Statistics and Probability Letters 80 (2010) 381–385

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

On the optimal amount of experimentation in sequential decision problems☆

Dinah Rosenberg [b], Eilon Solan [a,*], Nicolas Vieille [b]

[a] The School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel

[b] Département Finance et Economie, HEC, 1, rue de la Libération, 78 351 Jouy-en-Josas, France

ARTICLE INFO

Article history:
Received 12 July 2009
Accepted 18 November 2009
Available online 3 December 2009

MSC:
62C10
60G99
93E35

ABSTRACT

We provide a tight bound on the amount of experimentation under the optimal strategy in sequential decision problems. We show the applicability of the result by providing a bound on the cut-off in a one-arm bandit problem.

© 2009 Elsevier B.V. All rights reserved.

1. Introduction

A basic issue faced by the statistician in sequential decision problems is the trade-off between the cost of pursuing the experimentation and the informational benefit from doing so. For instance, in bandit problems, the decision maker chooses whether to pull an apparently optimal arm, or to pull some seemingly poorer one, in the hope of thereby getting valuable information.

Such problems lead to unwieldy analytical problems, rarely amenable to closed form solutions, which is arguably one reason why sequential methods are still seldom relied upon in practice (see Lai (2001) and Armitage (1975)). For bandit problems, while the optimal strategy is well characterized and consists in pulling the arm with highest dynamic allocation index (Gittins and Jones, 1974; Gittins, 1979), the explicit computation of these indices is rarely feasible, except for very specific cases where the risky arm yields a Bernoulli payoff (see for instance Bradt et al. (1956), Feldman (1962), Woodroofe (1979) and Berry and Fristedt (1985)).

Over the years, a number of approaches have been pursued: (i) computing approximate solutions of the corresponding dynamic programming equation, as in Berry (1972) or Fabius and van Zwet (1970); (ii) relying on close-by problems for which explicit solutions are known, as in Lai (1987); (iii) using extensively numerical computations, as in Lai (1988, for sequential testing of composite hypotheses); (iv) designing *ad hoc* policies, sometimes investigating their performance numerically, as in Cornfield et al. (1969), Berry and Sobel (1973), Berry (1978) and, more recently, (v) finding explicit *a priori* bounds, as in Brezzi and Lai (2000).

This paper contributes to the last category. Motivated by economic applications, (see, e.g. Dixit and Pindyck (1994), Bolton and Harris (1999), Bergemann and Välimäki (2000), Keller et al. (2005) and Rosenberg et al. (2007)), we consider general Bayesian, discounted sequential problems. The parameter θ has an initial distribution $\mathbf{P}$ (the belief of the economic agent). The agent repeatedly receives some information, chooses an action from a set A, and get a possibly unobserved instantaneous

☆ We thank Ehud Lehrer for the discussions we had on the subject.

* Corresponding author. Tel.: +972 528 787275; fax: +972 3 640 9357.
E-mail addresses: dinah@zeus.math.univ-paris13.fr (D. Rosenberg), eilons@post.tau.ac.il (E. Solan), vieille@hec.fr (N. Vieille).

0167-7152/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2009.11.014

reward $u(\theta, a)$. Future gains are discounted by a discount factor $\delta \in (0, 1)$. Given a decision rule σ, and a stage n, we define the *amount of experimentation* in stage n to be the difference Δ_n between the currently highest reward, and the current reward obtained when using σ.

We show that, for every *optimal* decision rule, the expected value of $\sum_{n=1}^{+\infty} \Delta_n$ does not exceed $C\delta/(1-\delta)$, where C is a bound[1] on the reward function u. The bound is valid irrespective of the prior belief $\mathbf{P}$, and no matter how information flows in to the decision maker. This result was used in Rosenberg et al. (2009) to show that the limit payoff of neighbors in connected social networks coincides, and to provide conditions that ensure consensus.

