A Note on Performance Limitations in Bandit Problems with Side Information

Alexander Goldenshluger and Assaf Zeevi

Abstract

We consider a sequential adaptive allocation problem which is formulated as a traditional two armed bandit problem but with one important modification: at each time step t, before selecting which arm to pull, the decision maker has access to a random variable X_t which provides information on the reward in each arm. Performance is measured as the fraction of time an inferior arm (generating lower mean reward) is pulled. We derive a minimax lower bound that proves that in the absence of sufficient statistical "diversity" in the distribution of the covariate X, a property that we shall refer to as lack of *persistent excitation*, no policy can improve on the best achievable performance in the traditional bandit problem *without* side information.

Keywords: Two-armed bandit, side information, inferior sampling rate, allocation rule, lower bound.

I. Introduction

Sequential allocation problems, otherwise known as multi-armed bandit problems, arise frequently in various areas of statistics, adaptive control, marketing, and economics. The first instance in this class of problems was introduced by Robbins (1952), and since then many variants thereof have been studied extensively in numerous different contexts [see, e.g., Berry and Fristedt (1985), Gittins (1989)].

In the prototypical two–armed bandit problem there are two statistical populations characterized by univariate density functions $f_{\theta_i}(x)$, $i = 1, 2$, where θ_i are unknown parameters belonging to a parameter set Θ. At each stage t one can sample an observation $Y_t = Y_t^{(i)}$ either from the first ($i = 1$) or from the second ($i = 2$) population. The *policy* π is a sequence of random

Research partially supported by NSF grant DMI-0447562, and by the US-Israel Binational Science Foundation (BSF) grant # 2006075

A. Goldenshluger is with the Department of Statistics, Haifa University, Haifa 31905, Israel e-mail: goldensh@stat.haifa.ac.il

A. Zeevi is with the Graduate School of Business, Columbia University, New York, NY 10027 USA e-mail: assaf@gsb.columbia.edu

variables $\pi_1, \pi_2, \ldots$ taking values in $\{1, 2\}$, and such that at each time t, π_t is only allowed to depend on past observations and allocation decisions. The total mean reward up to stage n associated with the policy π is

$$R_n(\pi, \theta) = \mathbb{E}_\theta^{\pi,n} \sum_{t=1}^{n} Y_t,$$

where $\mathbb{E}_\theta^{\pi,n}$ denotes expectation w.r.t. the joint distribution $\mathbb{P}_\theta^{\pi,n}$ of observations collected up to stage n when $\theta = (\theta_1, \theta_2)$, and under the policy π. The quality of a policy π is typically compared with the reward $R_n^*(\theta) = R_n(\pi^*, \theta)$ of the *oracle rule* π^*, that is, the rule which knows θ_1 and θ_2 and at each stage selects the best arm. The *regret* of a policy π is the defined as follows,

$$L_n(\pi, \theta) := R_n^*(\theta) - R_n(\pi, \theta),$$

and the goal is to develop a policy π such that the regret is as small as possible. In fact, it is not difficult to verify that the regret can be expressed as

$$L_n(\pi, \theta) = |\mu_1 - \mu_2| \, \mathbb{E}_\theta^{\pi,n}[T_{\text{inf}}(n)],$$

where μ_i is the mean reward under θ_i for $i = 1, 2$, and $T_{\text{inf}}(n)$ is the *inferior sampling rate*, or the total number of times the policy π sampled the inferior population (i.e., the one generating lower mean reward).

In the outlined setup, Lai and Robbins (1985) proposed a policy $\hat{\pi}$ such that

$$L_n(\hat{\pi}, \theta) \leq [C(\theta) + o(1)] \ln n, \quad n \to \infty, \tag{1}$$

where $C(\theta)$ is a constant depending on $\theta = (\theta_1, \theta_2)$ and the underlying density functions f_{θ_i}, $i = 1, 2$. It was also shown that the proposed policy cannot be improved upon in the following sense: among all policies such that for each fixed θ one has that $L_n(\pi, \theta) = o(n^a)$ for every $a > 0$, there does not exist a policy with regret smaller than the bound in (1). For related results and extensions we refer to Lai (1987), Anantharam, Varaiya and Warland (1987a, 1987b), Lai and Yakowitz (1995), Kulkarni and Lugosi (2000), and Auer, Cesa-Bianchi and Fischer (2002).

The described allocation model assumes sequential sampling from two "homogeneous" populations. However, in many practical situations some additional information can be utilized for allocation purposes. In particular, imagine that at each stage t a random covariate X_t is given, and the reward in each arm now depends also on the value of this side observation. In this manner, the first arm may be superior for one value of the covariate but inferior for a different one. For examples of such formulations we refer to Woodroofe (1979, 1982), Clayton (1989), Sarkar (1991), Yang and Zhu (2002), Wang, Kulkarni and Poor (2005), Langford and Zhang (2008), Goldenshluger and Zeevi (2009) and references therein. One of the main questions that arise in this context is whether additional information translates into performance improvement

vis-a-vis the traditional bandit setting summarized above. Some instances where it is possible to achieve *bounded* regret were given in Lai and Robbins (1984), Wang, Kulkarni and Poor (2005) and Goldenshluger and Zeevi (2009).

This paper identifies a key property of the side information which affects the intrinsic complexity of the allocation problem. In particular, we show that the expected inferior sampling rate *cannot be bounded* unless a suitable assumption is imposed on the variability or "diversity" that characterizes the covariates. To that end, we consider a very simple setting of a two-armed bandit problem with discrete valued covariates such that the rewards in each arm are governed by linear regression models. We derive a minimax lower bound on the expected number of inferior arm selections which is shown to be of the order $\ln n$ when the total number of samples taken is n; see Theorem 1 in Section 2. Our proof introduces a new bounding technique predicated on information theoretic arguments, which helps in elucidating the fundamental complexity of sequential allocation problems.

