Ann Inst Stat Math (2013) 65:993–1006
DOI 10.1007/s10463-013-0401-5

One-armed bandit process with a covariate

You Liang · Xikui Wang · Yanqing Yi

Received: 8 March 2012 / Revised: 7 February 2013 / Published online: 12 April 2013

Abstract We generalize the bandit process with a covariate introduced by Woodroofe in several significant directions: a linear regression model characterizing the unknown arm, an unknown variance for regression residuals and general discounting sequence for a non-stationary model. With the Bayesian regression approach, we assume a normal-gamma conjugate prior distribution of the unknown parameters. It is shown that the optimal strategy is determined by a sequence of index values which are monotonic and determined by the observed value of the covariate and updated posterior distributions. We further show that the myopic strategy is not optimal in general. Such structural properties help to understand the tradeoff between information gathering and immediate expected payoff and may provide certain insight for covariate adjusted response adaptive design of clinical trials.

Keywords Bandit process · Bayesian regression · Markov decision process · Optimal strategy

Y. Liang · X. Wang (✉)
Department of Statistics, University of Manitoba, Winnipeg, MB R3T 2N2, Canada
e-mail: Xikui.Wang@ad.umanitoba.ca

Y. Liang
e-mail: umlian33@cc.umanitoba.ca

Y. Yi
Division of Community Health and Humanities, Faculty of Medicine, Memorial University of Newfoundland, St. John's, NF 1B 3V6, Canada
e-mail: Yanqing.Yi@med.mun.ca

1 Introduction

Multi-armed bandit processes are an important class of sequential decision problems with applications in areas such as statistics, clinical trials, operations research, engineering and economics. A decision-maker is faced with several statistical populations, called arms, and some (or all) of these populations have unknown statistical distributions. At each point (in time or space), one and only one arm is selected for observation. There are two consequences. On one hand, a reward is randomly generated according to the statistical distribution characterizing the selected arm. On the other hand, the observation of the reward offers potentially useful information for making statistical inference of the unknown statistical distribution and hence making better informed decisions in the future. The overall objective is to maximize the total of expected, possibly discounted, rewards from all selections over a given (finite or infinite) horizon.

The two consequences of selecting an arm are often antagonistic. For example, selecting an arm with unknown distribution offers valuable information for statistical inference but does not necessarily provide the highest immediate expected reward. Therefore bandit processes are typically characterized by the competing goals of information gathering (so as to make better informed decisions in the future) and immediate payoff of the highest expected reward. Berry and Fristedt (1985) and Gittins (1989) are standard references on bandit problems with complete observations. Bandit processes with censored (hence incomplete) observations are treated in Eick (1988), Wang (2000) and Wang and Bickis (2003).

The majority of literature on bandit processes assumes homogeneous statistical populations. However in many applications such as clinical trials, certain covariates (such as demographic variables like age and gender) may offer useful information for making sequential selections, and random rewards may depend on the covariates. Woodroofe (1979) initiated research on bandit processes with covariates. Woodroofe's framework assumes two arms and the probability distribution of one arm is known. Such a model is typically named a one-armed bandit.

Let $(X_n, Y_{0,n}, Y_{1,n}), n = 1, 2, \ldots, N$, be a sequence of random variables, where X_n is the covariate for the n^{th} selection (or subject), $Y_{i,n}, i = 0, 1$, is a random reward if arm i is observed for the n^{th} selection, and N is the horizon of the sequential selection problem. The horizon N may be finite or infinite, or even random. It is assumed that $(X_n, Y_{0,n}, Y_{1,n}), n = 1, 2, \ldots, N$, are conditionally independent and identically distributed as (X, Y_0, Y_1) given the unknown distribution, but the density function $f(x)$ of the covariate X is assumed to be known.

