A LINEAR RESPONSE BANDIT PROBLEM*

By Alexander Goldenshluger and Assaf Zeevi

University of Haifa and Columbia University

We consider a two–armed bandit problem which involves sequential sampling from two *non-homogeneous* populations. The response in each is determined by a random covariate vector and a vector of parameters whose values are not known a priori. The goal is to maximize cumulative expected reward. We study this problem in a minimax setting, and develop rate-optimal polices that combine myopic action based on least squares estimates with a suitable "forced sampling" strategy. It is shown that the regret grows logarithmically in the time horizon n and no policy can achieve a slower growth rate over all feasible problem instances. In this setting of linear response bandits, the identity of the sub-optimal action changes with the values of the covariate vector, and the optimal policy is subject to sampling from the inferior population at a rate that grows like $\sqrt{n}$.

1. Introduction. Sequential allocation problems, otherwise known as multi-armed bandit problems, arise frequently in various areas of statistics, adaptive control, marketing, economics and machine learning. The problem can be described as that of choosing between arms of a slot machine, where each time an arm is pulled a random reward which is arm-dependent is realized. The goal is to maximize the cumulative expected reward. Since the mean reward rate for each arm is not known, the gambler is faced with the classical dilemma between *exploration* and *exploitation*.

The first instance of these sequential allocation problems was introduced by Robbins (1952), and since then numerous variants thereof have been studied extensively in many different contexts; we refer to Berry and Fristedt (1985), Gittins (1989), Lai (2001) and the recent book by Cesa–Bianchi and Lugosi (2006), as well as references therein. A stream of such literature has focused on the characterization of optimal procedures under Bayesian formulations, but the complexity of the problem has led many researchers to seek approximate solutions that perform well in a suitable asymptotic sense; see, e.g., Lai and Robbins (1985), Lai (1987), Lai (1988), Lai and Yakowitz (1995), Auer et al. (2002a) among others.

*The work was supported in part by BSF Grant 2006075.

AMS 2000 subject classifications: Primary 62L05; secondary 60G40, 62C20

Keywords and phrases: sequential allocation, estimation, bandit problems, regret, minimax, rate–optimal policy

The prototypical bandit model assumes sequential sampling from two "homogeneous" populations, where the distribution of the realized random reward depends only on the chosen arm. However, in many practical situations some additional information can be utilized for allocation purposes. In particular, imagine that at each stage t additional side information, in the form of a random covariate X_t, is given, and the reward in each arm depends also on the value of this side observation. This model is a more accurate reflection of many instances treated in the bandit literature. Consider for example the problem of clinical trials which motivated several of the original bandit papers. In that setting, see Lai (1987), patients enter sequentially and receive one of, say, two possible treatments whose efficacy is yet to be determined. The objective is to allocate the "better" treatment to each patient. Viewing the patients as a *homogeneous population* ignores many important patient specifics (e.g., age, weight, blood pressure, comorbidities, etc), that can be used to improve treatment allocation. Such additional information can be encoded as a *covariate*, i.e., an auxiliary explanatory variable, that is observable just prior to the selection of an arm, and whose value can influence the mean response of each arm. Because the identity of the preferred arm depends on the revealed covariate value, the theory of sequential allocation problems with such side information is significantly different from the traditional setting where no covariates are present.

Although sequential allocation problems have been the subject of a voluminous literature, the study of bandit problems with side information (covariates) is in a far more nascent stage. The pioneering work of Woodroofe (1979) provides recursive expressions for the optimal Bayesian policy in a *one–armed* bandit problem. This setting involves one arm whose response is known a priori, and another arm whose response depends in a very specific manner on a single covariate; see also Woodroofe (1982) and Sarkar (1991). The one–armed allocation model was recently studied in a minimax setting by Goldenshluger and Zeevi (2009); we refer to that paper for further references and links to antecedent literature.

In the present paper we formulate and study a two–armed bandit problem which is a natural extension of the setups in Lai and Robbins (1985), Woodroofe (1979) and Goldenshluger and Zeevi (2009). At each stage t a covariate $X_t \in \mathbb{R}^d$ is observed, and the mean reward of each arm is given by a linear function of the covariate, where the parameters characterizing this response function are unknown. After the value of X_t is observed, one of the two arms is selected and one obtains a noisy observation of the mean reward for that particular value of X_t; this is the response variable. (One can view this setting as a linear regression model with X_t playing the role of a vec-

tor of explanatory variables.) The objective is to maximize the cumulative expected reward over a finite horizon of length n. We refer to this setup as the *linear response bandit problem*. In this setting we develop an allocation policy, study its performance, and derive a lower bound on the minimax regret for a natural class of joint distributions of covariates and rewards. We demonstrate that the proposed policy is optimal in terms of the dependence of its performance on the horizon n. That is, no other admissible policy can achieve a faster growth rate of cumulative expected rewards. These results characterize the complexity of the linear response bandit problem.

In antecedent literature there is some discussion of a univariate version of the linear response bandit problem by Gooley and Lattin (2000) in the context of dynamic customization of marketing messages. The most closely related papers to ours are Auer (2002), Mersereau et al. (2009) and Rusmevichientong and Tsitsiklis (2010); it will be useful to defer further discussion of connections between our work and the aforecited literature to Section 6, after our model and main results have been developed. We also refer the interested reader to Ginebra and Clayton (1995), Yang and Zhu (2002), Wang et al. (2005), Langford and Zhang (2008), and Lu et al. (2010), where other related models were considered.

The rest of the paper is organized as follows. In Section 2 we present the problem formulation. Section 3 describes the proposed allocation policy, and Section 4 contains our main results. In Section 5 we present results characterizing properties of the least squares estimators under the proposed policy; these are the key ingredients that are later used in the proof of the upper bound on the regret. Section 6 contains discussion and adds some concluding remarks. The proofs are given in Section 7. An appendix contains auxiliary results.

2. Problem formulation. Consider the following two–armed bandit problem. One observes a sequence $X_1, X_2, \ldots$ of independent random vectors with common distribution P_X. At each stage t, one can allocate the covariate vector $X_t \in \mathbb{R}^d$ to the i-th arm ($i = 1, 2$) of the bandit machine; this will be referred to as selecting the ith arm. Following that selection, one obtains the response $Y_t = Y_t^{(i)}$,

$$Y_t^{(i)} = \beta_i^T Z_t + \varepsilon_t^{(i)}, \quad i = 1, 2,$$

where $\beta_i \in \mathbb{R}^{d+1}$ are unknown parameters, $Z_t = (1, X_t) \in \mathbb{R}^{d+1}$ and $\varepsilon_t^{(i)}$ are iid normal random variables with zero mean and variance σ^2, independent of X_t. In that manner, at each stage t if the i-th arm is selected then the obtained reward is equal to $Y_t^{(i)}$. The goal is to maximize cumulative expected

rewards up to stage n. Here and in what follows, all random variables are assumed to be defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and $\mathbb{E}$ stands for the expectation operator with respect to $\mathbb{P}$.

An admissible *allocation rule (policy)* π is a sequence of random variables $\pi_1, \pi_2, \ldots$ taking values in the set $\{1, 2\}$ such that π_t is measurable with respect to the σ–field $\mathcal{F}_{t-1}^{+}$ generated by the previous observations and actions $\big\{\pi_s, X_s, Y_s^{(\pi_s)}, s = 1, \ldots, t-1\big\}$ and by the current covariate vector X_t,

$$\mathcal{F}_{t-1}^{+} = \sigma\big(\pi_1, X_1, Y_1^{(\pi_1)}, \ldots, \pi_{t-1}, X_{t-1}, Y_{t-1}^{(\pi_{t-1})}, X_t\big).$$

We also denote $\mathcal{F}_{t-1} = \sigma\big(\pi_1, X_1, Y_1^{(\pi_1)}, \ldots, \pi_{t-1}, X_{t-1}, Y_{t-1}^{(\pi_{t-1})}\big)$.

Let $\pi^* = (\pi_t^*, t \geq 1)$ be the *oracle rule* which "knows" the arm parameters and at each stage t prescribes

$$\pi_t^* = \operatorname*{argmax}_{i=1,2}\{\beta_i^T Z_t\}, \quad t = 1, 2, \ldots$$

Performance of an allocation rule π will be measured by the regret relative to the oracle performance:

$$R_n(\pi, \pi^*) = \mathbb{E}\sum_{t=1}^{n} Y_t^{(\pi_t^*)} - \mathbb{E}\sum_{t=1}^{n} Y_t^{(\pi_t)} = \mathbb{E}\sum_{t=1}^{n} |\beta_1^T Z_t - \beta_2^T Z_t| I\{\pi_t \neq \pi_t^*\}.$$

Let $\mathcal{P}$ be a class of pairs $(P_{X,Y^{(1)}}, P_{X,Y^{(2)}})$ of joint distributions of $(X, Y^{(1)})$ and $(X, Y^{(2)})$. The maximal regret of a rule π is defined as

$$R_n(\pi; \mathcal{P}) := \sup\Big\{R_n(\pi, \pi^*) : (P_{X,Y^{(1)}}, P_{X,Y^{(2)}}) \in \mathcal{P}\Big\},$$

and the objective is to develop a *rate–optimal rule* $\hat{\pi}$ such that

$$R_n(\hat{\pi}; \mathcal{P}) \leq C \inf_{\pi} R_n(\pi, \mathcal{P}) \tag{1}$$

for some positive constant C that does not depend on n. The infimum on the right hand side is taken over all admissible policies π.

3. The proposed policy. Now we are in a position to define the proposed policy. The policy combines *forced sampling* from both arms at prescribed time instances with *myopic action* between said times; the terms will become clear in what follows. We begin with a description of the sequences of time instants when the policy will perform forced sampling from both arms.

3.1. *Forced sampling sequences.* Fix a real number $q > 0$, which will serve as a design parameter to be specified. Define the sequence $\mathcal{T} = (\tau_s, s \geq 1)$ of positive integers by

$$\tau_1 = 1, \quad \tau_s = \lfloor \exp\{qs\} \rfloor, \quad s = 2, 3, \ldots$$

For a given natural number t, the number of elements $N(t)$ of the sequence $\mathcal{T}$ that are less than or equal to t satisfies the following inequalities:

$$\frac{1}{q} \ln t - 1 \leq N(t) \leq \frac{1}{q} \ln(t+1). \tag{2}$$

It is easily verified that if

$$s > \nu := 1 + \left\lceil \frac{1}{q} \ln_+ \left(\frac{2}{e^q - 1} \right) \right\rceil, \tag{3}$$

then $\tau_s - \tau_{s-1} > 1$. This fact implies that the subsequences

$$\mathcal{T}_1 := (\tau_s,\, s \geq \nu + 1), \quad \mathcal{T}_2 := \mathcal{T}_1 + 1 = (\tau_s + 1,\, s \geq \nu + 1)$$

are disjoint.

Now we are in a position to define the forced sampling sequences. Let

$$\nu_0 := \nu \vee (d+1), \tag{4}$$

and define

$$\begin{aligned} \mathcal{T}_{1,t} &:= \{s \in \mathcal{T}_1 : 2\nu_0 + 1 \leq s \leq t\} \cup \{s : s = 2j - 1,\ j = 1, \ldots, \nu_0\}, \\ \mathcal{T}_{2,t} &:= \{s \in \mathcal{T}_2 : 2\nu_0 + 1 \leq s \leq t\} \cup \{s : s = 2j,\ j = 1, \ldots, \nu_0\}. \end{aligned} \tag{5}$$

Here $\mathcal{T}_{i,t}$ denotes the set of time instances up until stage t when the policy performs forced sampling from the ith arm. We set also $T_i(t) = \#\{\mathcal{T}_{i,t}\}$, where from now on $\#\{\cdot\}$ stands for the cardinality of a set.