We next show, by means of an example, that this bound is tight. We also illustrate how to use this bound in practice to derive a priori estimates for specific sequential problems. For simplicity, we focus on an instance of a one-arm bandit problem, for which no explicit solution is available, and give an estimate of the optimal boundary in the associated optimal stopping problem. In contrast to Brezzi and Lai (2000), who provide a bound on the Gittins' index in bandit problems, our bound is on the cut-off of the optimal strategy.

2. Setup and results

The parameter set[2] is a measurable space $(\Theta, \mathcal{A})$, endowed with a prior distribution $\mathbf{P}$. At each stage $n \geq 1$, a decision maker first gets an observation drawn from a (measurable) set S, then chooses an action a out of a (compact metric) set A, and gets a reward $u(\theta, a)$. The decision maker discounts future rewards at the rate $\delta \in [0, 1)$. The reward function $u : \Omega \times A \to \mathbf{R}$ is (jointly) measurable, and continuous w.r.t. a. In addition, we assume that the highest reward $\overline{u}\colon \theta \mapsto \max_{a\in A} u(\theta, a)$ and the lowest reward $\underline{u}\colon \theta \mapsto \min_{a\in A} u(\theta, a)$ have finite expectation.

We stress that we place no restriction whatsoever on the nature of observations[3]: e.g., they may depend, possibly in a random way, on the parameter θ, and on past observations and actions; they may or may not reveal past rewards; and they may be independent or not.

Note that we assume that the current reward is a *deterministic* function $u(\omega, a)$ of the parameter ω and of the action a. This assumption is made without loss of generality. Statistical models such as multi-armed bandit problems, where the decision maker observes her current reward that randomly depends on θ (and on a), can be cast into the above framework. Indeed, it suffices to re-label such a random reward as the "observation", and to define the reward to be the expectation of the "observation". Such a change does not affect the optimal decision rules, nor the optimal value of the problem.

For a decision[4]rule σ, $\mathbf{P}_\sigma$ is the joint distribution of θ and of the infinite sequence of observations and decisions. Expectation w.r.t. $\mathbf{P}_\sigma$ is denoted by $\mathbf{E}_\sigma$.

We focus on the amount of experimentation that optimal decisions entail. To be specific, let a decision rule σ be given. Given a stage n, we denote by $\mathcal{H}_n$ the information available at stage n, that is, the σ-field induced by past observations and actions. When using the decision rule σ prior to stage n, the expectation $\mathbf{E}_\sigma[u(\theta, a)|\mathcal{H}_n]$ is the expected reward when choosing a in stage n, given all available information, and $\overline{u}_n := \max_{a\in A} \mathbf{E}_\sigma[u(\theta, a)|\mathcal{H}_n]$ is the myopically optimal reward. Thus, letting a_n denote the action of the decision maker in stage n, $u_n = \mathbf{E}_\sigma[u(\theta, a_n)|\mathcal{H}_n]$ is the actual reward that the decision maker expects to get in stage n, when following σ. The difference $\Delta_n := \overline{u}_n - u_n$ provides a measure of the degree of experimentation performed in stage n. The infinite sum $\sum_{n\geq 1} \Delta_n$ therefore measures the overall amount of experimentation.

Theorem 2.1. *For any optimal decision rule σ, one has*

$$\mathbf{E}_\sigma\left[\sum_{n\geq 1} \Delta_n\right] \leq \left(\mathbf{E}\left[\overline{u}\right] - \mathbf{E}\left[\underline{u}\right]\right) \times \frac{\delta}{(1-\delta)}.$$

Beyond quantitative implications, this bound also yields qualitative implications. Consider for instance a multi-arm bandit problem. For simplicity, assume that the types of the various arms are first drawn, and that each arm then yields a sequence of rewards, which is conditionally i.i.d. given its type. For concreteness, assume that with probability 1 over the types, the expected outputs of the arms are all distinct.

