta > 1$,

$$\int_{[0,1]^d} \mathbb{1}(\phi(x) \le \delta M^\beta) \mathrm{d}x = 1 \,.$$

On the other hand, when $\delta M^\beta \le 1$, we find

$$\int_{[0,1]^d} \mathbb{1}(\phi(x) \le \delta M^\beta) \mathrm{d}x = 1 - \int_{[0,1]^d} \mathbb{1}(\|x\|_\infty \le 1 - M\delta^{1/\beta}) \mathrm{d}x = 1 - \left(1 - M\delta^{1/\beta}\right)^d \le dM\delta^{1/\beta} \,.$$

It yields

$$\begin{aligned} P_X(0 < |f_\omega(X) - 1/2| \le C_\phi \delta) &\le m M^{-d} \mathbb{1}(\delta M^\beta > 1) + m d M^{1-d} \delta^{1/\beta} \mathbb{1}(\delta M^\beta \le 1)\big) \\ &\le M^{-\alpha\beta} \mathbb{1}(M^{-\alpha\beta} < \delta^\alpha) + d M^{1-\alpha\beta} \delta^{1/\beta} \mathbb{1}(M \le \delta^{-1/\beta}) \\ &\le (1+d)\delta^\alpha \,, \end{aligned}$$

where we used the fact that $1 - \alpha\beta \ge 0$ to bound the second term in the last inequality. Thus, the margin condition is satisfied for any δ_0 and with $C_\delta = (1+d)/C_\phi^\alpha$.

We now prove (15) by observing that if we denote $\omega = (\omega_1, \dots, \omega_m) \in \Omega_m$, we have

$$\begin{aligned} \sup_{f \in \mathcal{C}} S_n(\pi, f^{(1)}, 1/2) &= \sup_{\omega \in \Omega_m} \sum_{t=1}^{n} \mathbb{E}_{\pi, f_\omega}^{t-1} P_X\left[\pi_t(X_t) \ne \mathrm{sign}(f_\omega(X_t))\right] \\ &= \sup_{\omega \in \Omega_m} \sum_{j=1}^{m} \sum_{t=1}^{n} \mathbb{E}_{\pi, f_\omega}^{t-1} P_X\left[\pi_t(X_t) \ne \omega_j, X_t \in B_j\right] \\ &\ge \frac{1}{2^m} \sum_{j=1}^{m} \sum_{t=1}^{n} \sum_{\omega \in \Omega_m} \mathbb{E}_{\pi, f_\omega}^{t-1} P_X\left[\pi_t(X_t) \ne \omega_j, X_t \in B_j\right] \end{aligned} \tag{16}$$

Observe now that for any $j = 1, \dots, m$, the sum $\sum_{\omega \in \Omega}[\cdots]$ in the previous display can be decomposed as

$$Q_j^t = \sum_{\omega_{[-j]} \in \Omega_{m-1}} \sum_{i \in \{-1,1\}} \mathbb{E}_{\pi, f_{\omega_{[-j]}^i}}^{t-1} P_X\left[\pi_t(X_t) \ne i, X_t \in B_j\right] \,,$$

where $\omega_{[-j]} = (\omega_1, \ldots, \omega_{j-1}, \omega_{j+1}, \ldots, \omega_m)$ and $\omega^i_{[-j]} = (\omega_1, \ldots, \omega_{j-1}, i, \omega_{j+1}, \ldots, \omega_m)$ for $i = -1, 1$. Using Theorem 2.2(iii) of Tsybakov (2009), and denoting by $P^j_X(\cdot)$ the conditional distribution $P_X(\cdot|X \in B_j)$, we get

$$\begin{aligned}
\sum_{i\in\{-1,1\}} \mathbb{E}^{t-1}_{\pi,f_{\omega^i_{[-j]}}} P_X\left[\pi_t(X_t) \neq i, X_t \in B_j\right] &= \frac{1}{M^d} \sum_{i\in\{-1,1\}} \mathbb{E}^{t-1}_{\pi,f_{\omega^i_{[-j]}}} P^j_X\left[\pi_t(X_t) \neq i\right] \\
&\geq \frac{1}{4M^d} \exp\left[-\mathcal{K}\big(\mathbb{P}^{t-1}_{\pi,f_{\omega^{-1}_{[-j]}}} \times P^j_X, \mathbb{P}^{t-1}_{\pi,f_{\omega^1_{[-j]}}} \times P^j_X\big)\right] \\
&= \frac{1}{4M^d} \exp\left[-\mathcal{K}\big(\mathbb{P}^{t-1}_{\pi,f_{\omega^{-1}_{[-j]}}}, \mathbb{P}^{t-1}_{\pi,f_{\omega^1_{[-j]}}}\big)\right] \qquad (17)
\end{aligned}$$