Note that by construction, for all $t \geq 2\nu_0 + 1$ one has $T_1(t) = N(t) - \nu + \nu_0$, while $T_2(t) = N(t) - \nu + \nu_0$ or $T_2(t) = N(t) - \nu + \nu_0 - 1$. Thus we always have $N(t) - \nu + \nu_0 - 1 \leq T_i(t) \leq N(t) - \nu + \nu_0$ which, in view of (2), yields

$$\frac{1}{q} \ln t - \nu + \nu_0 - 2 \ \leq\ T_i(t) \ \leq\ \frac{1}{q} \ln(t+1) + \nu_0 - \nu, \quad \forall t \geq 2\nu_0 + 1. \tag{6}$$

3.2. *Policy description.* In what follows we denote our proposed policy by $\hat{\pi} = (\hat{\pi}_t, t \geq 1)$. For a subset $\mathcal{J}$ of $\{1, 2, \ldots\}$ we define

$$\hat{b}(\mathcal{J}) := \hat{Q}^{-1}(\mathcal{J})\Big[\frac{1}{\#\{\mathcal{J}\}}\sum_{s\in\mathcal{J}} Z_s Y_s^{(\hat{\pi}_s)}\Big], \quad \hat{Q}(\mathcal{J}) := \frac{1}{\#\{\mathcal{J}\}}\sum_{s\in\mathcal{J}} Z_s Z_s^T. \tag{7}$$

Thus, given a set $\mathcal{J}$ of indices, $\hat{b}(\mathcal{J})$ denotes the least squares (LS) estimator of the parameter vector β in the linear regression model based on the observations $\{Z_s, Y_s^{(\hat{\pi}_s)}, s \in \mathcal{J}\}$. We will write Y_s and ε_s instead of $Y_s^{(\hat{\pi}_s)}$ and $\varepsilon_s^{(\hat{\pi}_s)}$ respectively when the value of $\hat{\pi}_s$ is clear from the context.

Let

$$\mathcal{S}_{i,t} = \{1 \leq s \leq t : \hat{\pi}_s = i\}, \quad \text{and} \quad S_i(t) := \#\{\mathcal{S}_{i,t}\}, \quad i = 1, 2, \tag{8}$$

be the set of indices for which the strategy $\hat{\pi}$ pulls arm $i = 1, 2$, and its cardinality, respectively. For this index set, and for $\mathcal{T}_{i,t}$ given in (5), we define the following estimators via (7):

$$\begin{aligned} \hat{\beta}_i(t) &:= \hat{b}(\mathcal{S}_{i,t}), \quad \hat{Q}_i(t) := \hat{Q}(\mathcal{S}_{i,t}) \\ \tilde{\beta}_i(t) &:= \hat{b}(\mathcal{T}_{i,t}), \quad \tilde{Q}_i(t) := \hat{Q}(\mathcal{T}_{i,t}). \end{aligned} \tag{9}$$

Thus we have two sets of the LS estimators: the estimators $\hat{\beta}_i(t)$ are based on the entire set of observations sampled from arm $i = 1, 2$ up until time t, while $\tilde{\beta}_i(t)$ are estimators based only on the observations collected through the forced sampling subsequence.

The proposed policy $\hat{\pi} = (\hat{\pi}_t, t \geq 1)$ uses these two sets of the LS estimators and requires two design parameters: a *forced sampling parameter* q; and a *localization parameter* h. These parameters will be specified later; see (15) in Theorem 1. The policy is given by the following algorithm.

Algorithm 1

1. Initialization: pull each arm ν_0 times, i.e., set

$$\hat{\pi}_{2j-1} = 1, \text{ and } \hat{\pi}_{2j} = 2, \quad j = 1, \ldots, \nu_0.$$

2. Set $t = 2\nu_0$.
3. If $t + 1 \in \mathcal{T}_i$

   then set $\hat{\pi}_{t+1} = i$ (pull arm i) and compute the estimates $\hat{\beta}_i(t+1), \tilde{\beta}_i(t+1)$, $i = 1, 2$;

   else

if $|(\tilde{\beta}_1(t) - \tilde{\beta}_2(t))^T Z_{t+1}| > h/2$ then set

$$\hat{\pi}_{t+1} = \arg\max_{i=1,2} \left\{ \tilde{\beta}_i^T(t) Z_{t+1} \right\};$$

if $|(\tilde{\beta}_1(t) - \tilde{\beta}_2(t))^T Z_{t+1}| \leq h/2$ then set

$$\hat{\pi}_{t+1} = \arg\max_{i=1,2} \left\{ \hat{\beta}_i^T(t) Z_{t+1} \right\};$$

and compute the estimates $\hat{\beta}_i(t+1)$, $i = 1, 2$.

4. $t \leftarrow t+1$ and repeat from step (iii).

If the time instant $t+1$ belongs to a forced sampling subsequence $\mathcal{T}_i$, then the subsequent observation is taken from the ith arm (i.e. $\hat{\pi}_{t+1}$ is set to i), and all estimates are updated. Otherwise, two different actions are possible. If Z_{t+1} falls outside the $h/2$–margin of the set $\{z : \tilde{\beta}_1^T(t)z = \tilde{\beta}_2^T(t)z\}$, then the myopic decision is made on the basis of the *forced sampling LS estimates*. If Z_{t+1} belongs to the $h/2$–margin of the set $\{z : \tilde{\beta}_1^T(t)z = \tilde{\beta}_2^T(t)z\}$, then $\hat{\pi}$ performs myopic action using the LS estimates based on the *entire set of observations*.

We note that the accuracy of the forced sampling LS estimates is easily controlled; this fact facilitates analysis of the algorithm performance. In particular, intuitively speaking, the forced sampling estimators are used to ensure that statistical information is gathered in a suitable manner from *both* arms, and preclude the myopic action from "under-sampling" any one of them. In this context, when myopic action is performed, the localization parameter h is used to distinguish whether the present covariate falls "close" to the decision boundary; if it does, then the "full information" LS estimates are used (these are the estimates computed based on forced sampling *and* observations collected via myopic actions); if it does not, the LS estimates based exclusively on the forced sampling instances are used.

4. Main results.

4.1. *Classes of joint distributions.* Since

$$Y^{(i)}|X = x \sim \mathcal{N}(\beta_i^T z, \sigma^2), \quad z = (1, x), \quad i = 1, 2, \tag{10}$$

the classes of pairs of joint distributions of $(X, Y^{(1)}), (X, Y^{(2)})$ can be defined by specification of parameter sets for (β_1, β_2) and a class of distributions P_X.

Definition 1 *We say that the pair $(P_{X,Y^{(1)}}, P_{X,Y^{(2)}})$ belongs to the class $\mathcal{P}$ if (10) holds and the following conditions are satisfied:*

(A1) $X_t = (X_{t,1}, \dots, X_{t,d})$, $t = 1, 2, \dots$ *are independent identically distributed random vectors having density with respect to Lebesgue measure on* $\mathbb{R}^d$, *and* $\max_{j=1,\dots,d} |X_{t,j}| \le r$, *for all* t. *Let* $\mu := \mathbb{E}X_t$, $V := \mathbb{E}X_t X_t^T$, *and*

$$Q = \begin{bmatrix} 1 & \mu^T \\ \mu & V \end{bmatrix} \in \mathbb{R}^{(d+1)\times(d+1)}. \tag{11}$$

There exists a positive real number $\underline{\lambda}$ *such that* $\lambda_{\min}\{Q\} \ge \underline{\lambda} > 0$.

(A2) (Margin condition). *There exist positive real numbers* ρ_0, L *such that*

$$\mathbb{P}\big\{|(\beta_1 - \beta_2)^T Z_t| \le \rho\big\} \le L\rho, \quad \forall \rho \in (0, \rho_0].$$

(A3) (Diversity outside margin). *There exists a real number* $\lambda_* > 0$ *such that*

$$\lambda_{\min}\big\{\mathbb{E}U_s^+ Z_s Z_s^T\big\} \wedge \lambda_{\min}\big\{\mathbb{E}U_s^- Z_s Z_s^T\big\} \ge \lambda_* > 0, \tag{12}$$

where $U_s^+ := I\big\{(\beta_1 - \beta_2)^T Z_s \ge \rho_0\big\}$ *and* $U_s^- := I\big\{(\beta_1 - \beta_2)^T Z_s < -\rho_0\big\}$.

(A4) (Parameter set). *The arm parameters* (β_1, β_2) *satisfy*

$$\|\beta_i\| \le b, \quad i = 1, 2.$$

Discussion of the assumptions. Assumption (A1) states that X_t is a random vector with non–degenerate distribution over $\mathbb{R}^d$, which ensures that the arm parameters β_1 and β_2 are identifiable. Moreover, because this distribution has a density with respect to Lebesgue measure on $\mathbb{R}^d$, it follows that after the initialization step in Algorithm 1, the LS estimators are well–defined, i.e., the corresponding matrices $\hat{Q}_i(\cdot)$ and $\tilde{Q}_i(\cdot)$ are non–singular with probability one.

Assumption (A2) is related to the behavior of the distribution of X_t "near" the decision boundary $\{x : \beta_1^T x = \beta_2^T x\}$; it is akin to the so-called *margin condition* that is widely used in the classification literature [see, e.g., Tsybakov (2004)]. Roughly speaking, the region near the decision boundary is where it is most "difficult" to distinguish between the two arms. At the same time, the contribution to the regret of making incorrect decisions is affected by the probability of covariates falling into that region. The margin condition therefore plays a central role in determining the complexity of the linear response bandit problem, much like its role in classification problems.

Assumption (A3) implies "persistence of excitation," in the language of automatic control. Under this condition it is possible to improve the accuracy of estimates of the arm parameters without necessitating errors in the

selection of the arms; for further discussion see Goldenshluger and Zeevi (2011). Assumption (A3) also implies that

$$p_* := \min\Big[\mathbb{P}\big\{(\beta_1-\beta_2)^T Z_t \ge \rho_0\big\},\ \mathbb{P}\big\{(\beta_1-\beta_2)^T Z_t < -\rho_0\big\}\Big] > 0. \tag{13}$$

That is, there is probability mass away from the region in which the margin condition holds. This implies that the distribution P_X is such that with positive probability the random variable $(\beta_1-\beta_2)^T Z_t$ can take positive and negative values, and the preference of arms is changed depending on the sign of $(\beta_1-\beta_2)^T Z_t$.

4.2. *Bounds on the regret.* The next result establishes an upper bound on the maximal regret of the policy defined in Algorithm 1.

Let t_* denote the minimal positive integer satisfying

$$t_* \ge (8\nu_0+1) \ \vee\ 4096\sigma^2(1\vee r^2)d(1+rd)^2\rho_0^{-2}\lambda_*^{-2}, \tag{14}$$

where ν_0 is given in (4), and ρ_0 and λ_* appear in Assumptions (A2) and (A3).

Theorem 1 *Let $\hat\pi = (\hat\pi_t, t\ge 1)$ be the policy defined by Algorithm 1 and associated with parameters h and q. If $\mathcal{P}$ is a class of pairs $(P_{X,Y^{(1)}}, P_{X,Y^{(2)}})$ of joint distributions given in Definition 1, and*

$$\rho_0 \ge h,\ \ \min\Big\{\frac{h^2\underline{\lambda}}{192\sigma^2(1+r^2d)}\ ,\ \frac{\underline{\lambda}^2}{24(d+1)^2(r^2\vee r^4)}\Big\} \ge q \tag{15}$$

then for every $n > t_*$

$$R_n(\hat\pi;\mathcal{P}) \le \frac{4b}{q}\sqrt{\lambda_{\max}(Q)}\,\ln(n+1) + C_1 L\sigma^2 d^2(1+r^2d)\lambda_*^{-2}(1\vee r)^2\ln n + C_2, \tag{16}$$

where C_1 is an absolute constant, and C_2 is a constant that can depend on parameters of the class $\mathcal{P}$, σ^2,d, r and b only.

The bound shows that the maximal regret of $\hat\pi$ grows at most logarithmically, i.e., $R_n(\hat\pi;\mathcal{P}) = O(\ln n)$ as $n\to\infty$. As the proof of the theorem indicates, the first term on the right hand side of (16) describes the contribution of the forced sampling scheme to the regret, while the second term is due to the "difficulty" of resolving the correct allocation when the covariate X_t falls "near" the decision boundary. It is also worth noting that the proposed policy does not require a priori knowledge of the length of the horizon n.