The diversity property alluded to above can be described also as *persistence of excitation*; a term that is often used in the adaptive control and system identification literature. In the context of our problem this property can be characterized informally as follows: the distribution of the covariates is such that under the an oracle sampling rule it is possible to learn about conditional distributions of the rewards associated with each arm, without having to incur errors in the arm selections; see Definition 1 in Section 2 for a more precise description. In the absence of this characteristic, any policy must sample the inferior arm a large number of times, as this is the only way to eventually distinguish which arm should be pulled for each covariate value. Our results indicate that without persistence of excitation it is not possible to construct policies in which the expected number of incorrect arm selections stays bounded as $n \to \infty$. In particular, according to Theorem 2 in Section 2, the expected inferior sampling rate diverges to infinity at the same rate as that characterizing the traditional bandit problem of Lai and Robbins (1985). In other words, in the absence of persistent excitation, the side information *does not* lead to any improvement in the expected inferior sampling rate.

II. Formulation and Main Results

A. Description of the model.

Consider the following formulation of a two–armed bandit problem. One observes a sequence $X_1, X_2, \ldots$ of i.i.d. random variables with common distribution P_X sequentially in time. At each stage t, one can allocate the covariate X_t to i-th arm ($i = 1, 2$) of the bandit machine obtaining the response $Y_t = Y_t^{(i)}$, where

$$Y_t^{(i)} = \alpha_i + \beta_i X_t + \varepsilon_t^{(i)}, \quad i = 1, 2. \tag{2}$$

Here $\theta_i = (\alpha_i, \beta_i) \in \mathbb{R}^2$ are unknown parameters, and $\varepsilon_t^{(i)}$ are the i.i.d. normal random variables with zero mean and variance σ^2, independent of X_t. If the i-th arm is selected at stage t, the obtained reward is equal to $Y_t^{(i)}$, and the goal is to maximize the total expected reward up to the stage n. In what follows we refer to $\theta = (\theta_1, \theta_2) = (\alpha_1, \beta_1, \alpha_2, \beta_2)$ as a *configuration* and regard it as a vector in $\mathbb{R}^4$.

By *policy* π we mean a sequence of random variables $\pi_1, \pi_2, \ldots$ taking values in $\{1, 2\}$ such that π_t is measurable with respect to the σ–field $\mathcal{F}_{t-1}$ generated by the previous observations X_1, Y_1, $\ldots, X_{t-1}, Y_{t-1}$, and by the current covariate value X_t. Let $\pi^* = \pi^*(\theta, x)$ be the *oracle rule*, which at each time t prescribes

$$\pi_t^* := \pi^*(\theta, X_t) = \operatorname*{argmax}_{i=1,2}\{\alpha_i + \beta_i X_t\}, \quad t = 1, 2, \ldots. \tag{3}$$

The inferior sampling rate of a policy π is given by $T_{\inf}(n) = \sum_{t=1}^n I\{\pi_t \neq \pi_t^*\}$, where $I\{\cdot\}$ is the indicator function. Let Θ be a set of parameter values. We will measure quality of a policy π by its maximal expected inferior sampling rate over Θ,

$$S_n(\pi, \Theta) := \sup_{\theta \in \Theta} \mathbb{E}_\theta^{\pi, n}\, [T_{\inf}(n)],$$

where $\mathbb{E}_\theta^{\pi,t}$ denotes expectation with respect to the distribution $\mathbb{P}_\theta^{\pi,t}$ of the observations $\mathcal{Y}_t = (X_1, Y_1, \ldots, X_{t-1}, Y_{t-1}, X_t)$ from the model (2) associated with configuration θ and policy π. The minimax expected inferior sampling rate is defined by

$$S_n^*(\Theta) = \inf_\pi S_n(\pi, \Theta),$$

where $\inf$ is taken over all possible policies $\pi = \{\pi_t\}$. We next study the behavior of $S_n^*(\Theta)$ for a natural choice of parameter set Θ.

B. A lower bound on the inferior sampling rate.

In order to state our first result we introduce the following notation and definitions.

Let $\theta = (\theta_1, \theta_2)$ be a configuration, and let η be a positive real number. We say that parameters of the arms $\theta_1, \theta_2 \in \mathbb{R}^2$ are η–separated if $\|\theta_1 - \theta_2\| \geq \eta$ where $\|\cdot\|$ is the Euclidean norm. Similarly, if θ and θ' are two different configurations then they are η–separated if $\|\theta - \theta'\| \geq \eta$. With each configuration θ we associate the real number $\tau(\theta) = (\alpha_2 - \alpha_1)/(\beta_1 - \beta_2)$ which is the x–coordinate of the intersection point of the two regression lines defined in (2).

(A) $\{X_t\}$ are i.i.d. random variables taking values ± 1 with probability $1/2$.

Theorem 1: Let Assumption (A) hold, and let

$$\Theta_\eta := \{\theta = (\theta_1, \theta_2) : \tau(\theta) = 0,\ \|\theta_1 - \theta_2\| \geq \eta\}.$$

Then for all n large enough

$$S_n^*(\Theta_\eta) \geq C\sigma^2\eta^{-2}\ln n, \tag{4}$$

where C is an absolute constant.

Remarks:

1) By definiton of Θ_η, for any configuration $\theta = (\theta_1, \theta_2) \in \Theta_\eta$, θ_1 and θ_2 are η–separated. Futhemore, condition $\tau(\theta) = 0$ along with Assumption (A) ensures that the best arm depends on X_t for all $\theta \in \Theta_\eta$.
2) The method of proof we employ does not rely on the change-of-measure argument that was first introduced by Lai and Robbins (1985), and since then adopted by many other subsequent papers.
3) The proof of Theorem 1 shows that the lower bound (4) already holds for $S_n^*\big(\{\theta^{(0)}, \theta^{(1)}\}\big)$ where the worst–case configurations $\theta^{(0)}, \theta^{(1)} \in \Theta_\eta$ are independent of n, and given by $\theta^{(0)} = (0, 0, 0, \eta)$ and $\theta^{(1)} = (\eta, \eta, \eta, 0)$.

The last remark leads to the following straightforward result.

Corollary 1: Let Assumption (A) hold; then for any policy π there exists a configuration $\theta \in \{\theta^{(0)}, \theta^{(1)}\}$ such that

$$\limsup_{n\to\infty} \frac{\mathbb{E}_\theta^{\pi,n}\left[T_{\inf}(n)\right]}{\ln n} > 0.$$

It is worth noting that in the setting where the covariate distribution P_X is discrete, Wang, Kulkarni and Poor (2005) showed that in certain cases the regret can be finite. The above results indicate that even in the simplest instances of this problem the regret *cannot* be bounded unless a suitable assumption on the variability or diversity of the covariates is imposed.