For any $t = 2, \ldots, n$, let $\mathcal{F}_t$ denote the σ-algebra generated by the information available at time t immediately *after* observing X_t, i.e., $\mathcal{F}_t = \sigma\big(X_t, (X_s, Y_s^{(\pi_s(X_s))}), s = 1, \ldots, t-1)\big)$. Define the conditional distribution $\mathbb{P}^{\cdot|\mathcal{F}_t}_{\pi,f}$ of the random couple $(X_t, Y_t^{(\pi_t(X_t))})$, conditioned on $\mathcal{F}_t$. Denote also by E_{X_t} the expectation with respect to the marginal distribution of X_t. Applying the chain rule for KL divergence, we find that for any $t = 1, \ldots, n$ and any $f, g : \mathcal{X} \to [0, 1]$, we have

$$\begin{aligned}
\mathcal{K}\big(\mathbb{P}^t_{\pi,f}, \mathbb{P}^t_{\pi,g}\big) &= \mathcal{K}\big(\mathbb{P}^{t-1}_{\pi,f}, \mathbb{P}^{t-1}_{\pi,g}\big) + \mathbb{E}^{t-1}_{\pi,f} E_{X_t}\left[\mathcal{K}\big(\mathbb{P}^{\cdot|\mathcal{F}_t}_{\pi,f}, \mathbb{P}^{\cdot|\mathcal{F}_t}_{\pi,g}\big)\right] \\
&= \mathcal{K}\big(\mathbb{P}^{t-1}_{\pi,f}, \mathbb{P}^{t-1}_{\pi,g}\big) + \mathbb{E}^{t-1}_{\pi,f} E_{X_t}\left[\mathcal{K}\big(\mathbb{P}^{Y_t^{(\pi_t(X_t))}|\mathcal{F}_t}_{\pi,f}, \mathbb{P}^{Y_t^{(\pi_t(X_t))}|\mathcal{F}_t}_{\pi,g}\big)\right],
\end{aligned}$$

where $\mathbb{P}^{Y_t^{(\pi_t(X_t))}|\mathcal{F}_t}_{\pi,f}$ denotes the conditional distribution of $Y_t^{(\pi_t(X_t))}$ given $\mathcal{F}_t$. Since, for any $f \in \mathcal{C}$, we have that $\mathbb{E}[Y_t^{(\pi_t(X_t))}|\mathcal{F}_t] = f^{(\pi_t(X_t))}(X_t) \in [1/2 - \tau, 1/2 + \tau]$, we can apply (12) to derive the following upper bound:

$$\begin{aligned}
\mathcal{K}\big(\mathbb{P}^{Y_t^{(\pi_t(X_t))}|\mathcal{F}_t}_{\pi,f_{\omega^{-1}_{[-j]}}}, \mathbb{P}^{Y_t^{(\pi_t(X_t))}|\mathcal{F}_t}_{\pi,f_{\omega^1_{[-j]}}}\big) &\leq \frac{1}{\kappa^2}\left(f_{\omega^1_{[-j]}}(X_t) - f_{\omega^{-1}_{[-j]}}(X_t)\right)^2 \mathbb{1}\left(\pi_t(X_t) = 1\right) \\
&\leq \frac{4}{\kappa^2} C_\phi^2 M^{-2\beta} \mathbb{1}\left(\pi_t(X_t) = 1, X_t \in B_j\right) \\
&\leq \frac{M^{-2\beta}}{4\kappa^2} \mathbb{1}\left(\pi_t(X_t) = 1, X_t \in B_j\right).
\end{aligned}$$

By induction, the last two displays yield that for any $t = 1, \ldots, n$,

$$\mathcal{K}(\mathbb{P}^{t-1}_{\pi,f_{\omega^1_{[-j]}}}, \mathbb{P}^{t-1}_{\pi,f_{\omega^{-1}_{[-j]}}}) \leq \frac{M^{-2\beta}}{4\kappa^2} \mathsf{N}_{j,\pi}, \qquad (18)$$

where

$$\mathsf{N}_{j,\pi} = \mathbb{E}^{n-1}_{\pi,f_{\omega^{-1}_{[-j]}}} E_X\left[\sum_{t=1}^n \mathbb{1}\left(\pi_t(X) = 1, X \in B_j\right)\right],$$

denotes the expected number of times t between time 1 and time n that $X_t \in B_j$ and $\pi_t(X_t) = 1$. Combining (17) and (18), we get

$$Q^t_j \geq \frac{2^{m-1}}{4M^d} \exp\left(-\frac{M^{-2\beta}}{4\kappa^2} \mathsf{N}_{j,\pi}\right). \qquad (19)$$