Another characteristic of the policy is the *maximal inferior sampling rate*, defined as the worst–case expected cumulative number of inferior arm selections:

$$\begin{aligned} M_n(\hat{\pi};\mathcal{P}) &:= \sup\Big\{M_n(\hat{\pi},\pi^*) : (P_{X,Y^{(1)}}, P_{X,Y^{(2)}}) \in \mathcal{P}\Big\} \\ M_n(\hat{\pi},\pi^*) &:= \mathbb{E}\sum_{t=1}^{n} I\{\pi_t \neq \pi_t^*\}, \end{aligned}$$

where $\pi^* = (\pi_t^*, t \geq 1)$ is the oracle rule. The inferior sampling rate of the policy $\hat{\pi}$ is bounded as follows

$$M_n(\hat{\pi};\mathcal{P}) \;\leq\; \frac{1}{q}\ln(n+1) \;+\; C_3 L d^{3/2}\lambda_*^{-1}\sigma\sqrt{1+r^2 d}\,(1\vee r)\sqrt{n} \;+\; C_4, \tag{17}$$

i.e., $M_n(\hat{\pi};\mathcal{P}) = O(\sqrt{n})$ as $n \to \infty$. In the end of the proof of Theorem 1 we present the arguments leading to this upper bound.

A natural question is whether there is a policy with maximal (over $\mathcal{P}$) regret that grows slower than $O(\ln n)$ as $n \to \infty$. We now show that the magnitude of the regret established in Theorem 1 is essentially best possible.

Theorem 2 *Let $\mathcal{P} = (P_{X,Y^{(1)}}, P_{X,Y^{(2)}})$ be the class of pairs of distributions defined in Definition 1. Then for any admissible policy $\tilde{\pi}$ and for n sufficiently large*

$$R_n(\tilde{\pi};\mathcal{P}) \geq C_3\sigma^2 \ln n,$$

where C_3 depends on the parameters of the class $\mathcal{P}$.

This theorem, along with Theorem 1, shows that the proposed policy $\hat{\pi}$ is rate–optimal in the sense of (1).

4.3. *Numerical illustration.* In this section we illustrate performance of the proposed policy and its dependence on the dimensionality of the covariate vector X_t. In our simulations the covariate vectors X_t were chosen to be uniformly distributed on $[-1,1]^d$ with $d \in \{2,4,6,8,10,12\}$. In each run the arm parameters $\beta_1, \beta_2 \in \mathbb{R}^{d+1}$ are chosen as follows: $\beta_1 = 0$ and $\beta_2 = \frac{1}{\sqrt{d}}(0,\delta_1,\ldots,\delta_d)$ where δ_i are independent random signs (Rademacher random variables). The policy is implemented with parameters $h = 1$ and $q = \min\{0.1, d^{-2}\}$.

In each run we compute the regret of our policy for the horizon $n = 250\times k$ where $k \in \{1,2,3,4,6,12,20,30\}$ and average the results over 50 runs. Figure 1 depicts the graphs of average regret against the logarithm of the horizon for $d \in \{2,6,12\}$. As predicted by the results of Theorem 1, the graphs indicate an approximately linear growth of the regret with $\ln n$. It is also

clearly seen that the policy performance deteriorates with the dimensionality d. Based on the regret samples obtained in 50 runs we estimated the

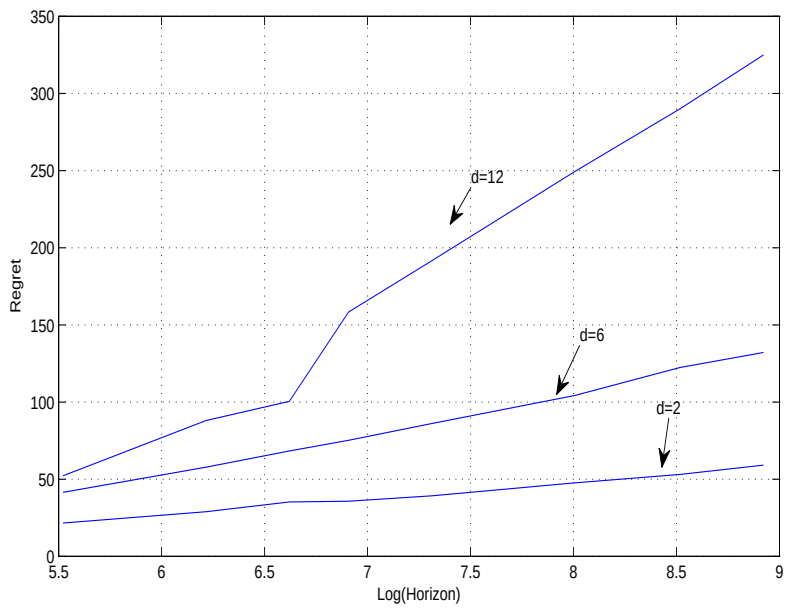

FIG 1. *Regret of the proposed policy averaged over 50 runs against the logarithm of the horizon* $\ln n$ *for dimensions* $d \in \{2, 6, 12\}$.

slopes using the least squares method. The slope estimates along with the standard deviations are given in Table 1. These results indicate that the upper bound of Theorem 1 is conservative in terms of its dependence on the dimensionality d.

dimensionality d	2	4	6	8	10	12
slope estimate	10.659	17.715	27.021	42.467	61.134	84.636
standard deviation	0.380	0.562	0.594	1.012	3.370	4.522

TABLE 1
The slope estimates for $\ln n$ *dependence of the regret.*

5. Statistical properties of the LS estimators under the proposed policy. The key to establishing upper bounds on the regret in Theorem 1 is an analysis of statistical properties of the least squares estimators

$\tilde{\beta}_i(t)$ and $\hat{\beta}_i(t)$, $i = 1, 2$ defined in (9). In this section we present several results to that effect and briefly discuss their implications.

Our results will be in the form of various probability bounds, and for that purpose it will be useful to define the following quantity that will be used repeatedly. For $\varkappa, s, u > 0$, put

$$P(\varkappa, s, u) \;:=\; \sqrt{\frac{2}{\pi}}(1 + r^2 d)^{1/2} \frac{\sqrt{s}}{\varkappa\sigma} \exp\Big\{ - \frac{\varkappa^2 \underline{\lambda}\, s}{4\sigma^2 u^2} \Big\} + 2d(d+3) \exp\Big\{ - \frac{\underline{\lambda}^2 s}{8(d+1)^2 (r^2 \vee r^4)} \Big\}. \tag{18}$$

The next statement establishes a probability bound related to deviations of the LS estimators along the forced sampling subsequence, $\tilde{\beta}_i(t)$, from the true parameter values.

Proposition 1 *Let Assumption (A1) hold; then for any $\varkappa > 0$, and any $t \geq 2\nu_0 + 1$*

$$\mathbb{P}\Big\{ \big|(\tilde{\beta}_i(t) - \beta_i)^T Z_{t+1}\big| \geq \varkappa \Big\} \;\leq\; P\big(\varkappa, T_i(t), \sqrt{1 + r^2 d}\big), \quad i = 1, 2.$$

Proposition 1 and the lower bounds on $T_i(t)$, $i = 1, 2$ in (6) imply that the probability $\mathbb{P}\big\{ \big|(\tilde{\beta}_i(t) - \beta_i)^T Z_{t+1}\big| \geq \varkappa \big\}$ can be made negligible by choice of the forced sampling parameter q. It turns out that this fact, together with Assumption (A3) that characterizes the diversity of X_t's outside the margin, ensures that under the proposed policy each arm is pulled a linearly increasing number of times with high probability. Moreover, under the proposed policy the information about the unknown parameters increases at a suitable rate. These facts are central for the proof of Theorem 1; they are subject of the next statement for which the following notation will be useful.

For $\mathcal{J} \subseteq \{1, 2, \ldots\}$ let $\Sigma(\mathcal{J}) := \sum_{s \in \mathcal{J}} Z_s Z_s^T$. Recall also that $\mathcal{S}_{i,t} = \{1 \leq s \leq t : \hat{\pi}_s = i\}$, and $S_i(t) = \#\{\mathcal{S}_{i,t}\}$, $i = 1, 2$.

Proposition 2 *Let Assumptions (A1) and (A3) hold, and let*

$$P_t := (d+1) \sum_{s=\lfloor t/2 \rfloor}^{t} \sum_{i=1}^{2} P\Big(\frac{\rho_0}{4(1 + rd)}, T_i(s-1), 1 \Big),$$

where $P(\cdot, \cdot, \cdot)$ is defined in (18). If $h \leq \rho_0$ then for all $t \geq 8\nu_0 + 1$ and

$i = 1, 2$

$$\mathbb{P}\Big\{S_i(t) \le \frac{1}{8}p_* t\Big\} \le P_t + \exp\Big\{-\frac{p_*^2 t}{128}\Big\}, \tag{19}$$

(20)

$$\mathbb{P}\Big\{\lambda_{\min}\big(\Sigma(\mathcal{S}_{i,t})\big) \;\le\; \frac{\lambda_* t}{8}\Big\} \le P_t + (d+1)(d+2)\exp\Big\{-\frac{\lambda_*^2 t}{128(1\vee r^2)^2(1+d)^2}\Big\},$$

where p_* and λ_* are given in (13) and in (12) respectively.

A couple of remarks on the statement of Proposition 2 are in order. First, it follows from (18) and (6) that if for any $k > 0$ we set $q \le C_1 k^{-1}$, where $C_1 = C_1(\rho_0, \underline{\lambda}, d, r)$ is some explicit constant, then $P_t \le \sqrt{\ln t}/t^k$ for all $t \ge 8\nu_0 + 1$. Under this choice of the forced sampling parameter q, with high probability the proposed policy pulls each arm a linearly growing number of times. Second, inequality (20) shows that with high probability the minimal eigenvalue of the matrix $\Sigma(\mathcal{S}_{i,t})$ grows linearly. This, in turn, implies that the accuracy of the estimates $\hat{\beta}_i(t)$ improves in a suitable manner when sampling according to the proposed policy. This last fact is quantified in the next proposition.

Proposition 3 *Let Assumptions (A1) and (A3) hold, and let $h \le \rho_0$; then for all $t \ge 2\nu_0 + 1$ and $i = 1, 2$ one has*

$$\begin{aligned}
&\mathbb{P}\Big\{\|\hat{\beta}_i(t) - \beta_i\| \ge \varkappa,\;\; \lambda_{\min}\big(\Sigma(\mathcal{S}_{i,t})\big) \ge \frac{\lambda_* t}{8}\Big\} \\
&\qquad\qquad \le\; 2\exp\Big\{-\frac{\varkappa^2\lambda_*^2 t}{256\sigma^2}\Big\} + 2d\exp\Big\{-\frac{\varkappa^2\lambda_*^2 t}{256\sigma^2 d r^2}\Big\}, \\
&\mathbb{P}\Big\{|(\hat{\beta}_i(t) - \beta_i)^T Z_t| \ge \varkappa,\;\; \lambda_{\min}\big(\Sigma(\mathcal{S}_{i,t})\big) \ge \frac{\lambda_* t}{8}\Big\} \\
&\qquad\qquad \le\; 2\exp\Big\{-\frac{\varkappa^2\lambda_*^2 t}{256\sigma^2(1+r^2 d)}\Big\} \\
&\qquad\qquad\quad + 2d\exp\Big\{-\frac{\varkappa^2\lambda_*^2 t}{256\sigma^2 d(1+r^2 d)r^2}\Big\}.
\end{aligned}$$

Proposition 3 establishes bounds on the deviations of the LS estimates $\hat{\beta}_i(t)$ from β_i, $i = 1, 2$, on the event that the minimal eigenvalue of the matrix $\Sigma(\mathcal{S}_{i,t})$ grows linearly in t. As stated above in Proposition 2, this event occurs with high probability under an appropriate choice of the force sampling parameter q.

6. Discussion and concluding remarks.

1. Extensions. As shown above, under suitable conditions on the class of joint distributions of covariates and responses, the minimax regret in the linear response bandit problem scales like $\ln n$ with the horizon length n. These results can be extended in different directions.