Sequentially, we observe $X_n = x_n$ at time $n = 1, 2, \ldots, N$. Based on the current knowledge, we select either arm 0 or arm 1, but not both, and receive a random reward $Y_{0,n}$ or $Y_{1,n}$ respectively depending on the arm selected. Let π_n be a (possibly randomized) selection rule at time n so that arm 1 is selected with probability π_n. Then a strategy $\pi = (\pi_1, \pi_2, \ldots)$ is a sequence of selection probabilities. The objective is to find an optimal strategy to maximize

$$W_N(\pi) = \sum_{n=1}^{N} \rho_n E_\pi[\pi_n Y_{1,n} + (1 - \pi_n) Y_{0,n}]$$

where $0 < \rho_n < 1$ is a discounting factor such that $\sum_{n=1}^{N} \rho_n < \infty$, and E_π is the expectation taken with respect to the strategy π.

Woodroofe (1979) assumed (a) $N = \infty$, (b) $\rho_n = \rho^n, 0 < \rho < 1$, (c) a known conditional distribution of Y_0 given X, and (d) $Y_1 = X - \theta + \epsilon$ where ϵ is a random variable with zero mean and (e) a known distribution, and (f) θ is an unknown parameter following a certain prior distribution. Later, Sarkar (1991) extended this model to an exponential family model for Y_1. Both Woodroofe and Sarkar followed the Bayesian approach and examined the asymptotic solution when the geometric discounting factor ρ approaches 1.

Goldenshluger and Zeevi (2009) revisited Woodroofe's model in a non-Bayesian, minimax setting with a finite horizon. They established specific non-asymptotic lower bounds on the minimax regret and proposed intuitive rate-optimal strategies that attain these bounds. These rate-optimal strategies are not myopic. They demonstrated that the regret grows at various rates with the time horizon, depending on certain local properties of the covariate distribution.

Yang and Zhu (2002) investigated the multi-armed bandit problem with covariates and used the nonparametric approach to estimate the functional relationship between the response variable and the covariates. They introduced a randomized allocation strategy that balances the tradeoff between using the currently most promising arm and exploring the arm which is truly the best. The proposed strategy was shown to be strongly consistent in that the total reward is asymptotically equivalent to the total reward from the best arm almost surely.

In this paper, we extend Woodroofe's model with three notable generalizations. Firstly, we extend Woodroofe's model $Y_1 = X - \theta + \epsilon$ to the standardized linear regression $Y_1 = \beta_1 X + \epsilon$. Further generalization to multiple linear regression is possible but tedious. Secondly, Woodroofe and Sarkar considered the infinite horizon model with a geometric discounting sequence $(1, \rho, \rho^2, \ldots)$. Mathematically such a model is stationary and hence more tractable than a non-stationary model. We examine the more challenging non-stationary model where N is a fixed, finite integer, and the discounting sequence $(\rho_1, \rho_2, \ldots, \rho_N, 0, 0, \ldots)$ may be general. Lastly, Woodroofe assumed a known variance for ϵ but we assume an unknown variance for ϵ.

This paper is organized as follows. We introduce the notations and formulate the models in the next section. The one-armed bandit is investigated in Section 3 and lengthy proofs are provided in the Appendix. Section 4 concludes the paper.

2 Notations and model formulation

In this paper, we assume two arms. If arm 1 is observed for the nth selection under a strategy π, we assume that the random reward is given by the standardized regression $Y_{1,n} = \beta X_n + \epsilon_n$, where $\epsilon_n, n = 1, 2, \ldots, N$, are independent and identically distributed following a normal distribution with mean 0 and unknown variance σ^2. We assume an unknown β and call arm 1 the unknown arm. On the other hand, if arm 0 is selected at any time n, we assume a known constant expected reward $E(Y_{0,n}|X_n) \equiv B$ and call arm 0 the known arm.