C. Relation to traditional bandit problems.

The order of the lower bound in Theorem 1 is identical to that of Lai and Robbins (1985, Theorem 2), derived in the traditional bandit problem without side information. An obvious question that arises in the context of Theorem 1 is whether the bound stated there is achievable. To that end, imagine that each of the values of the covariate indexes a distinct and independent bandit machine. Consider the policy $\hat{\pi}$ that for the sequence of times in which $X_t = 1$ follows the strategy given in Lai and Robbins (1985, Section 3), ignoring all actions and observations associated with the sequence of times in which the covariate value is $X_t = -1$, and vice versa. It then follows straightforwardly from Lai and Robbins (1985, Theorem 3) that this policy achieves

$$S_n(\hat{\pi}, \Theta_\eta) \leq C'\sigma^2\eta^{-2}\ln n,$$

where C' is an absolute constant. That is, the policy that decouples the problem into two independent traditional bandit problems is optimal up to a constant factor. Thus, if one views

the complexity of bandit problems as being characterized by the expected inferior sampling rate, then the problem we described above is essentially equivalent to traditional bandit problems without side information.

D. The role of persistent excitation.

The lower bound of Theorem 1 is a manifestation of a general fact: the expected inferior sampling rate cannot be bounded as n tends to infinity unless an assumption on suitable variability or "diversity" in the values of the covariates is imposed. To spell this out in mathematical terms, define first for a given policy π, the subset $J_i^\pi(t)$ of the set of indices $\{1,\dots,t-1\}$ when the policy π selects the i–th arm, $J_i^\pi(t) = \{1 \le s < t : \pi_s = i\}$.

Definition 1: **(persistence of excitation)** We say that a policy π does not induce persistent excitation for a configuration θ, if there exists a positive constant $K < \infty$ such that for $i = 1$ or $i = 2$ and for all t

$$\mathbb{P}_\theta^{\pi,t}\Big\{\lambda_{\min}\Big(\sum_{s\in J_i^\pi(t)} Z_s Z_s^T\Big) \le K\Big\} = 1, \tag{5}$$

where $\lambda_{\min}(\cdot)$ is the minimal eigenvalue of a matrix, and $Z_s = (1, X_s)^T$. By convention $\sum_\emptyset = 0$.

Remarks:

1) Our definition of persistent excitation pertains to the linear regression model (2), but it can be easily adapted to other settings. The absence of persistent excitation in the sense of Definition 1 means that the Fisher information about arm parameters does not grow when observations are sampled according the policy π.
2) The setting of Theorem 1 provides a concrete example of absence of persistent excitation. In particular, since there is only a single point mass within the set in which each arm is superior, the oracle policy satisfies (5) for any configuration $\theta \in \Theta_\eta$.

Armed with this notion, we state the following result.

Theorem 2: Suppose that X_t are non–degenerate i.i.d. random variables, $|X_t| \le r$, $\forall t$. Let π^* be the oracle rule, and assume that π^* does not induce persistent excitation for some configuration $\theta^{(0)} = (\theta_1, \theta_2)$ such that $\|\theta_1 - \theta_2\| = \eta$ for some $\eta > 0$. Then there exists a configuration $\theta^{(1)}$ satisfying $\|\theta^{(0)} - \theta^{(1)}\| = C_1\eta$ such that for all sufficiently large n

$$S_n^*(\{\theta^{(0)}, \theta^{(1)}\}) \ge C_2\sigma^2\eta^{-2}\,\ln n,$$

where constants C_1 and C_2 depend on r and P_X only.

Remarks:

1) Theorem 2 gives a *necessary* condition for boundedness of the expected inferior sampling rate. In particular, if the oracle rule π^* does not induce persistent excitation, then the expected inferior sampling rate is unbounded.
2) Note that in the traditional bandit setting of Lai and Robbins (1985), the oracle rule trivially does not induce persistent excitation in the sense of (5). Hence the growth of the expected inferior sampling rate can be inferred from the absence of this property.
3) The exact characterization of constants C_1 and C_2 is given the proof.

III. PROOFS

A. *The key lemma.*

The following general result plays a central role in our derivation of lower bounds on the expected inferior sampling rate.

Lemma 1: Let Θ be a parameter set, and let $\theta^{(0)}, \theta^{(1)} \in \Theta$ be a pair of configurations such that

$$\pi^*(\theta^{(0)}, x) \neq \pi^*(\theta^{(1)}, x), \quad \forall x \in \mathrm{supp}(P_X) \tag{6}$$

Let

$$\mathcal{K}\big(\mathbb{P}^{\pi,t}_{\theta^{(0)}}, \mathbb{P}^{\pi,t}_{\theta^{(1)}}\big) = \mathbb{E}^{\pi,t}_{\theta^{(0)}}\left[\ln \frac{\mathbb{P}^{\pi,t}_{\theta^{(0)}}}{\mathbb{P}^{\pi,t}_{\theta^{(1)}}}(\mathcal{Y}_t)\right]$$

be the Kullback–Leibler divergence between distributions of observation $\mathcal{Y}_t$ when $\theta = \theta^{(0)}$ and $\theta = \theta^{(1)}$, and policy π is applied. Then for arbitrary policy π and all n one has

$$S_n(\pi, \Theta) \geq \frac{1}{4}\sum_{t=1}^{n} \exp\Big\{ - \mathcal{K}\big(\mathbb{P}^{\pi,t}_{\theta^{(0)}}, \mathbb{P}^{\pi,t}_{\theta^{(1)}}\big)\Big\}. \tag{7}$$

Remarks:

1) The condition that the parameter set Θ contains a pair of configurations $(\theta^{(0)}, \theta^{(1)})$ satisfying (6) ensures that the preference of an arm can be changed at every $x \in \mathrm{supp}(P_X)$ by the choice of a configuration.
2) It is worth pointing out that lower bounds on $S_n(\pi, \Theta)$ can be established also in terms of other distance measures; see Tsybakov (2009, Chapter 2).

Proof of Lemma 1: The proof is based on a standard reduction to a hypothesis testing problem.