Observe that, whenever the decision maker pulls a specific arm infinitely often, she eventually learns the type of this arm. Therefore, whenever the decision maker pulls *two* specific arms infinitely often, she eventually learns both types. Since one of these two arms is "better" than the other, this implies that the sequence $(\Delta_n)_{n\geq 1}$ then does not converge to zero. By Footnote 1, this event must have probability 0, for every optimal decision rule. In other words: any optimal allocation rule samples finitely often all arms but one. This provides an alternative proof of Theorem 2 in Brezzi and Lai (2000).[5]

We next show that the bound in Theorem 2.1 is tight.

---

1 In particular, $\sum \Delta_n < \infty$ a.s., hence any optimal decision rule eventually stops to experiment.

2 In spite of the qualifier "parameter", our decision problems are non-parametric, since the space Θ is fully general.

3 Beyond the minimal, technical assumption that the observation in stage n is drawn according to a transition probability from $\Theta \times (S \times A)^{n-1}$ to S.

4 That is, a sequence (σ_n) of measurable functions, where $\sigma_n : (S \times A)^{n-1} \times S \to A$ is the decision in stage n.

5 Brezzi and Lai (2000) assumes that the states of the different arms are independent. Our argument dispenses with this assumption.

Proposition 2.2. *For every ε and for every discount factor δ, there is a decision problem with an optimal decision rule σ such that* $\mathbf{E}_\sigma[\sum_{n\geq 1}\Delta_n] \geq (\mathbf{E}[\overline{u}] - \mathbf{E}[\underline{u}]) \times \frac{\delta}{(1-\delta)} \times (1-\varepsilon)$.

The decision problem in Proposition 2.2 depends both on ε and on δ. The next proposition improves in this respect, at a slight cost in the speed of convergence. In this statement, and given $\varepsilon > 0$, we denote by $N(\varepsilon)$ the (random) number of stages in which Δ_n is at least ε: $N(\varepsilon) := |\{n \geq 1 : \Delta_n \geq \varepsilon\}|$. Plainly, $\sum_{n\geq 1}\Delta_n \geq \varepsilon N(\varepsilon)$ for every $\varepsilon > 0$.

Proposition 2.3. *There is a decision problem such that for every $\delta > 2/3$ there is a unique optimal decision rule σ that satisfies*

$$\lim_{\varepsilon\to 0} \varepsilon^\alpha \mathbf{E}_\sigma[N(\varepsilon)] = +\infty, \quad \text{for every } \alpha < 1.$$

That is, as ε decreases, the expected number $\mathbf{E}_\sigma[N(\varepsilon)]$ of experimentation stages increases faster than $1/\varepsilon^\alpha$, for every $\alpha < 1$.

3. Proofs

3.1. Proof of Theorem 2.1

Consider an optimal decision rule σ. Set $Y_n := (1-\delta)\sum_{k=n}^{+\infty}\delta^{k-n}\mathbf{E}_\sigma[u_k \mid \mathcal{H}_n]$: Y_n can be interpreted as the *continuation* reward under the optimal decision rule (discounted back to stage n). Since $u_k \leq \mathbf{E}_\sigma[\overline{u} \mid \mathcal{H}_k]$ for all $k \geq n$, one has $\mathbf{E}_\sigma[Y_n] \leq \mathbf{E}[\overline{u}]$.

Since one option available to the decision maker, from stage n on, is to ignore all future observations, and to keep choosing the action that was myopically optimal in stage n, we have

$$Y_n \geq \overline{u}_n. \tag{1}$$

Now, rewrite Y_n as

$$\begin{aligned} Y_n &= (1-\delta)u_n + \delta\mathbf{E}_\sigma[Y_{n+1} \mid \mathcal{H}_n] \\ &= (1-\delta)(\overline{u}_n - \Delta_n) + \delta\mathbf{E}_\sigma[Y_{n+1} \mid \mathcal{H}_n]. \end{aligned} \tag{2}$$