Assume that prior to time $n \geq 1$, the unknown arm is selected k times at decision times $1 \leq n_1 < n_2 < \cdots < n_k < n$ and $\mathcal{O}_n = \{(x_{n_j}, y_{1,n_j}), j = 1, 2, \ldots, k\}$ is the set of observations. Write $\gamma_n = \sum_{j=1}^{k} x_{n_j}^2$, $\tau_n = \sum_{j=1}^{k} y_{1,n_j}^2$ and $\eta_n = \sum_{j=1}^{k} x_{n_j} y_{1,n_j}$. Then the likelihood function of β and σ^2 given $\mathcal{O}_n$ is

$$\ell(\beta, \sigma^2 | \mathcal{O}_n) \propto \sigma^{-k} \exp\left[-\frac{1}{2\sigma^2}\left((k-1)\hat{\sigma}_n^2 + (\beta - \hat{\beta}^{(n)})^2 \gamma_n\right)\right],$$

where $\exp(z) = e^z$, $\hat{\beta}^{(n)}$ and $\hat{\sigma}_n^2$ are the ordinary least squares (OLS) estimates given by $\hat{\beta}^{(n)} = \frac{\eta_n}{\gamma_n}$ and $\hat{\sigma}_n^2 = \frac{\sum_{j=1}^{k}(y_{1,n_j} - \hat{\beta}^{(n)} x_{n_j})^2}{k-1} = \frac{1}{k-1}\left(\tau_n - \frac{\eta_n^2}{\gamma_n}\right)$ (Bansal 2007, Chapter 9).

Define the measurement precision $\delta = \frac{1}{\sigma^2}$ and assume the natural conjugate prior $g(\beta, \delta) = g(\beta|\delta)g(\delta)$ for (β, δ) where $g(\delta)$ is a gamma distribution $G(\mu_0, \nu_0)$ and $g(\beta|\delta)$ is a normal distribution with mean $\beta^{(0)}$ and precision $M\delta$, $M > 0$. According to Bansal (2007, Chapter 9), after observing $\mathcal{O}_n$ at time n, the joint posterior distribution of (β, δ) is given by $g(\beta, \delta|\mathcal{O}_n) = g(\beta|\delta, \mathcal{O}_n)g(\delta|\mathcal{O}_n)$ where $g(\delta|\mathcal{O}_n) = G(\mu_n, \nu_n)$ is the gamma distribution with $\mu_n = \mu_0 + \frac{k}{2}$ and $\nu_n = \nu_0 + \frac{(k-1)\hat{\sigma}_n^2}{2} + \frac{M\gamma_n(\beta^{(0)} - \hat{\beta}^{(n)})^2}{2(M+\gamma_n)}$, and $g(\beta|\delta, \mathcal{O}_n)$ is the normal distribution with mean $\beta^{(n)} = \frac{M\beta^{(0)} + \gamma_n \hat{\beta}^{(n)}}{M+\gamma_n}$ and precision $(M + \gamma_n)\delta$. Here the posterior mean of the normal distribution is a weighted average of the prior mean and the ordinary least squares estimate. Furthermore, the marginal distribution of β is a 3-parameter t-distribution with $df^{(n)} = k + 2\mu_0$ degrees of freedom, location parameter $\beta^{(n)}$ and scale parameter $s^{(n)} = \frac{(k+2\mu_0)(M+\gamma_n)}{2\nu_n}$. We denote this distribution as $t(df^{(n)}, \beta^{(n)}, s^{(n)})$.

Write the sequence of discounting factors as $A_n^N = (\rho_n, \rho_{n+1}, \ldots, \rho_N, 0, \ldots)$, $n = 1, 2, \ldots, N$, and the mean of X as $\lambda = E(X)$. We assume non-increasing discounting so that $\rho_1 \geq \rho_2 \geq \ldots \geq \rho_N$. Under the Bayesian approach, the one-armed bandit process becomes a Markov decision process where at time $n = 1, 2, \ldots, N$, the state $s_n = (x_n, d^{(n)})$ is jointly given by the observed covariate $X_n = x_n$ and the posterior distributions

$$d^{(n)} = \left(t(df^{(n)}, \beta^{(n)}, s^{(n)}), G(\mu_n, \nu_n)\right),$$

the action space is $\{0, 1\}$, the one step random reward is given by $Y_{i,n}$ if arm i is selected, and the state either changes to $s_{n+1} = (x_{n+1}, d^{(n+1)})$ where