Let π^* be the oracle rule. For any two configurations $\theta^{(0)}$ and $\theta^{(1)}$ from Θ and any fixed policy π we have

$$\begin{aligned} S_n(\pi,\Theta) &= \sup_{\theta\in\Theta}\mathbb{E}_\theta^{\pi,n}\left[T_{\inf}(n)\right] \\ &= \sup_{\theta\in\Theta}\sum_{t=1}^{n}\mathbb{P}_\theta^{\pi,t}(\pi_t\neq\pi_t^*) \\ &\geq \frac{1}{2}\sum_{t=1}^{n}\left[\mathbb{P}_{\theta^{(0)}}^{\pi,t}(\pi_t\neq\pi_t^*)+\mathbb{P}_{\theta^{(1)}}^{\pi,t}(\pi_t\neq\pi_t^*)\right]. \end{aligned} \tag{8}$$

Fix $t=1,\ldots,n$, and consider the problem of testing the hypothesis

$$H_0:\theta=\theta^{(0)}=(\alpha_1,\beta_1,\alpha_2,\beta_2) \quad \text{versus} \quad H_1:\theta=\theta^{(1)}=(\alpha_1',\beta_1',\alpha_2',\beta_2')$$

from observations $\mathcal{Y}_t$ collected under the policy π. Define the event $D_t=\{\omega:\alpha_1+\beta_1X_t\geq\alpha_2+\beta_2X_t\}$ and consider the following $\mathcal{F}_{t-1}$–measurable random variable

$$p_t=p_t(\omega)=\begin{cases}1, & \omega\in D_t\\ 2, & \omega\notin D_t.\end{cases}$$

By definition $p_t=\pi^*(\theta^{(0)},X_t)$, or, in other words, $p_t=\pi_t^*$ under hypothesis H_0. At the same time, it follows from (6) that $p_t\neq\pi^*(\theta^{(1)},X_t)$, or, equivalently, $p_t\neq\pi_t^*$ under H_1.

Now consider the following test $\psi_t=I(\pi_t\neq p_t)$. The meaning of the event $(\psi_t=1)$ is that H_0 is rejected while $(\psi_t=0)$ means that H_0 is accepted. The error probabilities of the test ψ_t are the following

$$\begin{aligned} \mathbb{P}_{\theta^{(0)}}^{\pi,t}(\psi_t=1)&=\mathbb{P}_{\theta^{(0)}}^{\pi,t}(\pi_t\neq p_t)=\mathbb{P}_{\theta^{(0)}}^{\pi,t}(\pi_t\neq\pi_t^*)\\ \mathbb{P}_{\theta^{(1)}}^{\pi,t}(\psi_t=0)&=\mathbb{P}_{\theta^{(1)}}^{\pi,t}(\pi_t=p_t)=\mathbb{P}_{\theta^{(1)}}^{\pi,t}(\pi_t\neq\pi_t^*). \end{aligned}$$

Therefore

$$\mathbb{P}_{\theta^{(0)}}^{\pi,t}(\pi_t\neq\pi_t^*)+\mathbb{P}_{\theta^{(1)}}^{\pi,t}(\pi_t\neq\pi_t^*)=\mathbb{P}_{\theta^{(0)}}^{\pi,t}(\psi_t=1)+\mathbb{P}_{\theta^{(1)}}^{\pi,t}(\psi_t=0). \tag{9}$$

Using the lower bound in terms of the Kullback–Leibler divergence on the sum of error probabilities in hypotheses testing [see, e.g., Tsybakov (2009, Theorem 2.2)] we obtain that

$$\mathbb{P}_{\theta^{(0)}}^{\pi,t}(\psi_t=1)+\mathbb{P}_{\theta^{(1)}}^{\pi,t}(\psi_t=0)\geq\frac{1}{2}\exp\Big\{-\mathcal{K}(\mathbb{P}_{\theta^{(0)}}^{\pi,t},\mathbb{P}_{\theta^{(1)}}^{\pi,t})\Big\}.$$

Combining this inequality with (8) and (9) we complete the proof. ∎

B. Proof of Theorem 1

Let us introduce the following notation. For fixed π we let $J_i^\pi(t)$ be the subset of indices from $\{1,2,\dots,t-1\}$ when the policy π selects i-th arm. Let $T_i(t)$ denote the cardinality of $J_i^\pi(t)$.

Let $\eta>0$, and consider the following two configurations

$$\theta^{(0)}=(\theta_1,\theta_2)=(\alpha_1,\beta_1,\alpha_2,\beta_2)=(0,0,0,\eta)$$

$$\theta^{(1)}=(\theta_1',\theta_2')=(\alpha_1',\beta_1',\alpha_2',\beta_2')=(\eta,\eta,\eta,0).$$

Clearly, $\|\theta_1-\theta_2\|=\eta$, $\|\theta_1'-\theta_2'\|=\eta$, and $\tau(\theta^{(0)})=\tau(\theta^{(1)})=0$; hence $\theta^{(0)},\theta^{(1)}\in\Theta_\eta$. Note also that under configuration $\theta^{(0)}$ the first arm is superior when $X_t=-1$ and inferior when $X_t=1$. The preference of arms is changed when $\theta=\theta^{(1)}$: now the first arm is superior when $X_t=1$ and inferior when $X_t=-1$. This means that $\pi^*(\theta^{(0)},x)\neq\pi^*(\theta^{(1)},x)$ for any $x\in\{-1,1\}$, so that condition (6) is fulfilled, and Lemma 1 can be applied.

For fixed t we have

$$\begin{aligned}
\ln\frac{\mathrm{d}\mathbb{P}^{\pi,t}_{\theta^{(0)}}}{\mathrm{d}\mathbb{P}^{\pi,t}_{\theta^{(1)}}}(\mathcal{Y}_t) &= -\frac{1}{2\sigma^2}\sum_{s\in J_1^\pi(t)}Y_s^2-\frac{1}{2\sigma^2}\sum_{s\in J_2^\pi(t)}(Y_s-\eta X_s)^2\\
&\quad+\frac{1}{2\sigma^2}\sum_{s\in J_1^\pi(t)}(Y_s-\eta-\eta X_s)^2+\frac{1}{2\sigma^2}\sum_{s\in J_2^\pi(t)}(Y_s-\eta)^2\\
&= -\frac{1}{2\sigma^2}\sum_{s\in J_1^\pi(t)}\eta(1+X_s)[2Y_s-\eta(1+X_s)]\\
&\quad-\frac{1}{2\sigma^2}\sum_{s\in J_2^\pi(t)}\eta(1-X_s)(2Y_s-\eta X_s-\eta).
\end{aligned}$$