From (1) and (2) we obtain:

$$\overline{u}_n \leq (1-\delta)(\overline{u}_n - \Delta_n) + \delta\mathbf{E}_\sigma[Y_{n+1} \mid \mathcal{H}_n],$$

so that after cancelling $\overline{u}_n$ from both sides and dividing by δ,

$$\overline{u}_n \leq \mathbf{E}_\sigma[Y_{n+1} \mid \mathcal{H}_n] - \frac{\Delta_n(1-\delta)}{\delta}. \tag{3}$$

Substituting (3) into (2), we obtain

$$\begin{aligned} Y_n &\leq (1-\delta)\left(\mathbf{E}_\sigma[Y_{n+1} \mid \mathcal{H}_n] - \Delta_n\left(\frac{1-\delta}{\delta} + 1\right)\right) + \delta\mathbf{E}_\sigma[Y_{n+1} \mid \mathcal{H}_n] \\ &\leq \mathbf{E}_\sigma[Y_{n+1} \mid \mathcal{H}_n] - \frac{1-\delta}{\delta}\Delta_n. \end{aligned}$$

Taking expectations, summing over $n = 1, \ldots, k$, using $\mathbf{E}[\underline{u}] \leq \mathbf{E}_\sigma[Y_n] \leq \mathbf{E}[\overline{u}]$, and taking the limit as k goes to infinity, we obtain

$$\mathbf{E}_\sigma\left[\sum_{n\geq 1}\Delta_n\right] \leq (\mathbf{E}[\overline{u}] - \mathbf{E}[\underline{u}]) \times \frac{\delta}{(1-\delta)},$$

as desired. □

3.2. Proof of Proposition 2.2

Fix $\delta > 0$. Note that if the statement holds for ε_0, then it holds for every $\varepsilon > \varepsilon_0$. We will prove that the statement holds for $\varepsilon = 1/m$, for any natural number $m > 1/\delta$. Let $\Theta = \{\theta_1, \theta_2, \ldots, \theta_m\}$ and $A = \{a_0, a_1, \ldots, a_m\}$ contain m and $m+1$ elements respectively. The prior belief on Θ is uniform, and the reward function is given by:

$$u(\theta_k, a_k) = 1, \quad k = 1, \ldots, m, \tag{4}$$

$$u(\theta_k, a_l) = 0, \quad k = 1, \ldots, m, l \neq k, \tag{5}$$

$$u(\theta_k, a_0) = 0, \quad k = 1, \ldots, m. \tag{6}$$

Thus, once the parameter is inferred with certainty, there is a unique optimal action, whereas *ex ante*, $a_1, \ldots, a_m$ are all myopically optimal, while a_0 is $(1/m)$-suboptimal.

Information is provided to the decision maker according to the following rules: if the decision maker has chosen a_0 in all previous stages, the true parameter is revealed with probability $c := \frac{(1-\delta)}{\delta(m-1)} < 1$; if the decision maker did not choose a_0 in all previous stages, no information is revealed, that is, no observation is made. Suppose the decision maker chooses a_0 until the state of the world is revealed, and then switches to the optimal action. The expected reward A satisfies $A = c\delta + (1-c)\delta A$, so that $A = \frac{c\delta}{1-(1-c)\delta}$. Substituting $c = \frac{(1-\delta)}{\delta(m-1)}$ we obtain that the expected reward is $1/m$, so that this strategy is optimal. However, for $\varepsilon = 1/m$ one has:

$$\mathbf{E}_\sigma\left[\sum_{n\geq 1}\Delta_n\right] = \mathbf{E}_\sigma[\varepsilon N(\varepsilon)] = \frac{\varepsilon}{c} = \frac{m-1}{m}\frac{\delta}{1-\delta}.$$

Since $\bar{u} = 1$ and $\underline{u} = 0$ we get the desired result. □

3.3. Proof of Proposition 2.3

We provide an example within the class of Gaussian models. Set $\Theta = \mathbf{R}$, and let the action set $A = \mathbf{R} \cup \{-\infty, +\infty\}$ be the set of extended real numbers, endowed with the usual topology. The reward function $u(\theta, a)$ is equal to one if $a \in \mathbf{R}$ and $|\theta - a| \leq 1$, and equal to zero otherwise.