The upper bound of Theorem 1 remains intact if condition (A2) in the definition of the class $\mathcal{P}$ is replaced by the following more general margin condition: there exist positive real numbers ρ_0, L and $\alpha \geq 1$ such that $\mathbb{P}\{|(\beta_1 - \beta_2)^T Z_t| \leq \rho\} \leq L\rho^\alpha$, $\forall \rho \in (0, \rho_0]$. Note, however, that because the covariate X_t is assumed to have a continuous distribution, the case $\alpha = 1$ is the most typical. When $\alpha > 1$, the policy proposed in this paper is no longer rate–optimal; see Goldenshluger and Zeevi (2009) that treats a particular one-armed setting from which the above conclusion can be drawn.

Although we considered the two–armed version of the linear response bandit problem, our results can be extended to the case of K arms. Indeed, one expects that under a suitable margin condition, of the type given in (A2) ensuring proper separation of the best arm from other arms for any given value of the covariate vector, a maximal regret of the order $\ln n$ is achievable. It is also worth noting that the results of this paper remain intact if instead of the Gaussian assumption, the errors $\varepsilon_t^{(i)}$, $i = 1, 2$ are assumed to have sub–Gaussian tails.

2. Relation to existing literature. The linear response model was previously studied by Auer (2002); however, his assumptions, analysis and results all differ from ours. In particular, Auer (2002) considers the K–armed bandit problem with bounded linear response, when the parameter vector is *common to all arms* (for the two–armed problem, $\beta_1 = \beta_2$ in our notation). In his model, the covariates are assumed to be arbitrary, deterministic and *arm–dependent*, i.e., for each arm there is an arm–specific sequence of the corresponding covariates. In this setup, Auer (2002) develops a policy and shows that the worst–case regret (with respect to the *covariates*) is bounded from above by $O(\sqrt{n})$. In this context it is also worth noting that even in the traditional bandit problem, the minimax regret has the order $\sqrt{n}$ under so-called adversarial and "gap–free" scenarios [see, e.g., Auer et al. (2002b) and Juditsky et al. (2008)].

The work by Mersereau et al. (2009) and Rusmevichientong and Tsitsiklis (2010) considers the problem of sequential constrained linear optimization based on noisy observations of function values. In particular, in their set up, the mean reward is given by an underlying linear function, whose parameters are not known a priori. At each stage, the decision maker needs to choose an argument value of that function, with the objective of maximizing the

cumulative mean rewards. In that setting, Rusmevichientong and Tsitsiklis (2010) propose a rate–optimal policy and demonstrate that the minimax regret is of the order $\sqrt{n}$ for a broad family of constraint sets.

In contrast to these results, we consider the case when the parameter vectors are arm–dependent, and the covariates are independent identically distributed vectors. To further clarify, what we term as a "covariate" is exogenously given, while in Mersereau et al. (2009) and Rusmevichientong and Tsitsiklis (2010) this is the decision variable. Thus our model is a more natural extension of the traditional line of research on bandit problems, and in particular directly extends the clinical trial model considered by Lai and Robbins (1985).

3. Regret, inferior sampling rates, and traditional bandit results. The order of maximal regret achieved by our policy exhibits a growth rate of $O(\ln n)$ that is identical to the best achievable asymptotic results in the traditional bandit literature; cf. Lai and Robbins (1985). The reasoning for this is markedly different in the linear response bandit problem. First note that in the traditional bandit problem, the regret is proportional to the inferior sampling rate. Hence the logarithmic growth is essentially driven by controlling the error probability in a sequential hypothesis testing problem. In contrast, in the linear response bandit problem the errors are essentially due to the difficulty of allocating covariates X_t in proximity of the decision boundary. In particular, the maximal inferior sampling rate here is of order $\sqrt{n}$, which means that the frequency at which one makes sub-optimal arm selections is quite high relative to the traditional bandit problem. In essence, when the values of X_t are in the vicinity of the decision boundary $\{z : \beta_1^T z = \beta_2^T z\}$, it is impossible to distinguish the two arms at stage t to any finer resolution than order $1/\sqrt{t}$. This is the determining factor behind the $\sqrt{n}$-growth in the inferior sampling rate. At the same time, the regret is also determined by the difference in the mean reward rate, which is smaller near the margin. Hence, eventually the large number of inferior arm selections does not translate into a large growth rate in the regret. One other comment pertains to the lower bounds which establishes the rate-optimality of our proposed policy. Unlike the change-of-measure argument due to Lai and Robbins (1985), which is ubiquitous in asymptotic optimality results found in the traditional bandit literature, our lower bound argument builds on a Bayesian analogue of a Cramér-Rao type inequality to bound the mean squared estimation error from below at each stage t.

4. The case of "fully separated" responses. We have focused on classes of joint distributions of covariates and rewards such that the preference of arms changes for different values of the covariate vector X_t. If the arm parameters

are well–separated and preference of the arms does not change for all values of X_t, and if the distribution of X_t is continuous, then the minimax regret is also of the order $\ln n$ as $n \to \infty$. This rate is achieved by the policy with forced sampling and myopic action, just on the basis of "forced sampling" least squares estimates $\tilde{\beta}_i(t)$, $i = 1, 2$ [see (9)]. We note, however, that under these conditions both the regret and the inferior sampling rate are of the order $\ln n$ which is similar to the setup considered in Lai and Robbins (1985).

7. Proofs. As stated in the discussion of the assumptions in section 4.1, under Assumption (A1) the initialization step of the proposed policy results in well defined LS estimates (i.e., the corresponding LS matrices are invertible) with probability one. This fact is used in all the proofs below without further explicit mentioning.

7.1. *Proof of Proposition 1.* Because the estimates $\tilde{\beta}_i(t)$ are based on the forced sampling observations only we have

$$\tilde{\beta}_i(t) - \beta_i \mid \{X_s : s \in \mathcal{T}_{i,t}\} \sim \mathcal{N}_{d+1}\Big(0, \frac{\sigma^2}{T_i(t)}\tilde{Q}_i^{-1}(t)\Big), \quad i = 1, 2,$$

where $\tilde{Q}_i(t)$ is given in (9). Therefore, for any vector $a \in \mathbb{R}^{d+1}$ and any real number $\varkappa > 0$

$$\begin{aligned} &\mathbb{P}\Big\{(\tilde{\beta}_i(t) - \beta_i)^T a \geq \varkappa \,\Big|\, \{X_s : s \in \mathcal{T}_{i,t}\}\Big\} \\ &\qquad \leq \frac{\sqrt{T_i(t)}}{\sqrt{2\pi}\varkappa\sigma[a^T\tilde{Q}_i^{-1}(t)a]^{1/2}} \exp\Big\{-\frac{\varkappa^2 T_i(t)}{2\sigma^2 a^T\tilde{Q}_i^{-1}(t)a}\Big\}. \end{aligned} \tag{21}$$

We need the following bounds on the eigenvalues of the matrices $\tilde{Q}_i^{-1}(t)$. Because the maximal eigenvalue of a symmetric matrix is less than the maximum of the sums of absolute values of its row elements, it follows from boundedness of the components of X_t that $\lambda_{\max}\{\tilde{Q}_i(t)\} \leq 1 + r^2 d$. Hence

$$\lambda_{\min}\{\tilde{Q}_i^{-1}(t)\} \geq \frac{1}{1 + r^2 d}. \tag{22}$$

Furthermore, we can write $\tilde{Q}_i(t) = Q + E_i(t)$, $i = 1, 2$, where Q is given in (11), and $E_i(t)$ is a symmetric matrix whose elements can be controlled using Lemma 1 [see Appendix]. In particular, letting $\Gamma_* = \Gamma(\varkappa_*, \mathcal{T}_{i,t})$ in (38) (see Lemma 1 in the Appendix), $\varkappa_* := [2(d+1)]^{-1}\underline{\lambda}$ with $\underline{\lambda}$ given in Assumption (A1) we have

$$\max_{k,l=1,\ldots,d+1} \big|[E_i(t)]_{kl}\big| I\{\Gamma_*\} \leq \varkappa_* = \frac{\underline{\lambda}}{2(d+1)}.$$

Here $[E_i(t)]_{kl}$ denotes the k,l entry in the matrix $E_i(t)$. Hence on the event Γ_*, all eigenvalues of $E_i(t)$ belong to the interval $[-\underline{\lambda}/2, \underline{\lambda}/2]$. The well–known result on perturbation of eigenvalues of $d \times d$ symmetric matrices states that if A and E are symmetric matrices, then $\lambda_j(A+E) \in [\lambda_j(A) + \lambda_{\min}(E), \lambda_j(A) + \lambda_{\max}(E)]$, for $j = 1, \ldots, d$ [see, e.g., (Stewart and Sun 1990, Corollary 4.9)]. Using this result we have that on the event Γ_*

$$\lambda_{\min}\{\tilde{Q}_i(t)\} \geq \frac{1}{2}\underline{\lambda} \quad \Rightarrow \quad \lambda_{\max}\{\tilde{Q}_i^{-1}(t)\} \leq \frac{2}{\underline{\lambda}}. \tag{23}$$

Using (22) and (23) we obtain from (21)

$$\begin{aligned} \mathbb{P}\Big\{ \big[(\tilde{\beta}_i(t) - \beta_i)^T a \geq \varkappa\big] \cap \Gamma_* \quad \Big| \quad \{X_s : s \in \mathcal{T}_{i,t}\} \Big\} & \\ \leq \quad \frac{\sqrt{(1+r^2 d) T_i(t)}}{\sqrt{2\pi}\varkappa\sigma\|a\|} \exp\Big\{ -\frac{\varkappa^2 \underline{\lambda} T_i(t)}{4\sigma^2\|a\|^2} \Big\}. & \end{aligned} \tag{24}$$

Here and in what follows $\|\cdot\|$ stands for the Euclidean norm of a vector. In addition, by Lemma 1

$$\begin{aligned} \mathbb{P}\Big\{ \big[(\tilde{\beta}_i(t) - \beta_i)^T a \geq \varkappa\big] \cap \Gamma_*^c \Big\} \quad &\leq \quad \mathbb{P}\{\Gamma_*^c\} \\ &\leq \quad d(d+3) \exp\Big\{ -\frac{\underline{\lambda}^2 T_i(t)}{8(d+1)^2 (r^2 \vee r^4)} \Big\}. \end{aligned} \tag{25}$$

Combining these inequalities we obtain for any $a \in \mathbb{R}^{d+1}$ that

$$\begin{aligned} \mathbb{P}\Big\{ (\tilde{\beta}_i(t) - \beta_i)^T a \geq \varkappa \Big\} \quad &\leq \quad \frac{\sqrt{(1+r^2 d) T_i(t)}}{\sqrt{2\pi}\varkappa\sigma\|a\|} \exp\Big\{ -\frac{\varkappa^2 \underline{\lambda} T_i(t)}{4\sigma^2\|a\|^2} \Big\} \\ &+ \quad d(d+3) \exp\Big\{ -\frac{\underline{\lambda}^2 T_i(t)}{8(d+1)^2 (r^2 \vee r^4)} \Big\}. \end{aligned} \tag{26}$$

In order to complete the proof of the proposition, we substitute Z_{t+1} for a in (24) while conditioning on X_{t+1}, use the evident bounds $\|Z_{t+1}\| \geq 1$ and $\|Z_{t+1}\| \leq \sqrt{1+r^2 d}$, and combine the obtained result with (25). ∎

7.2. *Proof of Proposition 2.* We will prove the proposition for $\mathcal{S}_{1,t}$ and $\Sigma(\mathcal{S}_{1,t})$. The proof for $\mathcal{S}_{2,t}$ and $\Sigma(\mathcal{S}_{2,t})$ is completely analogous.