On the other hand, from the definition of Q_j^t, we clearly have

$$\sum_{t=1}^{n} Q_j^t \geq 2^{m-1} \mathsf{N}_{j,\pi} \,. \tag{20}$$

Plugging the lower bounds (19) and (20) into (16) yields

$$\begin{aligned}
\sup_{f \in \mathcal{C}} S_n(\pi, f^{(1)}, 1/2) &\geq \frac{2^{m-1}}{2^m} \sum_{j=1}^{m} \max \left\{ \frac{n}{4M^d} \exp\left(-\frac{M^{-2\beta}}{4\kappa^2} \mathsf{N}_{j,\pi} \right), \mathsf{N}_{j,\pi} \right\} \\
&\geq \frac{1}{4} \sum_{j=1}^{m} \left\{ \frac{n}{4M^d} \exp\left(-\frac{M^{-2\beta}}{4\kappa^2} \mathsf{N}_{j,\pi} \right) + \mathsf{N}_{j,\pi} \right\} \\
&\geq \frac{m}{4} \inf_{z \geq 0} \left\{ \frac{n}{4M^d} \exp\left(-\frac{M^{-2\beta}}{4\kappa^2} z \right) + z \right\}
\end{aligned}$$

Notice now that

$$z^* = \operatorname*{argmin}_{z \geq 0} \left\{ \frac{n}{4M^d} \exp\left(-\frac{M^{-2\beta}}{4\kappa^2} z \right) + z \right\}$$

is strictly positive if and only if $n > 16\kappa^2 M^{2\beta+d}$, in which case

$$z^* = 4\kappa^2 M^{2\beta} \log\left(\frac{n}{16\kappa^2 M^{2\beta+d}} \right) .$$

Taking

$$M = \left\lceil \left(\frac{n}{16 e \kappa^2} \right)^{\frac{1}{2\beta+d}} \right\rceil$$

gives $z^* = c^* n^{\frac{2\beta}{2\beta+d}}$ for some positive constant c^*, so that

$$\sup_{f \in \mathcal{C}} S_n(\pi, f^{(1)}, 1/2) \geq C m z^* \geq C n^{1 - \frac{\alpha\beta}{2\beta+d}}.$$

This completes the proof. ∎

Notice that the rates obtained in Theorem 4.1, can be obtained in the full information case, where the operator observes the whole i.i.d sequence $(X_i, Y_i^{(1)}, Y_i^{(2)}), i = 1, \ldots, n$, even before the first round. Indeed, such bounds have been obtained by Audibert and Tsybakov (2007) in the classification setup, i.e., when the rewards are Bernoulli random variables. However, we state a different technique, tailored for bandit policies in a partial information setup. While the final result is the same, we believe that it sheds light on the technicalities encountered in proving such a lower bound.

References

Audibert, J.-Y. and Tsybakov, A. B. (2007). Fast learning rates for plug-in classifiers. *Ann. Statist.*, **35** 608–633.

Auer, P., Cesa-Bianchi, N. and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, **47** 235–256.

Boucheron, S., Bousquet, O. and Lugosi, G. (2005). Theory of classification: a survey of some recent advances. *ESAIM Probab. Stat.*, **9** 323–375 (electronic).

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge University Press, Cambridge.

Goldenshluger, A. and Zeevi, A. (2009). Woodroofe's one-armed bandit problem revisited. *Ann. Appl. Probab.*, **19** 1603–1633.

Hazan, E. and Megiddo, N. (2007). Online learning with prior knowledge. In *Learning theory*, vol. 4539 of *Lecture Notes in Comput. Sci.* Springer, Berlin, 499–513.

Kakade, S., Shalev-Shwartz, S. and Tewari, A. (2008). Efficient bandit algorithms for online multiclass prediction. In *Proceedings of the 25th Annual International Conference on Machine Learning (ICML 2008)* (A. McCallum and S. Roweis, eds.). Omnipress, 440–447.

Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Adv. in Appl. Math.*, **6** 4–22.

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. Platt, D. Koller, Y. Singer and S. Roweis, eds.). MIT Press, Cambridge, MA, 817–824.

Lu, T., Pál, D. and Pál, M. (2009). Showing relevant ads via context multi-armed bandits. Tech. rep.

Mammen, E. and Tsybakov, A. B. (1999). Smooth discrimination analysis. *Ann. Statist.*, **27** 1808–1829.

Rigollet, P. and Vert, R. (2009). Fast rates for plug-in estimators of density level sets. *Bernoulli*, **15** 1154–1178.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.*, **58** 527–535.

Slivkins, A. (2009). Contextual bandits with similarity information. *Arxiv preprint arXiv:0907.3986*.

Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, **32** 135–166.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated.

Wang, C.-C., Kulkarni, S. and Poor, H. (2005). Bandit problems with side observations. *Automatic Control, IEEE Transactions on*, **50** 338–355.

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

Yang, Y. and Zhu, D. (2002). Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates. *Ann. Statist.*, **30** 100–121.

Philippe Rigollet
Department of Operations Research
and Financial Engineering
Princeton University
Princeton, NJ 08544, USA
rigollet@princeton.edu

Assaf Zeevi
Graduate School of Business
Columbia University
New York, NY 10027
assaf@gsb.columbia.edu