Given a normal distribution μ with precision ρ (that is, with variance $1/\rho$), define $\bar{u}(\rho)$ to be the highest reward that the decision maker may achieve, when holding the belief μ. Observe that $\bar{u}(\rho)$ does not depend on the mean of μ. Plainly, the map $\rho \mapsto \bar{u}(\rho)$ is continuous and increasing, with $\lim_{\rho\to 0}\bar{u}(\rho) = 0$, and $\lim_{\rho\to+\infty}\bar{u}(\rho) = 1$.

The signalling structure of the decision problem is designed in such a way that the decision maker's belief is always a normal distribution. In addition, she keeps receiving additional information about θ as long as she follows a pre-specified sequence of suboptimal actions.

To be specific, let $(\varepsilon_n)_{n\geq 1}$ be a decreasing sequence of positive numbers that satisfies (i) $\sum_{n=1}^{\infty}\varepsilon_n \in (1/2, 1)$, (ii) $\varepsilon_n n^\beta \to +\infty$, for every $\beta > 1$, and[6] (iii) $\frac{\varepsilon_{n-1}}{\varepsilon_n} > \frac{2}{3}$. The sequence $(\rho_n)_{n\geq 1}$ is defined recursively by the condition

$$\bar{u}(\rho_1 + \cdots + \rho_n) = \varepsilon_1 + \cdots + \varepsilon_n.$$

Let the prior distribution $\mathbf{P}$ be a normal distribution with precision ρ_1, and let $(\xi_n)_{n\geq 2}$ be a sequence of independent normally distributed variables with precision ρ_n, and independent from θ.

Observe that, in the absence of any information about θ, the decision maker's myopically optimal reward is $\bar{u}(\rho_1) = \varepsilon_1$. We set $a_1 = +\infty$. On the other hand, if she receives the observations $s_k := \theta + \xi_k$, $k = 2, \ldots, n$ ($n \geq 2$), her belief over θ is normally distributed, with precision $\rho_1 + \cdots + \rho_n$. Hence, her myopically optimal reward is $\bar{u}(\rho_1 + \cdots + \rho_n) = \varepsilon_1 + \cdots + \varepsilon_n$, and there is an action a_n (which depends on $s_2, \ldots, s_n$), which yields an expected reward equal to $\varepsilon_1 + \cdots + \varepsilon_{n-1}$.

We now define the information received by the decision maker:

- Prior to stage 1, the decision maker receives no observation.
- Prior to stage 2, she receives the observation $s_2 = \theta + \xi_2$ if she played $a_1 = +\infty$ at the first stage, and no observation otherwise.
- Prior to stage $n > 2$, she receives the observation $s_n = \theta + \xi_n$ if she played $a_1, a_2, \ldots, a_{n-1}$ at the previous stages. Otherwise, she receives no observation.

Playing the sequence (a_n) of actions is the unique optimal decision rule. Indeed, if the decision maker first deviates from that sequence at stage $k \geq 1$, she receives no further information, hence her optimal reward in all later stages is $\varepsilon_1 + \cdots + \varepsilon_k$; if she sticks to the sequence (a_n), her continuation reward (discounted back to stage k) is

$$(1-\delta)\sum_{n=k}^{\infty}\delta^{n-k}(\varepsilon_1 + \cdots + \varepsilon_{n-1}).$$

By (iii), this reward is higher than $\varepsilon_1 + \cdots + \varepsilon_k$.