Therefore

$$\begin{aligned}
\mathcal{K}(\mathbb{P}^{\pi,t}_{\theta^{(0)}},\mathbb{P}^{\pi,t}_{\theta^{(1)}}) &= -\frac{1}{2\sigma^2}\mathbb{E}^{\pi,t}_{\theta^{(0)}}\Big\{\sum_{s\in J_1^\pi(t)}\eta(1+X_s)[2\varepsilon_s^{(1)}-\eta(1+X_s)]\Big\}\\
&\quad-\frac{1}{2\sigma^2}\mathbb{E}^{\pi,t}_{\theta^{(0)}}\Big\{\sum_{s\in J_2^\pi(t)}\eta(1-X_s)[2\varepsilon_s^{(2)}-\eta(1-X_s)]\Big\}\\
&= \frac{\eta^2}{2\sigma^2}\mathbb{E}^{\pi,t}_{\theta^{(0)}}\Big\{\sum_{s\in J_1^\pi(t)}(1+X_s)^2\Big\}+\frac{\eta^2}{2\sigma^2}\mathbb{E}^{\pi,t}_{\theta^{(0)}}\Big\{\sum_{s\in J_2^\pi(t)}(1-X_s)^2\Big\},
\end{aligned}\tag{10}$$

where the last equality follows because

$$\begin{aligned}
\mathbb{E}^{\pi,t}_{\theta^{(0)}} \sum_{s\in J_i^\pi(t)} \varepsilon_s^{(i)} &= \mathbb{E}^{\pi,t}_{\theta^{(0)}} \sum_{s=1}^{t-1} \mathbb{E}^{\pi,t}_{\theta^{(0)}} \Big[\varepsilon_s^{(i)} I(\pi_s=2)|\mathcal{F}_{s-1}\Big] \\
&= \mathbb{E}^{\pi,t}_{\theta^{(0)}} \sum_{s=1}^{t-1} I(\pi_s=2)\mathbb{E}^{\pi,t}_{\theta^{(0)}} \big[\varepsilon_s^{(i)}\big] = 0 \\
\mathbb{E}^{\pi,t}_{\theta^{(0)}} \sum_{s\in J_i^\pi(t)} \varepsilon_s^{(i)} X_s &= \mathbb{E}^{\pi,t}_{\theta^{(0)}} \sum_{s=1}^{t-1} \mathbb{E}^{\pi,t}_{\theta^{(0)}} \Big[\varepsilon_s^{(i)} X_s I(\pi_s=2)|\mathcal{F}_{s-1}\Big] \\
&= \mathbb{E}^{\pi,t}_{\theta^{(0)}} \sum_{s=1}^{t-1} X_s I(\pi_s=2)\mathbb{E}^{\pi,t}_{\theta^{(0)}} \big[\varepsilon_s^{(i)}\big] = 0.
\end{aligned}$$

Let $T_i^{(x)}(t)$ denote the number of times the i-th arm was pulled up until time t when $X_s = x$, i.e., $T_i^{(x)}(t) = \sum_{s=1}^{t-1} I(\pi_s = i, X_s = x)$. With this notation the expression on the right hand side of (10) takes the form

$$\begin{aligned}
\mathcal{K}(\mathbb{P}^{\pi,t}_{\theta^{(0)}}, \mathbb{P}^{\pi,t}_{\theta^{(1)}}) &= \frac{\eta^2}{2\sigma^2}\mathbb{E}^{\pi,t}_{\theta^{(0)}}\Big\{ \sum_{s\in J_1^\pi(t)\cap\{s:X_s=1\}} (1+X_s)^2\Big\} \\
&\quad + \frac{\eta^2}{2\sigma^2}\mathbb{E}^{\pi,t}_{\theta^{(0)}}\Big\{ \sum_{s\in J_2^\pi(t)\cap\{s:X_s=-1\}} (1-X_s)^2\Big\} \\
&= \frac{2\eta^2}{\sigma^2}\mathbb{E}^{\pi,t}_{\theta^{(0)}}\big[T_1^{(1)}(t) + T_2^{(-1)}(t)\big] = \frac{2\eta^2}{\sigma^2}\mathbb{E}^{\pi,t}_{\theta^{(0)}}\big[T_{\inf}(t)\big], \qquad (11)
\end{aligned}$$

where the last equality is a consequence of the fact that under configuration $\theta^{(0)}$ the first arm is inferior at $x = 1$ and the second arm is inferior at $x = -1$.

Combining this result with (7) we obtain

$$\begin{aligned}
S_n &\geq \frac{1}{4}\sum_{t=1}^{n} \exp\Big\{ -\frac{2\eta^2}{\sigma^2}\ \mathbb{E}^{\pi,t}_{\theta^{(0)}}[T_{\inf}(t)]\Big\} \\
&\geq \frac{1}{4}\sum_{t=1}^{n} \exp\Big\{ -\frac{2\eta^2}{\sigma^2}\sup_{\theta\in\Theta}\mathbb{E}^{\pi,t}_{\theta}[T_{\inf}(t)]\Big\} \\
&= \frac{1}{4}\sum_{t=1}^{n} \exp\Big\{ -\frac{2\eta^2}{\sigma^2}S_t\Big\}, \\
&\geq \frac{n}{4}\exp\Big\{ -\frac{2\eta^2}{\sigma^2}S_n\Big\} \qquad (12)
\end{aligned}$$

where for brevity we write $S_t := S_t(\pi, \Theta)$, and the last step follows since $\{S_t\}$ is a non-decreasing sequence. Note that (12) holds for any policy π and for all n. Then the assertion of the theorem follows from the fact that the inequality (12) is fulfilled only if for any $\varepsilon \in (0,1)$ the numerical sequence $\{S_n\}$ satisfies for all n large enough $S_n \geq \frac{1}{2}\sigma^2\eta^{-2}(1-\varepsilon)\ln n$. ∎

C. Proof of Theorem 2

For a given policy π let us denote

$$Q_i^{\pi}(t) = \sum_{s \in J_i^{\pi}(t)} Z_s Z_s^T, \quad i = 1, 2.$$

By the premise of the theorem the oracle rule π^* does not induce persistent excitation for a configuration $\theta^{(0)} = (\theta_1, \theta_2) = (\alpha_1, \beta_1, \alpha_2, \beta_2)$; hence (5) holds with θ replaced by $\theta^{(0)}$ for $i = 1$ or $i = 2$. Without loss of generality assume (5) holds only for the second arm: for all t

$$\lambda_{\min}\{Q_2^{\pi^*}(t)\} \leq K, \quad \mathbb{P}_{\theta^{(0)}}^{\pi^*, t} - \text{a.s.} \tag{13}$$

It is easily seen from Step 1 below that if (5) holds for both arms, then the minimax lower bound can be constructed as in the proof of Theorem 1. For the sake of definiteness we assume also that $\beta_2 > 0$.