Let $\xi_i(s) = \tilde{\beta}_i(s) - \beta_i$, $i = 1, 2$. We have the following set inclusions

$$\begin{aligned}
\mathcal{S}_{1,t} &= \{1 \le s \le t : \hat{\pi}_s = 1\} \\
&\overset{\text{(a)}}{\supseteq} \{2\nu_0 + 1 \le s \le t : \hat{\pi}_s = 1,\ (\tilde{\beta}_1(s-1) - \tilde{\beta}_2(s-1))^T Z_s \ge h/2\} \\
&\overset{\text{(b)}}{\supseteq} \{2\nu_0 + 1 \le s \le t : (\tilde{\beta}_1(s-1) - \tilde{\beta}_2(s-1))^T Z_s \ge \rho_0/2\} \\
&= \Big\{2\nu_0 + 1 \le s \le t : (\beta_1 - \beta_2)^T Z_s \ge \frac{\rho_0}{2} - \xi_1^T(s-1)Z_s + \xi_2^T(s-1)Z_s\Big\} \\
&\supseteq \Big\{2\nu_0 + 1 \le s \le t : (\beta_1 - \beta_2)^T Z_s \ge \rho_0, |\xi_1^T(s-1)Z_s| \le \rho_0/4, \\
&\qquad |\xi_2^T(s-1)Z_s| \le \rho_0/4\Big\} \\
&=: \tilde{\mathcal{S}}_{1,t},
\end{aligned}$$

where (a) follows from truncation of the index set, and (b) follows from the structure of the policy $\hat{\pi}$ and $h \le \rho_0$.

1^0. Recall that $S_1(t) = \#\{\mathcal{S}_{1,t}\}$. Using the above inclusions we have

$$S_1(t) \ge \#\{\tilde{\mathcal{S}}_{1,t}\} \ge \sum_{s=2\nu_0+1}^{t} I\big\{(\beta_1 - \beta_2)^T Z_s \ge \rho_0\big\}\, w_s,$$

where $w_s := I(A_s)$, and the event A_s is defined by

$$A_s := \Big\{\max_{j=1,\dots,d+1} |\xi_{1,j}(s-1)| \le \frac{\rho_0}{4(1+rd)},\ \max_{j=1,\dots,d+1} |\xi_{2,j}(s-1)| \le \frac{\rho_0}{4(1+rd)}\Big\}.$$

Therefore for any $c > 0$

$$\begin{aligned}
\mathbb{P}\{S_1(t) \le ct\} &\le \mathbb{P}\Big\{\sum_{s=2\nu_0+1}^{t} I[(\beta_1 - \beta_2)^T Z_s \ge \rho_0]\, w_s \le ct\Big\} \\
&\le \mathbb{P}\Big\{\sum_{s=2\nu_0+1}^{t} I\big[(\beta_1 - \beta_2)^T Z_s \ge \rho_0\big]\, w_s \le ct, \\
&\qquad \sum_{s=2\nu_0+1}^{t} w_s \ge 2ct/p_*\Big\} \\
&\qquad + \mathbb{P}\Big\{\sum_{s=2\nu_0+1}^{t} w_s < 2ct/p_*\Big\} =: J_1 + J_2.
\end{aligned}$$

Denote $p' = \mathbb{P}\{(\beta_1 - \beta_2)^T Z_s \ge \rho_0\}$ and observe that by (13), $p' \ge p_*$. Note also that w_s is $\mathcal{F}_{s-1}$-measurable; hence $\big(\sum_{s=2\nu_0+1}^{t}[p' - I\{(\beta_1 -$

$\beta_2)^T Z_s \geq \rho_0\}]w_s, \mathcal{F}_s\big)$ is a martingale with bounded differences. Therefore by the Azuma–Hoeffding inequality

$$\begin{aligned} J_1 &\leq \mathbb{P}\Bigg\{ \sum_{s=2\nu_0+1}^{t} \Big[p' - I\{(\beta_1-\beta_2)^T Z_s \geq \rho_0\}\Big] w_s \geq p' \sum_{s=2\nu_0+1}^{t} w_s - ct, \\ &\qquad p_* \sum_{s=2\nu_0+1}^{t} w_s \geq 2ct \Bigg\} \\ &\leq \mathbb{P}\Bigg\{ \sum_{s=2\nu_0+1}^{t} \Big[p' - I\{(\beta_1-\beta_2)^T Z_s \geq \rho_0\}\Big] w_s \geq ct \Bigg\} \\ &\leq \exp\Bigg\{ -\frac{c^2t^2}{2(t-2\nu_0)} \Bigg\}. \end{aligned}$$

In order to bound J_2 note that

$$\begin{aligned} &\Bigg\{ \sum_{s=2\nu_0+1}^{t} I\{A_s^c\} > t - 2d - 2 - \frac{2ct}{p_*} \Bigg\} \subseteq \Bigg\{ \sum_{s=2\nu_0+1}^{t} I\{A_s^c\} > \frac{t}{2} \Bigg\} \\ &\subseteq \bigcup_{s=\lfloor t/2 \rfloor}^{t} A_s^c, \end{aligned}$$

provided that $t \geq 8\nu_0 + 1$ and $c \leq p_*/8$. Therefore using a union bound and (25) with the vector a taken to be the standard basis vectors in $\mathbb{R}^{d+1}$ we obtain

$$\begin{aligned} J_2 &\leq \sum_{s=\lfloor t/2 \rfloor}^{t} \mathbb{P}\{A_s^c\} \leq \sum_{s=\lfloor t/2 \rfloor}^{t} \sum_{i=1}^{2} \mathbb{P}\Big\{ \max_{j=1,\ldots,d+1} |\xi_{i,j}(s-1)| \geq \frac{\rho_0}{4(1+rd)} \Big\} \\ &\leq (d+1) \sum_{s=\lfloor t/2 \rfloor}^{t} \sum_{i=1}^{2} P\Big(\frac{\rho_0}{4(1+rd)}, T_i(s-1), 1 \Big), \end{aligned}$$

where we have used Proposition 1; recall that $P(\cdot,\cdot,\cdot)$ is defined in (18). Setting $c = p_*/8$ we arrive to (19).

2^0. By Lemma 3 we have that for any two sets of indices $\mathcal{J}_2 \subseteq \mathcal{J}_1$, $\lambda_{\min}\{\Sigma(\mathcal{J}_2)\} \leq \lambda_{\min}\{\Sigma(\mathcal{J}_1)\}$. Hence

$$\begin{aligned} \lambda_{\min}\{\Sigma(\mathcal{S}_{1,t})\} &\geq \lambda_{\min}\{\Sigma(\tilde{\mathcal{S}}_{1,t})\} \\ &\geq \lambda_{\min}\Bigg\{ \sum_{s=2\nu_0+1}^{t} w_s Z_s Z_s^T I\{(\beta_1-\beta_2)^T Z_s \geq \rho_0\} \Bigg\} \\ &= \lambda_{\min}\Bigg\{ \sum_{s=2\nu_0+1}^{t} w_s U_s^+ Z_s Z_s^T \Bigg\}. \end{aligned}$$

We can write

$$\sum_{s=2\nu_0+1}^{t} w_s U_s^+ Z_s Z_s^T = \sum_{s=2\nu_0+1}^{t} w_s \mathbb{E}\{U_s^+ Z_s Z_s^T\} + \sum_{s=2\nu_0+1}^{t} w_s E(s),$$

where

$$E(s) = U_s^+ Z_s Z_s^T - \mathbb{E} U_s^+ Z_s Z_s^T.$$

Thus we have

$$\begin{aligned}\lambda_{\min}(\Sigma(\tilde{\mathcal{S}}_{1,t})) \quad &\geq \quad \lambda_* \sum_{s=2\nu_0+1}^{t} w_s \\ &\quad - \max\Big\{\Big|\lambda_{\min}\big(\sum_{s=2\nu_0+1}^{t} w_s E(s)\big)\Big|, \Big|\lambda_{\max}\big(\sum_{s=2\nu_0+1}^{t} w_s E(s)\big)\Big|\Big\}.\end{aligned}$$

Define the event

$$D_t := \Big\{\max_{k,l=1,\ldots,d+1}\Big|\sum_{s=2\nu_0+1}^{t} w_s E_{k,l}(s)\Big| \leq \frac{\lambda_* t}{8(d+1)}\Big\},$$

where $E_{k,l}(s)$ are the components of the matrix $E(s)$. Note that for any k and l,
$\big(\sum_{s=2\nu_0+1}^{t} w_s E_{k,l}(s), \mathcal{F}_s\big)$ is a martingale with bounded differences, $|E_{k,l}(s)| \leq 2(1 \vee r^2)$. Therefore by a union bound and the Azuma–Hoeffding inequality

$$\mathbb{P}\{D_t^c\} \leq (d+1)(d+2)\exp\Big\{-\frac{\lambda_*^2 t}{128(1\vee r^2)^2(1+d)^2}\Big\}.$$

On the event D_t

$$\max\Big\{\Big|\lambda_{\min}\Big(\sum_{s=2\nu_0+1}^{t} w_s E(s)\Big)\Big|, \Big|\lambda_{\max}\Big(\sum_{s=2\nu_0+1}^{t} w_s E(s)\Big)\Big|\Big\} \leq \frac{\lambda_* t}{8}.$$

Therefore,

$$\begin{aligned}\mathbb{P}\Big\{\lambda_{\min}\big(\Sigma(\tilde{\mathcal{S}}_{1,t})\big) \leq \frac{\lambda_* t}{8}\Big\} \quad &\leq \quad \mathbb{P}\Big\{\lambda_{\min}\big(\Sigma(\tilde{\mathcal{S}}_{1,t})\big) \leq \frac{\lambda_* t}{8}, D_t\Big\} + \mathbb{P}\{D_t^c\} \\ &\leq \quad \mathbb{P}\Big\{\lambda_* \sum_{s=2\nu_0+1}^{t} w_s \leq \frac{\lambda_* t}{4}, D_t\Big\} + \mathbb{P}\{D_t^c\} \\ &\leq \quad (d+1)\sum_{s=\lfloor t/2\rfloor}^{t}\sum_{i=1}^{2} P\Big(\frac{\rho_0}{4(1+rd)}, T_i(s-1), 1\Big) \\ &\qquad + (d+1)(d+2)\exp\Big\{-\frac{\lambda_*^2 t}{128(1\vee r^2)^2(1+d)^2}\Big\},\end{aligned}$$

where the last inequality follows from the bound on J_2. This concludes the proof. ∎

7.3. *Proof of Proposition 3.* Observe that

$$\hat{\beta}_i(t) - \beta_i = \Sigma^{-1}(\mathcal{S}_{i,t}) \sum_{s=1}^{t} Z_s \varepsilon_s I(\hat{\pi}_s = i), \quad i = 1, 2. \tag{27}$$

Furthermore, by Lemma 2 in the Appendix for any $\varkappa > 0$

$$\mathbb{P}\Big\{\Big|\sum_{s=1}^{t} \varepsilon_s I(\hat{\pi}_s = i)\Big| \geq \varkappa\Big\} \leq 2\exp\Big\{-\frac{\varkappa^2}{2t\sigma^2}\Big\},$$

$$\mathbb{P}\Big\{\max_{k=1,\ldots,d}\Big|\sum_{s=1}^{t} X_{s,k}\varepsilon_s I(\hat{\pi}_s = i)\Big| \geq \varkappa\Big\} \leq 2d\exp\Big\{-\frac{\varkappa^2}{2t\sigma^2 r^2}\Big\}.$$

Consider the event

$$B_{i,t} := \Big\{\lambda_{\min}\big(\Sigma(\mathcal{S}_{i,t})\big) \geq \frac{\lambda_* t}{8}\Big\}, \quad i = 1, 2. \tag{28}$$

Then it follows from Lemma 2 in the Appendix, the previous two inequalities, and (27) that

$$\begin{aligned}
&\mathbb{P}\Big\{\|\hat{\beta}_i(t) - \beta_i\| \geq \varkappa, B_{i,t}\Big\} \\
&\leq \mathbb{P}\Big\{\lambda_{\max}\{\Sigma^{-1}(\mathcal{S}_{i,t})\}\Big\|\sum_{s=1}^{t} Z_s\varepsilon_s I(\hat{\pi}_s = i)\Big\| \geq \varkappa,\ B_{i,t}\Big\} \\
&\leq \mathbb{P}\Big\{\Big\|\sum_{s=1}^{t} Z_s\varepsilon_s I(\hat{\pi}_s = i)\Big\| \geq \varkappa\lambda_{\min}\{\Sigma(\mathcal{S}_{i,t})\},\ B_{i,t}\Big\} \\
&\leq \mathbb{P}\Big\{\Big\|\sum_{s=1}^{t} Z_s\varepsilon_s I(\hat{\pi}_s = i)\Big\| \geq \frac{\varkappa\lambda_* t}{8}\Big\} \\
&\leq \mathbb{P}\Big\{\Big|\sum_{s=1}^{t} \varepsilon_s I(\hat{\pi}_s = i)\Big| \geq \frac{\varkappa\lambda_* t}{16}\Big\} \\
&\quad + \mathbb{P}\Big\{\max_{k=1,\ldots,d}\Big|\sum_{s=1}^{t} X_{s,k}\varepsilon_s I(\hat{\pi}_s = i)\Big| \geq \frac{\varkappa\lambda_* t}{16\sqrt{d}}\Big\} \\
&\leq 2\exp\Big\{-\frac{\varkappa^2\lambda_*^2 t}{256\sigma^2}\Big\} + 2d\exp\Big\{-\frac{\varkappa^2\lambda_*^2 t}{256\, d\sigma^2 r^2}\Big\},
\end{aligned}$$

where the second to last inequality follows from the fact that $Z_s = (1, X_s)$. The second statement of the proposition follows from the above bound and the Cauchy–Schwarz inequality.