Note that a_n is (myopically) ε_n-optimal, for each $n \geq 1$. Since the sequence (ε_n) is decreasing, there are exactly n rounds in which the decision maker does not play a myopically ε_n-optimal action, so that by (ii) $(\varepsilon_n)^\alpha N(\varepsilon_n) = n(\varepsilon_n)^\alpha$ converges to infinity for every $\alpha < 1$. □

4. Application

We here illustrate how Theorem 2.1 can be used to derive *a priori* bounds on the optimal decision rules in specific decision problems. Since our goal is here purely illustrative, we restrict ourselves to the analysis of a specific one-arm bandit problem, where the risky arm has two possible types, a good type and a bad type, and observations are i.i.d. In such a problem, the optimal decision rule consists of pulling the risky arm as long as the posterior probability assigned to the good type exceeds a specific cut-off, and then in switching permanently to the safe arm.

[6] For instance, choose $\varepsilon_n = \frac{(n\ln^2 n)^{-1}}{\sum_{k=1}^{\infty}(k\ln^2 k)^{-1}}$ for n sufficiently large.

We set the problem so as to depart as little as possible from a Bernoulli problem, for which a closed form expression for the optimal cut-off is known. We also make no attempt at optimizing our final bound.

The type θ of the risky arm takes values in the two-point set $\{\theta_0, \theta_1\}$. Both types are *ex ante* equally likely. The safe arm yields zero. Given $\theta = \theta_i$, the risky arm may yield three different rewards, a, b and c, with probabilities p_a^i, p_b^i and p_c^i. These probabilities are such that (i) the expected reward of the risky arm is 1 if $\theta = \theta_1$, and -1 if $\theta = \theta_0$; (ii) one has $\ln \frac{p_a^1}{p_a^0} = \alpha$, $\ln \frac{p_b^1}{p_b^0} = 2\alpha$, and $\ln \frac{p_c^1}{p_c^0} = -\alpha$, for some $\alpha > 0$.

Denote by π_n the posterior belief that $\theta = \theta_1$, based on all observations prior to stage n, and let $Z_n = \ln \frac{\pi_n}{1-\pi_n}$ be the log-likelihood ratio. Conditional on $\theta = \theta_0$, the sequence (Z_n) follows a random walk, which moves up by α (with probability p_a^0), by 2α, or moves down by α between any two stages.

The optimal decision rule consists in pulling the risky arm until the first stage σ^* where $Z_n = -k^*\alpha$, for some $k^* \in \mathbf{N}$, and then in pulling repeatedly the safe arm. We will derive an upper bound on k^* using Theorem 2.1.

The amount of experimentation in stage n is $\Delta_n = \max\{0, 1/2 - \pi_n\}$. For $k < k^*$, let $N(k)$ be the number of passage of the sequence (Z_n) at the level $-k\alpha$, and denote by $\varepsilon(k) = 1/2 - \frac{e^{-k\alpha}}{1+e^{-k\alpha}}$ the corresponding value of Δ_n. Thus,

$$\sum_{n=1}^{+\infty} \Delta_n = \sum_{k<k^*} \varepsilon(k) N(k). \tag{7}$$

Observe now that whenever $Z_n = -k\alpha$, the expected number of visits (including stage n) to $-k\alpha$ before Z_n moves below $-k\alpha$ is $1/(1 - p_a^0)$. On the other hand, it is then the case that the sequence (Z_n) moves down to $-(k+1)\alpha$. Hence, the probability that (Z_n) will move back to $-k\alpha$ before hitting $-k^*\alpha$ is[7] at least p_a^0. Therefore,

$$\mathbf{E}_{\theta_0}[N(k)] \geq \frac{p_a^0}{1 - p_a^0}. \tag{8}$$

By Theorem 2.1 one has $\frac{1}{2}\mathbf{E}_{\theta_0}\left[\sum_{n=1}^{+\infty} \Delta_n\right] + \frac{1}{2}\mathbf{E}_{\theta_0}\left[\sum_{n=1}^{+\infty} \Delta_n\right] \leq \frac{2\delta}{1-\delta}$. Therefore, (7) and (8) yield

$$\sum_{k=0}^{k^*-1} \frac{1}{2} \frac{1 - e^{-k\alpha}}{1 + e^{-k\alpha}} = \sum_{k=0}^{k^*-1} \varepsilon(k) \leq 4 \frac{1 - p_a^0}{p_a^0(1-\delta)}. \tag{9}$$