The proof proceeds in two steps. First, we show that under conditions of the theorem the following statement holds: either (i) $Q_2^{\pi^*}(t) = O$; or (ii) $\lambda_{\min}\{Q_2^{\pi^*}(t)\} = 0$ for all t (here O is 2×2 zero matrix). Second, we construct worst–case configurations and apply Lemma 1.

Step 1: Assume that (13) holds. Because X_t, $t = 1, \ldots . n$ are i.i.d. random variables, the following two situations are possible:

(i) $\mathbb{P}(\theta_2^T Z_s \geq \theta_1^T Z_s) = 0$, i.e., the second arm is never pulled under the oracle rule π^*;

(ii) $\mathbb{P}(\theta_2^T Z_s \geq \theta_1^T Z_s) > 0$, i.e., under the oracle rule π^* the second arm is pulled a number of times that tends to infinity as $t \to \infty$, but the absense of excitation is caused by insufficient diversity of values of X_t allocated to the second arm.

In the case (i) the second arm is inferior for all x, $J_2^{\pi^*}(t) = \emptyset$, and hence $Q_2^{\pi^*}(t) = O$ for all t.

Now we consider the case (ii). Write for brevity $\tilde{X}_s = X_s I(\theta_2^T Z_s \geq \theta_1^T Z_s)$ and note that $\tilde{X}_s = X_s$ for those indices s where the oracle rule π^* selects the second arm. Write also $\tilde{T}_2(t) = \sum_{s=1}^{t-1} I(\theta_2^T Z_s \geq \theta_1^T Z_s)$. We have $\tilde{T}_2(t) \to \infty$ as $t \to \infty$, and since X_t are i.i.d. random variables by the strong law of large numbers

$$\frac{\tilde{T}_2(t)}{t} \overset{\text{a.s.}}{\to} \mathbb{P}(\theta_2^T Z_s \geq \theta_1^T Z_s), \quad \text{as } t \to \infty.$$

With introduced notation one has

$$Q_2^{\pi^*}(t) = \begin{bmatrix} \tilde{T}_2(t) & \sum_{s=1}^{t-1} \tilde{X}_s \\ \sum_{s=1}^{t-1} \tilde{X}_s & \sum_{s=1}^{t-1} \tilde{X}_s^2 \end{bmatrix}, \quad \forall t.$$

A straightforward calculation shows that

$$\lambda_{\min}\{Q_2^{\pi^*}(t)\} = \frac{1}{2}\Big\{ \sum_{s=1}^{t-1} \tilde{X}_s^2 + \tilde{T}_2(t) - \Big[\Big(\sum_{s=1}^{t-1} \tilde{X}_s^2 + \tilde{T}_2(t)\Big)^2 - 4\tilde{T}_2(t) \sum_{s=1}^{t-1} \tilde{X}_s^2 + 4\Big(\sum_{s=1}^{t-1} \tilde{X}_s\Big)^2\Big]^{1/2}\Big\}$$

which together with (13) implies that for any t

$$0 \le \frac{1}{\tilde{T}_2(t)} \sum_{s=1}^{t-1} \tilde{X}_s^2 - \Big(\frac{1}{\tilde{T}_2(t)} \sum_{s=1}^{t-1} \tilde{X}_s \Big)^2 \le \frac{4K}{\tilde{T}_2(t)} \Big(\frac{1}{\tilde{T}_2(t)} \sum_{s=1}^{t-1} \tilde{X}_s^2 + 1 \Big).$$

Letting $t \to \infty$ and using the strong law of large number we obtain that the expression on the right hand side tends to 0 while the left hand side converges to

$$\frac{\mathbb{E}\tilde{X}_s^2}{\mathbb{P}(\theta_2^T Z_s \ge \theta_1^T Z_s)} - \left[\frac{\mathbb{E}\tilde{X}_s}{\mathbb{P}(\theta_2^T Z_s \ge \theta_1^T Z_s)} \right]^2$$

almost surely. Hence

$$\mathbb{E}\big(X_s I\{\theta_2^T Z_s \ge \theta_1^T Z_s) \big)^2 = \frac{\big[\mathbb{E} X_s I\{\theta_2^T Z_s \ge \theta_1^T Z_s\} \big]^2}{\mathbb{P}\{\theta_2^T Z_s \ge \theta_1^T Z_s\}}. \tag{14}$$

Note that the Cauchy–Schwarz inequality implies that the left hand side of (14) is always greater than or equal to the right hand side. The equality in (14) is possible only if for some constant x_0, $X_s(\omega) = x_0$ on the event $\{\omega : \theta_2^T Z_s(\omega) \ge \theta_1^T Z_s(\omega)\}$. This shows that if (13) holds and X_t's are i.i.d. random variables then in case (ii) the distribution of X_t has only a single atom x_0 in the set $\{x : \alpha_2 + \beta_2 x \ge \alpha_1 + \beta_1 x\}$. Under these circumstances

$$Q_2^{\pi^*}(t) = \tilde{T}_2(t) \begin{bmatrix} 1 & x_0 \\ x_0 & x_0^2 \end{bmatrix}, \quad \forall t, \tag{15}$$

and $\lambda_{\min}\{Q_2^{\pi^*}(t)\} = 0$, $\forall t$. Moreover, since the distribution of X_t is non-degenerate, the two lines intersect in the interval $[-r, r]$. Recall that $\beta_2 > 0$. The following two cases are possible: $\beta_2 > \beta_1$ and $\beta_2 < \beta_1$. If $\beta_2 > \beta_1$ then arm 2 is inferior at any point $x \in [-r, x_0) \cap \mathrm{supp}(P_X)$. In addition, because (5) does not hold for arm 1, $\mathbb{P}\{X_t \in (-r, x_0)\} > 0$. If $\beta_2 < \beta_1$ then arm 2 is inferior at any point $x \in (x_0, r]$ and a similar conclusion holds for that interval. For the sake of definiteness we suppose that $\beta_2 > \beta_1$; the proof for the case of $\beta_2 < \beta_1$ goes along the same lines.