$$d^{(n+1)} = \left(t(df^{(n+1)}, \beta^{(n+1)}, s^{(n+1)}), G(\mu_{n+1}, \nu_{n+1})\right)$$

according to Bayes' law if the unknown arm is selected or to $s_{n+1} = (x_{n+1}, d^{(n)})$ where $d^{(n)}$ remains unchanged if the known arm is selected. According to Berry and Fristedt (1985, Page 14), the set of all distributions forms a Borel space. By standard theory of Markov decision processes with a Borel state space and a finite action space, there exists a Markov deterministic strategy which is optimal. Moreover the optimal strategy is characterized by the optimality equation based on which we can iteratively

derive the optimal strategy and investigate structural and other properties of the optimal strategy.

3 Main results

For any $n = 1, 2, \ldots, N$, let $s_n = (x_n, d^{(n)})$ be the observed state and

$$V\left(s_n, B, A_n^N\right) = V\left((x_n, t(df^{(n)}, \beta^{(n)}, s^{(n)}), G(\mu_n, \nu_n)), B, A_n^N\right)$$

be the optimal value of the bandit from time n until N, starting with the state s_n. Then the principle of optimality states that

$$\begin{aligned} &V\left((x_n, d^{(n)}), B, A_n^N\right) \\ &\quad = \max\left\{V^{(0)}\left((x_n, d^{(n)}), B, A_n^N\right), V^{(1)}\left((x_n, d^{(n)}), B, A_n^N\right)\right\} \end{aligned} \tag{1}$$

where $V^{(i)}\left((x_n, d^{(n)}), B, A_n^N\right) = V^{(i)}\left((x_n, t(df^{(n)}, \beta^{(n)}, s^{(n)}), G(\mu_n, \nu_n)), B, A_n^N\right)$ is the value of the strategy that selects arm $i = 0, 1$, at the initial state $(x_n, d^{(n)})$ and then always follows an optimal strategy starting at time $n + 1$. By the principle of dynamic programming, we have

$$\begin{aligned} &V^{(1)}\left((x_n, d^{(n)}), B, A_n^N\right) \\ &\quad = V^{(1)}\left((x_n, t(df^{(n)}, \beta^{(n)}, s^{(n)}), G(\mu_n, \nu_n)), B, A_n^N\right) = \rho_n \beta^{(n)} x_n \\ &\qquad + E\left[V\left((X_{n+1}, t(df^{(n+1)}, \beta^{(n+1)}, s^{(n+1)}), G(\mu_{n+1}, \nu_{n+1})), B, A_{n+1}^N\right) | s_n\right], \end{aligned}$$

where $df^{(n+1)} = df^{(n)} + 1$, $\beta^{(n+1)} = \beta^{(n+1)}(s_n, X_n, Y_{1,n})$, $s^{(n+1)} = s^{(n+1)}(s_n, X_n, Y_{1,n})$, $\mu_{n+1} = \mu_n + \frac{1}{2}$ and $\nu_{n+1} = \nu_{n+1}(s_n, X_n, Y_{1,n})$. On the other hand,

$$\begin{aligned} &V^{(0)}\left((x_n, d^{(n)}), B, A_n^N\right) = V^{(0)}\left((x_n, t(df^{(n)}, \beta^{(n)}, s^{(n)}), G(\mu_n, \nu_n)), B, A_n^N\right) \\ &\quad = \rho_n B + E\left[V\left((X_{n+1}, t(df^{(n)}, \beta^{(n)}, s^{(n)}), G(\mu_n, \nu_n)), B, A_{n+1}^N\right)\right] \end{aligned}$$

because the distributions $d^{(n)}$ do not change when the known arm is selected.