∎

7.4. *Proof of Theorem 1.* In the proof below $c_1, c_2, \ldots$ denote the absolute constants, while $C_1, C_2, \ldots$ stand for positive constants that can depend on $\underline{\lambda}$, σ^2, d, ρ_0, b, r.

1^0. Let $\bar{\mathcal{T}}_n = \mathcal{T}_{1,n} \cup \mathcal{T}_{2,n}$ denote the set of time instances where the policy $\hat{\pi}$ performs forced sampling from the first and second arm; and recall that $T_i(t) = \#\{\mathcal{T}_{i,t}\}$. Using the Cauchy–Schwarz inequality and Assumptions (A1) and (A4) we have

$$
\begin{aligned}
R_n(\hat{\pi}, \pi^*) &= \mathbb{E} \sum_{t=1}^{n} |\beta_1^T Z_t - \beta_2^T Z_t| I\{\pi_t \neq \pi_t^*\} \\
&\leq 2b\sqrt{\lambda_{\max}(Q)}\, t_* + \mathbb{E} \sum_{\substack{t=t_*+1 \\ t \in \bar{\mathcal{T}}_n}}^{n} |\beta_1^T Z_t - \beta_2^T Z_t| I\{\pi_t \neq \pi_t^*\} \\
&\quad + \mathbb{E} \sum_{\substack{t=t_*+1 \\ t \notin \bar{\mathcal{T}}_n}}^{n} |\beta_1^T Z_t - \beta_2^T Z_t| I\{\pi_t \neq \pi_t^*\} \\
&\leq 2b\sqrt{\lambda_{\max}(Q)} \left[t_* + T_1(n) + T_2(n)\right] + \mathbb{E} \sum_{\substack{t=t_*+1 \\ t \notin \bar{\mathcal{T}}_n}}^{n} |\beta_1^T Z_t - \beta_2^T Z_t| I\{\pi_t \neq \pi_t^*\}. \qquad (29)
\end{aligned}
$$

By definition of the policy $\hat{\pi}$,

$$
\{\hat{\pi}_t \neq \pi_t^*\} = \bigcup_{i=1}^{4} D_i(t)
$$

where

$$
\begin{aligned}
D_1(t) &:= \big\{(\tilde\beta_1(t-1)-\tilde\beta_2(t-1))^T Z_t > \frac{h}{2},\ (\beta_1-\beta_2)^T Z_t \le 0\big\}\\
D_2(t) &:= \big\{(\tilde\beta_1(t-1)-\tilde\beta_2(t-1))^T Z_t < -\frac{h}{2},\ (\beta_1-\beta_2)^T Z_t \ge 0\big\}\\
D_3(t) &:= \big\{\big|(\tilde\beta_1(t-1)-\tilde\beta_2(t-1))^T Z_t\big| \le \frac{h}{2},\\
&\qquad (\hat\beta_1(t-1)-\hat\beta_2(t-1))^T Z_t \le 0,\ (\beta_1-\beta_2)^T Z_t > 0\big\}\\
D_4(t) &:= \big\{\big|(\tilde\beta_1(t-1)-\tilde\beta_2(t-1))^T Z_t\big| \le \frac{h}{2},\\
&\qquad (\hat\beta_1(t-1)-\hat\beta_2(t-1))^T Z_t \ge 0,\ (\beta_1-\beta_2)^T Z_t < 0\big\}.
\end{aligned}
$$

The event $D_1(t)\cup D_2(t)$ occurs when at the step t the policy makes an error taking a myopic action based on the forced sampling estimates $\tilde\beta_i(t-1)$, $i=1,2$. The event $D_3(t)\cup D_4(t)$ describes the situation when at step t the policy incurs an error taking myopic action based on the estimates $\hat\beta_i(t-1)$, $i=1,2$.

Write

$$
\begin{aligned}
\mathbb{E}\sum_{\substack{t=t_*+1\\ t\notin\bar{\mathcal{T}}_n}}^{n} |\beta_1^T Z_t-\beta_2^T Z_t| I\{\pi_t\ne\pi_t^*\} &\le \mathbb{E}\sum_{\substack{t=t_*+1\\ t\notin\bar{\mathcal{T}}_n}}^{n} |\beta_1^T Z_t-\beta_2^T Z_t| I\{D_1(t)\cup D_2(t)\}\\
&\quad +\mathbb{E}\sum_{\substack{t=t_*+1\\ t\notin\bar{\mathcal{T}}_n}}^{n} |\beta_1^T Z_t-\beta_2^T Z_t| I\{D_3(t)\cup D_4(t)\}\\
&=: E_1+E_2
\end{aligned}
$$

2^0. First we bound E_1. We have

$$
\begin{aligned}
\mathbb{P}\{D_1(t)\} &\le \mathbb{P}\big\{(\tilde\beta_1(t-1)-\beta_1)^T Z_t+(\beta_2-\tilde\beta_2(t-1))^T Z_t > h/2\big\}\\
&\le \mathbb{P}\big\{|(\tilde\beta_1(t-1)-\beta_1)^T Z_t| > h/4\big\}+\mathbb{P}\big\{|(\beta_2-\tilde\beta_2(t-1))^T Z_t| > h/4\big\}\\
&\le P(h/4, T_1(t-1), \sqrt{1+r^2d})+P(h/4, T_2(t-1), \sqrt{1+r^2d}), \qquad (30)
\end{aligned}
$$

where we have used Proposition 1. The same upper bound holds for $\mathbb{P}\{D_2(t)\}$. Thus in view of (30)

$$
E_1 \le 2b(1+r^2d)^{1/2}\sum_{t=t_*+1}^{n}\sum_{i=1}^{2} P(h/4, T_i(t-1), \sqrt{1+r^2d}).
$$

By definition, $t_* \geq 8\nu_0 + 1$; then (6) and (18) imply that if q satisfies

$$q \leq \min\Big\{\frac{h^2\underline{\lambda}}{128\sigma^2(1+r^2d)},\ \frac{\underline{\lambda}^2}{16(d+1)^2(r^2\vee r^4)}\Big\}$$

then

$$E_1 \leq C_1 \sum_{t=t_*+1}^{n} \frac{\sqrt{\ln t}}{(t-1)^2} \leq C_2.$$

3^0. Now we bound E_2. For brevity we will write $\Delta_\beta = \beta_1 - \beta_2$ and $\hat{\Delta}_\beta(t) = \hat{\beta}_1(t) - \hat{\beta}_2(t)$. We have

$$\begin{aligned}\mathbb{E}|\Delta_\beta^T Z_t| I\{D_3(t)\} &= \mathbb{E}|\Delta_\beta^T Z_t| I\big\{D_3(t), \Delta_\beta^T Z_t > \rho_0\big\} \\ &\quad + \mathbb{E}|\Delta_\beta^T Z_t| I\big\{D_3(t), \Delta_\beta^T Z_t \leq \rho_0\big\} \\ &=: J_1(t) + J_2(t).\end{aligned}$$

In words, $J_1(t)$ and $J_2(t)$ are the t–step contributions to the regret due to myopic action errors when the covariate X_t falls inside and outside the ρ_0–margin of the decision boundary $\Delta_\beta^T Z_t = 0$.

We have

$$\begin{aligned}J_1(t) &\leq 2b(1+r^2d)^{1/2}\,\mathbb{P}\Big\{(\hat{\beta}_1(t-1)-\beta_1)^T Z_t + (\beta_2 - \hat{\beta}_2(t-1))^T Z_t \leq -\rho_0\Big\} \\ &\leq 2b(1+r^2d)^{1/2} \sum_{i=1}^{2}\Big[\mathbb{P}\Big\{\big|(\hat{\beta}_i(t-1)-\beta_i)^T Z_t\big| \geq \rho_0/2,\ B_{i,t-1}\Big\} + \mathbb{P}\{B_{i,t-1}^c\}\Big],\end{aligned} \tag{31}$$

where $B_{i,t}$ is defined in (28). By inequality (20) of Proposition 2 and (6),

$$\mathbb{P}\{B_{i,t}^c\} \leq C_3(t-1)^{-2}\sqrt{\ln(t-1)}, \tag{32}$$

provided that

$$q \leq \min\Big\{\frac{\rho_0^2\underline{\lambda}}{192\sigma^2(1+rd)^2},\ \frac{\underline{\lambda}^2}{24(d+1)^2(r^2\vee r^4)}\Big\}.$$

In addition, by the second inequality of Proposition 3 the first probability on the RHS of (31) decreases exponentially in t. Therefore $\sum_{t=t_*+1}^{n} J_1(t) \leq C_4$, i.e., we have finite contribution to the regret from myopic actions errors when X_t's are far from the decision boundary.

Now we proceed with $J_2(t)$. Let $B_t := B_{1,t} \cap B_{2,t}$; then

$$\begin{aligned}J_2(t) &= \mathbb{E}|\Delta_\beta^T Z_t| I\big\{D_3(t), \Delta_\beta^T Z_t \leq \rho_0\big\} I\{B_{t-1}\} \\ &\quad + \mathbb{E}|\Delta_\beta^T Z_t| I\big\{D_3(t), \Delta_\beta^T Z_t \leq \rho_0\big\} I\{B_{t-1}^c\} \\ &\leq \mathbb{E}|\Delta_\beta^T Z_t| I\big\{D_3(t), \Delta_\beta^T Z_t \leq \rho_0\big\} I\{B_{t-1}\} + 2(1+r^2d)^{1/2}\mathbb{P}\{B_{t-1}^c\} \\ &=: J_2^{(1)}(t) + J_2^{(2)}(t).\end{aligned}$$

By (32) and our choice of q

$$J_2^{(2)}(t) \le C_5 \frac{\sqrt{\ln(t-1)}}{(t-1)^2}, \quad \text{and} \quad \sum_{t=t_*+1}^{n} J_2^{(2)}(t) \le C_6.$$

It remains to bound $J_2^{(1)}(t)$. Define the events

$$M_k = \Big\{ 2^{-k-1}\rho_0 < (\beta_1 - \beta_2)^T Z_t \le 2^{-k}\rho_0 \Big\}, \quad k = 0, 1, 2, \ldots$$

Then

$$\begin{aligned}
J_2^{(1)}(t) &= \mathbb{E} \sum_{k=0}^{\infty} |\Delta_\beta^T Z_t| I\Big\{ M_k, (\hat\beta_1(t-1) - \beta_1)^T Z_t - (\hat\beta_2(t-1) - \beta_2)^T Z_t \le -\Delta_\beta^T Z_t,\ B_{t-1} \Big\} \\
&\le \rho_0\, \mathbb{E} \sum_{k=0}^{\infty} 2^{-k} I\big\{ (\beta_1 - \beta_2)^T Z_t \le \rho_0 2^{-k} \big\} I\Big\{ \sum_{i=1}^{2} \big| (\hat\beta_i(t-1) - \beta_i)^T Z_t \big| \ge 2^{-k-1}\rho_0,\ B_{t-1} \Big\} \\
&\le \rho_0 \sum_{k=0}^{\infty} 2^{-k} \mathbb{P}\big\{ (\beta_1 - \beta_2)^T Z_t \le \rho_0 2^{-k} \big\} \sum_{i=1}^{2} \mathbb{P}\Big\{ \|\hat\beta_i(t-1) - \beta_i\| \ge \frac{2^{-k-2}\rho_0}{\sqrt{1+r^2 d}},\ B_{t-1} \Big\} \\
&\le 2(2d+2) L \rho_0^2 \sum_{k=0}^{\infty} 2^{-2k} \exp\{-2^{-2k} z\},
\end{aligned}$$

where

$$z = \frac{\rho_0^2 \lambda_*^2 t}{4096\, \sigma^2 (r^2 \vee 1) d (1 + r^2 d)}\ . \tag{33}$$

Here the second inequality is obtained by bounding $\|Z_t\| \le \sqrt{1+r^2 d}$, conditioning on $\mathcal{F}_{t-1}$ and using the fact that Z_t is independent of $\mathcal{F}_{t-1}$; and the last inequality follows from Assumption (A2) and Proposition 3. Taking into account that $z \ge 1$ by (14), and using Lemma 4 we finally obtain

$$J_2^{(1)}(t) \le c_1 K_2 L d^2 \lambda_*^{-2} \sigma^2 (1 + r^2 d)(1 \vee r^2) t^{-1},$$

where c_1 is an absolute constant, and K_α is given in Lemma 4 in Appendix. Combining the above bounds we obtain that

$$\sum_{t=t_*+1}^{n} \mathbb{E} |\Delta_\beta^T Z_t| I\{D_3(t)\} \ \le\ c_2 L \sigma^2 d^2 (1 + r^2 d) \lambda_*^{-2} (1 \vee r^2) \ln n + C_7, \tag{34}$$

where c_2 is an absolute constant.