Step 2: Let π be an arbitrary policy. Using the defintion of $Q_i^\pi(t)$, we can write

$$\begin{aligned} Q_i^\pi(t) &= \sum_{s \in J_i^\pi(t) \cap J_i^{\pi^*}(t)} Z_s Z_s^T + \sum_{s \in J_i^\pi(t) \setminus J_i^{\pi^*}(t)} Z_s Z_s^T \\ &=: G_i(t) + B_i(t). \end{aligned}$$

In words, $G_i(t)$ is the design matrix corresponding to the pulls of the i-th arm when it is superior, while $B_i(t)$ is the design matrix corresponding to the inferior pulls of the i-th arm. Let $T_{\mathrm{inf},i}(t)$ denote the number of pulls of i–th arm up until time t when it is inferior; then the inferior sampling rate is $T_{\mathrm{inf}}(t) = T_{\mathrm{inf},1}(t) + T_{\mathrm{inf},2}(t)$.

Let $\theta^{(1)} = (\theta_1', \theta_2')$ be another configuration that will be specified later. Using previously introduced notation, we can write

$$\begin{aligned}\mathcal{K}(\mathbb{P}^{\pi,t}_{\theta^{(0)}}, \mathbb{P}^{\pi,t}_{\theta^{(1)}}) &= \frac{1}{2\sigma^2}\mathbb{E}^{\pi,t}_{\theta^{(0)}}\Big\{(\theta_1-\theta_1')^T Q_1^{\pi}(t)(\theta_1-\theta_1') + (\theta_2-\theta_2')^T Q_2^{\pi}(t)(\theta_2-\theta_2')\Big\} \\ &= \frac{1}{2\sigma^2}\mathbb{E}^{\pi,t}_{\theta^{(0)}}\Big\{(\theta_1-\theta_1')^T G_1(t)(\theta_1-\theta_1') + (\theta_1-\theta_1')^T B_1(t)(\theta_1-\theta_1') \\ &\quad + (\theta_2-\theta_2')^T G_2(t)(\theta_2-\theta_2') + (\theta_2-\theta_2')^T B_2(t)(\theta_2-\theta_2')\Big\}. \qquad (16)\end{aligned}$$

Because $J_2^{\pi}(t) \cap J_2^{\pi^*}(t) \subseteq J_2^{\pi^*}(t)$ we have that $Q_2^{\pi^*}(t) - G_2(t)$ is a non–negative definite matrix for all t. Therefore

$$(\theta_2-\theta_2')^T G_2(t)(\theta_2-\theta_2') \le (\theta_2-\theta_2')^T Q_2^{\pi^*}(t)(\theta_2-\theta_2') \quad \text{a.s.} \qquad (17)$$

Now we are in a position to specify the worst–case configuration $\theta^{(1)} = (\theta_1', \theta_2')$. We always choose $\theta_1' = \theta_1$. The choice of θ_2' is the following.

Case (i): Here we take $\theta_2' = (\alpha_2', \beta_2') = (\alpha_2 + 2\eta(r \vee 1), \beta_2)$. Because arm 2 is inferior at all $x \in \mathrm{supp}(P_X)$ and $\|\theta_1 - \theta_2\| = \eta$, the above choice of θ_2' ensures that the preference of arms is switched at all $x \in \mathrm{supp}(P_X)$ for the configuration $\theta^{(1)}$, i.e., $\pi^*(\theta^{(0)}, x) \neq \pi^*(\theta^{(1)}, x)$, $\forall x \in \mathrm{supp}(P_X)$. Hence Lemma 1 can be applied.

Case (ii): We consider the following construction. Let $e = (e_1, e_2)^T$ be a unit vector such that

$$e_1 + x_0 e_2 = 0. \qquad (18)$$

In view of (15), e belongs to the null space of the matrices $Q_2^{\pi^*}(t)$, $\forall t$, and $e = e^{(1)}$ or $e = e^{(2)}$ where

$$e^{(1)} := \left(-\frac{x_0}{\sqrt{1+x_0^2}}, \frac{1}{\sqrt{1+x_0^2}}\right)^T, \quad e^{(2)} := \left(\frac{x_0}{\sqrt{1+x_0^2}}, -\frac{1}{\sqrt{1+x_0^2}}\right)^T. \qquad (19)$$

For $\gamma > 0$ put

$$\theta_2' = \theta_2 + \gamma e = (\alpha_2 + \gamma e_1, \beta_2 + \gamma e_2). \qquad (20)$$

The idea is to choose $\gamma > 0$ and $e \in \{e^{(1)}, e^{(2)}\}$ (and configuration θ_2' via (20)) so that the preferences of arms left of point x_0 are switched. For this purpose it is sufficient to equate the two regression lines under configuration $\theta^{(1)}$ at the point $x = -r$.

First we recall that the second arm is inferior at any $x \in [-r, x_0)$ under $\theta^{(0)}$; in particular, $\alpha_2 - \beta_2 r < \alpha_1 - \beta_1 r$. In order to have $\alpha_2' - \beta_2' r = \alpha_1 - \beta_1 r$ we should choose $e_2 < 0$ because, by (18) and (20),

$$\alpha_2' - \beta_2' r = \alpha_2 - \beta_2 r - \gamma e_2 (x_0 + r).$$

Therefore we put $e = e^{(2)}$. Solving the system of equations

$$\alpha_2' + \beta_2' x_0 = \alpha_2 + \beta_2 x_0$$
$$\alpha_2' - \beta_2' r = \alpha_1 - \beta_1 r.$$

with respect to α_2' and β_2' we obtain

$$\beta_2 - \beta_2' = \frac{\alpha_2 - \alpha_1 + (\beta_1 - \beta_2)r}{x_0 + r}, \quad \alpha_2' - \alpha_2 = (\beta_2 - \beta_2')x_0,$$

which, in turn, yields

$$\gamma^2 = \|\theta_2 - \theta_2'\|^2 = \frac{1 + x_0^2}{(x_0 + r)^2}\Big[\alpha_2 - \alpha_1 + (\beta_1 - \beta_2)r\Big]^2. \tag{21}$$