By monotonicity, the left-hand side of (9) is at least equal to

$$\frac{1}{2}\int_0^{k^*-1} \tanh \frac{x\alpha}{2} dx = \frac{1}{\alpha} \ln \cosh \frac{\alpha(k^*-1)}{2} \geq \frac{1}{\alpha} \ln \frac{e^{\frac{\alpha(k^*-1)}{2}}}{2} = \frac{(k^*-1)}{2} - \frac{\ln 2}{\alpha}.$$

Thus,

$$k^* \leq 4\left(1 + 2\frac{\ln 2}{\alpha} + 2\frac{1 - p_a^0}{p_a^0(1-\delta)}\right).$$

References

Armitage, P., 1975. Sequential medical trials, 2nd edition. Blackwell, Oxford.
Bergemann, D., Välimäki, J., 2000. Experimentation in markets. Rev. Econom. Stud. 67, 213–234.
Berry, D.A., 1972. A Bernoulli two-armed bandit. Ann. Math. Statist. 43, 871–897.
Berry, D.A., 1978. Modified two-armed Bandit strategies for certain clinical trials. J. Amer. Statist. Assoc. 73, 339–345.
Berry, D.A., Fristedt, B., 1985. Bandit Problems: Sequential Allocation of Experiments. Springer.
Berry, D.A., Sobel, M., 1973. An improved procedure for selecting the better of two Bernoulli populations. J. Amer. Statist. Assoc. 68, 979–984.
Bolton, P., Harris, C., 1999. Strategic experimentation. Econometrica 67, 349–374.
Bradt, R.N., Johnson, S.M., Karlin, S., 1956. On sequential designs for miximizing the sum of n observations. Ann. Math. Statist. 27, 1060–1070.
Brezzi, M., Lai, T.L., 2000. Incomplete learning from endogenous data in dynamic allocation. Econometrica 68, 1511–1516.
Cornfield, J., Halperin, M., Greenhouse, S.W., 1969. An adaptive procedure for sequential clinical trials. J. Amer. Statist. Assoc. 64, 759–770.
Dixit, A.K., Pindyck, R.S., 1994. Investment under Uncertainty. Princeton University Press, Princeton.
Fabius, J., van Zwet, W.R., 1970. Some remarks on the two-armed bandit. Ann. Math. Statist. 41, 1906–1916.
Feldman, D., 1962. Contributions to the 'two armed bandit' problem. Ann. Math. Statist. 2, 615–629.
Gittins, J.C., 1979. Bandit processes and dynamic allocation indices. J. Roy. Statis. Soc. Ser. B 41, 148–177.
Gittins, J.C., Jones, D.M., 1974. A dynamic allocation index for the sequential design of experiments. In: Gani, J., et al. (Eds.), Progress in Statistics. North Holland, Amsterdam, pp. 241–266.
Keller, G., Rady, S., Cripps, M., 2005. Strategic experimentation with exponential bandits. Econometrica 73, 39–68.
Lai, T.L., 1987. Adaptive treatment allocation and the multi-armed bandit problem. Ann. Statist. 15, 1091–1114.
Lai, T.L., 1988. Nearly optimal sequential tests of composite hypotheses. Ann. Statist. 16, 855–886.
Lai, T.L., 2001. Sequential analysis: Some classical problems and new challenges. Statist. Sin. 11, 303–408.
Rosenberg, D., Solan, E., Vieille, N., 2007. Social learning in one-arm bandit problems. Econometrica 75, 1591–1611.
Rosenberg, D., Solan, E., Vieille, N., 2009. Informational externalities and emergence of consensus. Games Econom. Behav. 66, 979–994.
Woodroofe, M., 1979. A one-armed Bandit problem with a concomitant variable. J. Amer. Statist. Assoc. 74, 799–806.

[7] This bound is admittedly very crude.