We note that $\sum_{t=t_*+1}^{n} \mathbb{E}|\Delta_\beta^T Z_t| I\{D_4(t)\}$ is also bounded by the expression on the RHS of (34); the proof of this bound follows verbatim to the above. Combining these results with (29) and taking into account the upper bound in (6) we complete the proof. ∎

Remark on the proof of (17). The proof follows Theorem 1 with the following changes. Instead of (29) we have

$$M_n(\hat{\pi}, \pi^*) \leq \big[t_* + T_1(n) + T_2(n)\big] + \mathbb{E} \sum_{\substack{t=t_*+1 \\ t \notin \mathcal{T}_n}}^{n} I\{\pi_t \neq \pi_t^*\}.$$

The first term on the RHS above is upper bounded by $t_* + q^{-1}\ln(n+1)$. Bounding the second term on the RHS goes along the same lines. The dominant term here will be $J_2^{(1)}(t)$ that now takes the form

$$J_2^{(1)}(t) \leq 2(2d+2)L\rho_0 \sum_{k=0}^{\infty} 2^{-k} \exp\{-2^{-2k} z\},$$

where z is given by (33). Then application of Lemma 4 leads to (17). ∎

7.5. *Proof of Theorem 2.* Let $X_t = (X_{t,1}, \ldots, X_{t,d})$ be a random vector with independent components, each distributed with a density f_X with respect to the Lebesgue measure such that $\mathrm{supp}(f_X) = [-1, 1]$, and

$$0 < c_f \leq f_X(x) \leq C_f < \infty, \quad \forall x \in [-1, 1].$$

Denote $\beta = (\beta_1, \beta_2) \in \mathbb{R}^{d+1} \times \mathbb{R}^{d+1}$ the vector of the arm parameters; we refer to $\beta = (\beta_1, \beta_2)$ as the *configuration*. Let $B \subseteq \mathbb{R}^{d+1} \times \mathbb{R}^{d+1}$ be a set of configurations, and let λ be a probability measure on B, with $\mathbb{E}_\lambda$ denoting the expectation with respect to λ. (In what follows B and λ will be chosen so that assumptions defining the class $\mathcal{P}$ hold.)

Put $\Delta_\beta := \beta_1 - \beta_2$ and fix an arbitrary policy π. The regret of π can be

bounded from below as follows

$$\begin{aligned}
&\sup\Big\{R_n(\pi,\pi^*) : \big(P_{X,Y^{(1)}}, P_{X,Y^{(2)}}\big) \in \mathcal{P}\Big\} \\
&\geq \sup_{\beta\in B} \mathbb{E}\sum_{t=1}^n |\Delta_\beta^T Z_t| \Big[I\{\Delta_\beta^T Z_t \geq 0, \pi_t = 2\} + I\{\Delta_\beta^T Z_t < 0, \pi_t = 1\}\Big] \\
&\geq \mathbb{E}\sum_{t=1}^n \Bigg(\mathbb{E}_\lambda\Big[\Delta_\beta^T Z_t I\{\Delta_\beta^T Z_t \geq 0\} \mid \mathcal{F}_{t-1}^+\Big] I\{\pi_t = 2\} \\
&\quad - \mathbb{E}_\lambda\Big[\Delta_\beta^T Z_t I\{\Delta_\beta^T Z_t < 0\} \mid \mathcal{F}_{t-1}^+\Big] I\{\pi_t = 1\}\Bigg).
\end{aligned}$$

The Bayesian decision rule is: choose $\hat{\pi}_t = 1$ if

$$\mathbb{E}_\lambda\Big[\Delta_\beta^T Z_t I\{\Delta_\beta^T Z_t \geq 0\} \mid \mathcal{F}_{t-1}^+\Big] \geq -\mathbb{E}_\lambda\Big[\Delta_\beta^T Z_t I\{\Delta_\beta^T Z_t < 0\} \mid \mathcal{F}_{t-1}^+\Big]$$

and $\hat{\pi}_t = 2$ otherwise. Set $\gamma_{t-1} := \mathbb{E}_\lambda[\Delta_\beta \mid \mathcal{F}_{t-1}^+]$. Now, since Z_t is independent of both $\mathcal{F}_{t-1}$ and β, then $\mathbb{E}_\lambda[\Delta_\beta \mid \mathcal{F}_{t-1}^+] = \mathbb{E}_\lambda[\Delta_\beta \mid \mathcal{F}_{t-1}]$. Therefore, the Bayesian policy takes the form $\hat{\pi}_t = I\{\gamma_{t-1}^T Z_t \geq 0\} + 2I\{\gamma_{t-1}^T Z_t < 0\}$. Thus we have

$$R_n(\pi;\mathcal{P}) \geq \mathbb{E}\sum_{t=1}^n \mathbb{E}_\lambda |\Delta_\beta^T Z_t| I\Big\{\mathrm{sign}(\Delta_\beta^T Z_t) \neq \mathrm{sign}(\gamma_{t-1}^T Z_t)\Big\}. \tag{35}$$

Now we specify the probability measure λ on $B \subseteq \mathbb{R}^{d+1} \times \mathbb{R}^{d+1}$; λ will be taken as the product measure of the form $\lambda = (\delta_0 \times \delta_1 \times \delta_0 \times \cdots \times \delta_0) \times (\lambda_0 \times \delta_0 \times \cdots \times \delta_0)$, where δ_a denotes the point mass at a, and λ_0 is a probability distribution on $[-1/2, 1/2]$ to be specified. Under these circumstances $\beta_1 = (0, 1, 0, \ldots, 0)$ and $\beta_2 = (\theta, 0, \ldots, 0)$, where θ is a random variable distributed according to λ_0. With this construction, the observations are realized as: $Y_t^{(1)} = X_{t,1} + \varepsilon_t^{(1)}$ and $Y_t^{(2)} = \theta + \varepsilon_t^{(2)}$ for $t = 1, 2, \ldots$. Note that with the above choice of primitives, for any probability distribution λ_0 on $[-1/2, 1/2]$ we have that: Assumption (A1) holds with $r = 1$; Assumption (A2) holds with $L = 2C_f$ and $\rho_0 = 1/4$; Assumption (A3) holds with $p_* = c_f/4$; and Assumption (A4) holds with $b = 1$.

With the above setting, $\gamma_{t-1} = (\hat{\theta}_{t-1}, 1, \ldots, 0)$ with $\hat{\theta}_{t-1} := -\mathbb{E}_{\lambda_0}[\theta|\mathcal{F}_{t-1}]$, and note that $|\hat{\theta}_{t-1}| \leq 1/2$. Therefore

$$\begin{aligned}
&\mathbb{E}_\lambda |\Delta_\beta^T Z_t| I\big\{\mathrm{sign}(\Delta_\beta^T Z_t) \neq \mathrm{sign}(\gamma_{t-1}^T Z_t)\big\} \\
&= \mathbb{E}_{\lambda_0} |X_{t,1} - \theta| \Big[I\big\{X_{t,1} \geq \theta, X_{t,1} < \hat{\theta}_{t-1}\big\} + I\big\{X_{t,1} > \theta, X_{t,1} \leq \hat{\theta}_{t-1}\big\}\Big].
\end{aligned}$$

Because $\hat{\theta}_{t-1}$ is $\mathcal{F}_{t-1}$–measurable, and $X_{t,1}$ is independent of $\mathcal{F}_{t-1}$ we have from (35)

$$R_n(\pi;\mathcal{P}) \geq \mathbb{E}\mathbb{E}_{\lambda_0} \sum_{t=1}^{n} \int_{\theta\wedge\hat{\theta}_{t-1}}^{\theta\vee\hat{\theta}_{t-1}} |x-\theta| f_X(x)\mathrm{d}x \geq \frac{c_f}{2} \sum_{t=1}^{n} \mathbb{E}\,\mathbb{E}_{\lambda_0} |\hat{\theta}_{t-1}-\theta|^2.$$

Let $\ell(\theta)$ denote the density of λ_0. Let $\mathcal{F}^*_{t-1} = \sigma(X_{1,1},\ldots,X_{t-1,1},Y_1^{(1)},\ldots,Y_{t-1}^{(1)},Y_1^{(2)},\ldots,Y_{t-1}^{(2)})$; because $\mathcal{F}_{t-1} \subset \mathcal{F}^*_{t-1}$,

$$\sum_{t=1}^{n} \mathbb{E}\,\mathbb{E}_{\lambda_0} |\hat{\theta}_{t-1}-\theta|^2 \geq \sum_{t=1}^{n} \inf_{\delta_{t-1}} \int \mathbb{E}|\delta_{t-1}-\theta|^2 \ell(\theta) d\theta,$$

where the inf above is taken over all sequences $\delta = (\delta_t)$ such that δ_{t-1} is $\mathcal{F}^*_{t-1}$–measurable. Therefore we have

$$R_n(\pi;\mathcal{P}) \geq \frac{c_f}{2} \sum_{t=1}^{n} \inf_{\delta_{t-1}} \int \mathbb{E}|\delta_{t-1}-\theta|^2 \ell(\theta) d\theta. \tag{36}$$

Thus the problem is reduced to establishing a lower bound on the Bayesian risk in the problem of estimating the scalar parameter $\theta \in [-1/2, 1/2]$ from observations $\mathcal{Y}_{t-1} = \{(X_{s,1}, Y_s^{(1)}, Y_s^{(2)}), s = 1,\ldots,t-1\}$, where $Y_s^{(1)} = X_{s,1} + \varepsilon_s^{(1)}$ and $Y_s^{(2)} = \theta + \varepsilon_s^{(2)}$ for $s = 1,\ldots,t-1$ and $\{\varepsilon_t^{(1)}, \varepsilon_t^{(2)}\}$ are sequences of mutually independent iid zero mean Gaussian random variables with variance σ^2. This problem is well–studied, and there are different methods for establishing such lower bounds. In particular, by the van Trees inequality [see (Gill and Levit 1995)]

$$\inf_{\delta_{t-1}} \int \mathbb{E}|\delta_{t-1}-\theta|^2 \ell(\theta) d\theta \geq \frac{1}{I_{t-1} + I(\ell)}, \tag{37}$$

where I_{t-1} is the expected Fisher information for θ associated with the conditional density of the observations $\mathcal{Y}_{t-1}$ given θ; and $I(\ell)$ is the Fisher information for the location parameter in ℓ. Thus,

$$I_{t-1} := \mathbb{E}\Big[\frac{\partial}{\partial\theta} \ln f(\mathcal{Y}_{t-1}|\theta)\Big]^2 = \mathbb{E}\Big[-\frac{1}{\sigma^2}\sum_{s=1}^{t-1}(Y_s^{(2)}-\theta)\Big]^2 = \frac{t-1}{\sigma^2}.$$

A standard choice of the distribution λ_0 is

$$\ell(\theta) = 2\cos^2(\pi\theta) I\{|\theta| \leq 1/2\}.$$

With this choice $I(\ell) = 4\pi^2$. Therefore applying the inequality in (37) for each summand in (36) we obtain

$$R_n(\pi;\mathcal{P}) \geq \frac{c_f}{2}\sum_{t=1}^{n}\frac{1}{I_{t-1}+I(\ell)} \geq \sigma^2\frac{c_f}{2}\sum_{t=2}^{n}\frac{1}{t-1+4\pi^2\sigma^2} \geq Cc_f\sigma^2\ln n.$$

for n large enough. This completes the proof. ∎

APPENDIX A: AUXILIARY RESULTS.