Taking into account that $\|\theta_1 - \theta_2\| = \eta$ we have the following straightforward bounds on γ

$$\frac{(r \wedge 1)\eta\sqrt{1 + x_0^2}}{x_0 + r} \le \gamma \le \frac{\sqrt{2}(r \vee 1)\eta\sqrt{1 + x_0^2}}{x_0 + r}. \tag{22}$$

Thus, the choice $\theta_2' = \theta_2 + \gamma e$ with $e = e^{(2)}$ [see (19)] and γ given in (21) ensures that arm 2 is no longer inferior at any $x \in [-r, x_0)$. Since, by Step 1, $m := \mathbb{P}\{X_t \in (-r, x_0)\} > 0$, under configuration $\theta^{(1)}$ preference of the arms is changed for all $x \in A := (-r, x_0) \cap \mathrm{supp}(P_X)$. Then, restricting steps (8) and (9) to the event A, a straightforward calculation shows that the result of Lemma 1 holds with a factor m multiplying the RHS in (7).

Note that in both cases by construction $(\theta_2 - \theta_2')^T Q_2^{\pi^*}(t)(\theta_2 - \theta_2') = 0$, $\forall t$. Therefore using (17) we have the following upper bound on the expression under the expectation sign on the right hand side of (16)

$$\begin{aligned}
&(\theta_1 - \theta_1')^T Q_1^{\pi}(t)(\theta_1 - \theta_1') + (\theta_2 - \theta_2')^T Q_2^{\pi}(t)(\theta_2 - \theta_2') \\
&\quad\le\ (\theta_2 - \theta_2')^T G_2(t)(\theta_2 - \theta_2') + (\theta_2 - \theta_2')^T B_2(t)(\theta_2 - \theta_2') \\
&\quad\le\ (\theta_2 - \theta_2')^T Q_2^{\pi^*}(t)(\theta_2 - \theta_2') + (\theta_2 - \theta_2')^T B_2(t)(\theta_2 - \theta_2') \\
&\quad=\ (\theta_2 - \theta_2')^T B_2(t)(\theta_2 - \theta_2') \\
&\quad\le\ \gamma^2(1 + r^2)T_{\mathrm{inf},2}(t),
\end{aligned}$$

where the last inequality follows from the fact that $\|\theta_2 - \theta_2'\| = \gamma$, and

$$\lambda_{\max}\{B_2(t)\} \le \mathrm{tr}\{B_2(t)\} = \sum_{s \in J_2^{\pi}(t) \setminus J_2^{\pi^*}(t)} Z_s^T Z_s \le (1 + r^2)T_{\mathrm{inf},2}(t).$$

We have used also the fact that $|X_t| \le r$, $\forall t$, and the cardinality of the set $J_2^{\pi}(t) \setminus J_2^{\pi^*}(t)$ equals $T_{\mathrm{inf},2}(t)$. Thus, it follows from (16) that

$$\mathcal{K}(\mathbb{P}_{\theta^{(0)}}^{\pi,t}, \mathbb{P}_{\theta^{(1)}}^{\pi,t}) \le \frac{\gamma^2}{2\sigma^2}(1 + r^2)\mathbb{E}_{\theta^{(0)}}^{\pi,t}[T_{\mathrm{inf}}(t)],$$

and therefore by the modification of Lemma 1 discussed above,

$$S_n \geq \frac{m}{4} \sum_{t=1}^{n} \exp\Big\{ - \frac{\gamma^2(1+r^2)}{2\sigma^2} S_t \Big\}.$$

This inequality and (22) yield the announced result. ∎

REFERENCES

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

ANANTHARAM, V., VARAIYA, P. and WARLAND, J. (1987a). Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays–Part I: i.i.d. rewards. *IEEE Trans. Automat. Contr.* **34**, 968–976.

ANANTHARAM, V., VARAIYA, P. and WARLAND, J. (1987b). Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays–Part II: Markovian rewards. *IEEE Trans. Automat. Contr.* **34**, 977–982.

BERRY, D. A. and FRISTEDT, B. (1985). *Bandit Problems.* London, Chapman and Hall.

CLAYTON, M.K. (1989). Covariate models for Bernoulli bandits. *Sequential Anal.* **8** (1989), 405–426.

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 revisted. *Ann. Appl. Probab.* **19**, 1603-1633.

KULKARNI, S. AND LUGOSI, G. (2000). Finite time lower bounds for the two-armed bandit problem. *IEEE Trans. on Automatic Control*, **45**, 711–714.

LAI, T. L. and ROBBINS, H. (1984). Asymptotically optimal allocation of treatments in sequential experiments. In: *Design of experiments,* 127–142, Dekker, New York.

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

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

LAI, T. L. and YAKOWITZ, S. (1995). Machine learning and nonparametric bandit theory. *IEEE Trans. Automatic 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**, editors J.C. Platt, D. Koller, Y. Singer and S. Roweis, 817-824, MIT Press, Cambridge, MA.

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

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

TSYBAKOV, A. (2009). *Introduction to Nonparametric estimation*. Springer, New York.

WANG, C.-C, KULKARNI, S. and POOR, V. H. (2005). Bandit problems with side observations. *IEEE Trans. Automat. Control* **50**, 338-354.

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 banidt problem with covariates. *Annals of Statis.* **30**, 100–121.

Alexander Goldenshluger received the D.Sc. degree in operations research from the Technion–Israel Institute of Technology, Haifa, Israel, in 1996. In 1997 he joined the Department of Statistics, University of Haifa where he is currently an Associate Professor. His research interests include nonparametric statistics, time series and statistical signal processing.

Assaf Zeevi is the Henry Kravis Professor of Business at the Graduate School of Business, Columbia University. He is broadly interested in the formulation and analysis of mathematical models of complex systems. Zeevi received his Ph.D. from Stanford University in 2001, and has been a faculty at Columbia University ever since, while also holding a visiting position at Stanford University. Assaf is currently a departmental editor at Management Science and holds several other editorial positions in INFORMS journals in his areas of expertise. His research interests include stochastic modeling, applied probability, statistics and operations research.