Lemma 1 *Let $X_s = (X_{s,1},\ldots,X_{s,d})$, $s = 1,\ldots,n$ be a sequence of independent identically distributed random vectors from $\mathbb{R}^d$ with bounded components, $\max_{j=1,\ldots,d}|X_{s,j}| \leq r$. Let $\mathbb{E}X_s = \mu$ and $\mathbb{E}X_sX_s^T = V$. For any $\varkappa > 0$ and any non–random index set $\mathcal{J} \subseteq \{1,\ldots,n\}$ define the event*

$$\begin{aligned}\Gamma(\varkappa,\mathcal{J}) \;&=\; \Big\{\max_{i=1,\ldots,d}\Big|\frac{1}{\#\{\mathcal{J}\}}\sum_{s\in\mathcal{J}}X_{s,i}-\mu_i\Big| \leq \varkappa\Big\}\\ &\quad\bigcap\Big\{\max_{i,j=1,\ldots,d}\Big|\frac{1}{\#\{\mathcal{J}\}}\sum_{s\in\mathcal{J}}X_{s,i}X_{s,j}-V_{ij}\Big| \leq \varkappa\Big\}. \qquad (38)\end{aligned}$$

Then

$$\mathbb{P}\{\Gamma^c(\varkappa,\mathcal{J})\} \;\leq\; d(d+3)\exp\Big\{-\frac{\varkappa^2\#\{\mathcal{J}\}}{2(r^2\vee r^4)}\Big\}.$$

Proof: The proof is an immediate consequence of the Hoeffding inequality, the union bound, and straightforward algebra. ∎

Lemma 2 *Let $\{w_t\}$ be a sequence of random variables such that w_s is $\mathcal{F}_{s-1}$–measurable and $|w_s| \leq L_w$ for all s almost surely. Let $\{\varepsilon_t\}$ be iid , $\varepsilon_t \sim \mathcal{N}(0,\sigma^2)$ and ε_t is independent of $\mathcal{F}_{t-1}$. Then for any $a > 0$ one has*

$$\mathbb{P}\Big\{\sum_{s=1}^{t}w_s\varepsilon_s > a\Big\} \leq \exp\Big\{-\frac{a^2}{2t\sigma^2L_w^2}\Big\}, \quad \forall t.$$

Proof: The proof is straightforward; we provide it for completeness. For any $\lambda > 0$ we have

$$\begin{aligned}\mathbb{E}\exp\Big\{\lambda\sum_{s=1}^{t}w_s\varepsilon_s\Big\} \;&=\; \mathbb{E}\prod_{s=1}^{t}\exp\{\lambda w_s\varepsilon_s\} \;=\; \mathbb{E}\Big\{\mathbb{E}\Big[\prod_{s=1}^{t}\exp\{\lambda w_s\varepsilon_s\}\Big|\mathcal{F}_{t-1}\Big]\Big\}\\ &=\; \mathbb{E}\Big[\prod_{s=1}^{t-1}\exp\{\lambda w_s\varepsilon_s\}\mathbb{E}\Big\{\exp\big(\lambda w_t\varepsilon_t\big)\Big|\mathcal{F}_{t-1}\Big\}\Big].\end{aligned}$$

Furthermore,

$$\mathbb{E}\Big\{ \exp\left(\lambda w_t \varepsilon_t\right) \Big| \mathcal{F}_{t-1} \Big\} = \exp\left(\frac{\sigma^2}{2}\lambda^2 w_t^2\right) \leq \exp\left(\frac{\sigma^2}{2}\lambda^2 L_w^2\right),$$

because the random variable $\lambda w_t \varepsilon_t$ is conditionally (on $\mathcal{F}_{t-1}$) Gaussian. Iterating this argument we finally obtain that

$$\mathbb{E}\exp\left\{\lambda \sum_{s=1}^{t} w_s \varepsilon_s\right\} \leq \exp\left(\frac{t}{2}\lambda^2\sigma^2 L_w^2\right).$$

∎

Lemma 3 *For $\mathcal{J} \subseteq \{1, 2, \ldots\}$ define $\Sigma(\mathcal{J}) := \sum_{s\in\mathcal{J}} Z_s Z_s^T$. If $\mathcal{J}_2 \subseteq \mathcal{J}_1$ then $\lambda_{\min}\{\Sigma(\mathcal{J}_2)\} \leq \lambda_{\min}\{\Sigma(\mathcal{J}_1)\}$.*

Proof: The proof is immediate and it is omitted. ∎

Lemma 4 *For $\alpha, z > 0$, let $V(\alpha, z) := \sum_{k=0}^{\infty} 2^{-\alpha k} \exp\{-z2^{-2k}\}$. Then for all $0 < \alpha \leq 2z$*

$$V(\alpha, z) \leq K_\alpha z^{-\alpha/2}, \quad K_\alpha = \left[\frac{(\alpha/2)^{\alpha/2}}{1-2^{-\alpha}} + \frac{\Gamma(\alpha/2)}{2\ln 2}\right], \tag{39}$$

where $\Gamma(\cdot)$ is the Gamma–function.

Proof: Let

$$I(\alpha, z) := \int_0^\infty 2^{-\alpha y} \exp\{-z2^{-2y}\}\mathrm{d}y. \tag{40}$$

Note that the integrand above has a unique (global) maximum at $y^* = -(1/2)\log_2(\alpha/2z)$ provided that $\alpha \leq 2z$. Put $k^* := \lfloor y^* \rfloor$ and write

$$\begin{aligned} V(\alpha, z) &= \sum_{k=0}^{k^*} 2^{-\alpha k} \exp\{-z2^{-2k}\} + \sum_{k=k^*+1}^{\infty} 2^{-\alpha k} \exp\{-z2^{-2k}\} \\ &=: V_1(\alpha, z) + V_2(\alpha, z). \end{aligned}$$

It follows that

$$V_2(\alpha, z) \;\leq\; \frac{2^{-\alpha y^*}}{1-2^{-\alpha}} \;=\; \frac{(\alpha/2)^{\alpha/2}}{1-2^{-\alpha}} z^{-\alpha/2}.$$

Since the integrand in (40) is monotone increasing on $[0, y^*)$, we have that

$$\begin{aligned} V_1(\alpha, z) &\leq \int_0^{y^*} 2^{-\alpha y} \exp\{-z2^{-2y}\}\mathrm{d}y \;\leq\; I(\alpha, z) \\ &= \frac{1}{\ln 2}\int_0^1 u^{\alpha-1}\exp\{-zu^2\}\mathrm{d}u \;\leq\; \frac{\Gamma(\alpha/2)}{2\ln 2} z^{-\alpha/2}, \end{aligned}$$

where $\Gamma(\cdot)$ denotes the gamma function. Thus for all $0 < \alpha \leq 2z$

$$V(\alpha, z) \leq z^{-\alpha/2}\left[\frac{(\alpha/2)^{\alpha/2}}{1-2^{-\alpha}} + \frac{\Gamma(\alpha/2)}{2\ln 2}\right].$$

as claimed. ∎

REFERENCES

Auer, P. (2002). Using confidence bounds for exploitation–exploration trade–offs. *J. Mach. Learn. Res.* **3**, 397–422.

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

Auer, P., Cesa-Bianchi, N., Freund, Y. and Schapire, R. (2002b). The nonstochastic multiarmed bandit problem. SIAM J. Comput. **32**, 48–77.

Berry, D. A. and Fristedt, B. (1985). *Bandit Problems.* Chapman and Hall, London.

Cesa–Bianchi, N. and Lugosi, G. (2006). *Prediction, Learning and Games.* Cambridge University Press, Cambridge.

Ginebra, J. and Clayton, M. K. (1995). Response surface bandits. *J. Roy. Statist. Soc. Ser. B* **57**, 771–784.

Gill, R. D. and Levit, B. Y. (1995). Applications of the Van Trees inequality: a Bayesian Cramer-Rao bound. *Bernoulli* **1**, 59–79.

Gittins, J. C. (1989) *Multi-armed Bandit Allocation Indices.* Wiley-Interscience Series in Systems and Optimization. John Wiley & Sons, Chichester.

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

Goldenshluger, A. and Zeevi, A. (2011). A note on performance limitations in bandit problems with side information. *IEEE Trans. Inf. Theory* **57**, 1707–1713.

Gooley, C. and Lattin, J. (2000). Dynamic customization of marketing messages in interactive media. *Research Paper No. 1664*, Research Paper Series, Graduate School of Business, Stanford University. Available at `https://gsbapps.stanford.edu/researchpapers/Library/RP1664.pdf`

Juditsky, A., Nazin, A., Tsybakov, A. and Vayatis, N. (2008). Gap–free bounds for stochastic multi–armed bandit. *IFAC World Congress*, 2008.

Lai, T. L. (1987). Adaptive treatment allocation and the multi-armed bandit problem. *Ann. Statist.* **15**, 1091–1114.

Lai, T. L. (1988). Asymptotic solutions of bandit problems. *Stochastic differential systems, stochastic control theory and applications (Minneapolis, Minn., 1986)*, 275–292, IMA Vol. Math. Appl., **10**, Springer, New York.

Lai, T. L. (2001). Sequential analysis: some classical problems and new challenges. *Statist. Sinica* **11**, 303–408.

Lai, T. L. and Robbins, H. (1985). Asymptotically efficient allocation rules. *Adv. Applied Math.* **6**, 4-22.

Lai, T. L. and Yakowitz, S. (1995). Machine learning and nonparametric bandit theory. *IEEE Trans. Automat. Control* **40**, 1199–1209.

Langford, J. and Zhang, T. (2008). The epoch–greedy algorithm for multiarmed bandits with side information. *Advances in Neural Information Processing Systems* **20**, 817–824, Cambridge, MIT Press.

Lu, T., Pál, D. and Pál, M. (2010). Contextual multi–armed bandits. Proceedings of the *13th International Conference on Artificial Intelligence and Statistics*. Available at `http://research.google.com/pubs/archive/37042.pdf`

Mersereau, A. J., Rusmevichientong, P. and Tsitsiklis, J.N. (2009). A structured multiarmed bandit problem and the greedy policy, *IEEE Trans. Automatic Control* **54** 2787-2802.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.* **55**, 527-535.

Rusmevichientong, P. and Tsitsiklis J.N. (2010). Linearly parametrized bandits. *Math. Oper. Res.* **35**, 395-411.

Sarkar, J. (1991). One-armed bandit problems with covariates. *Ann. Statist.* **19**, 1978–2002.

Stewart, G. W. and Sun, J. G. (1990). *Matrix Perturbation Theory.* Academic Press, Inc., Boston, MA.

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

Wang, C.-C, Kulkarni, S. and Poor, V. H. (2005). Bandit problems with side observations. *IEEE Trans. Automat. Control* **50**, 799-806.

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

Woodroofe, M. (1982). Sequential allocation with covariates. *Sankhyā Ser. A* **44**, 403–414.

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

Department of Statistics,
University of Haifa,
Haifa 31905, Israel
E-mail: goldensh@stat.haifa.ac.il

Graduate School of Business,
Columbia University,
New York, NY 10027, USA
E-mail: assaf@gsb.columbia.edu