In the expression of $V^{(1)}\left((x_n, d^{(n)}), B, A_n^N\right)$, the term

$$V\left((X_{n+1}, t(df^{(n+1)}, \beta^{(n+1)}, s^{(n+1)}), G(\mu_{n+1}, \nu_{n+1})), B, A_{n+1}^N\right)$$

depends on the random variables X_{n+1} and $Y_{1,n}$ where the density function of X_{n+1} is $f(x)$ and conditional on $s_n = (x_n, d^{(n)})$, the marginal density function of $Y_{1,n}$ is given as $h(y|s_n)$. Hence

$$V^{(1)}\left((x_n, d^{(n)}), B, A_n^N\right) = \rho_n \beta^{(n)} x_n + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} V\left((x, d^{(n+1)}(y)), B, A_{n+1}^N\right) h(y|s_n) f(x) \mathrm{d}x \mathrm{d}y, \quad (2)$$

where $d^{(n+1)}(y)$ indicates that the posterior distributions depend on y and $h(y|s_n)$ is the conditional distribution of $Y_{1,n}$ given the observed state s_n.

Similarly in $V^{(0)}\left((x_n, d^{(n)}), B, A_n^N\right)$, the term

$$V\left((X_{n+1}, t(df^{(n)}, \beta^{(n)}, s^{(n)}), G(\mu_n, \nu_n)), B, A_{n+1}^N\right)$$

depends on the random variable X_{n+1} and hence

$$V^{(0)}\left((x_n, d^{(n)}), B, A_n^N\right) = \rho_n B + \int_{-\infty}^{\infty} V\left((x, d^{(n)}), B, A_{n+1}^N\right) f(x) \mathrm{d}x. \quad (3)$$

Given any state $s_n = (x_n, d^{(n)})$ at time $n = 1, 2, \ldots, N$, define

$$\Delta\left((x_n, d^{(n)}), B, A_n^N\right) = V^{(1)}\left((x_n, d^{(n)}), B, A_n^N\right) - V^{(0)}\left((x_n, d^{(n)}), B, A_n^N\right) \quad (4)$$

as the advantage of the unknown arm over the known arm, given the currently observed covariate x_n and updated posterior distributions $d^{(n)}$. Then the unknown (known respectively) arm is optimal at state $s_n = (x_n, d^{(n)})$ if and only if $\Delta\left((x_n, d^{(n)}), B, A_n^N\right) \geq (\leq) 0$. But there is no analytic solution to the inequality $\Delta\left((x_n, d^{(n)}), B, A_n^N\right) \geq 0$. Our main results show that, for each given observed covariate x_n and posterior distributions $d^{(n)}$ (i.e., for each given state $s_n = (x_n, d^{(n)})$), the root $B^*(x_n, d^{(n)}, A_n^N)$ of B for the equation $\Delta\left((x_n, d^{(n)}), B, A_n^N\right) = 0$ characterizes the optimal selection of arm at state s_n, the sequence $B^*(x, d, A_1^N)$, $N = 1, 2, \ldots$, displays a monotonic structure in N for the same state $s = (x, d)$, and the myopic strategy is not optimal in general. A strategy is said to be myopic at the state $s_n = (x_n, d^{(n)}), n = 1, 2, \ldots, N$, when the known (unknown respectively) arm is selected if and only if $B \geq (\leq) \beta^{(n)} x_n = E(\beta|d^{(n)}) x_n$. That is, the myopic strategy selects the arm with the highest immediate expected payoff and so is optimal if there is only one selection to make. However, the myopic strategy ignores information gathering.

The next theorem states our first main result characterizing the index value which defines the optimal initial selection of arm. Such an existence result greatly simplifies the form of the optimal strategy.

Theorem 1 *At any given time n, $n = 1, \ldots, N$, and for any given state $s_n = (x_n, d^{(n)})$ and discounting sequence A_n^N, there exists an index value $B^*(x_n, d^{(n)}, A_n^N)$ such that*

$$\Delta\left((x_n, d^{(n)}), B^*(x_n, d^{(n)}, A_n^N), A_n^N\right) = 0.$$

The index value is unique if the discounting sequence is strictly decreasing, otherwise the set of index values may form an interval. Moreover, for the bandit process $\left((x_n, d^{(n)}), B, A_n^N\right)$*, the unknown arm is optimal if and only if* $B \leq B^*(x_n, d^{(n)}, A_n^N)$ *while the known arm is optimal if and only if* $B \geq B^*(x_n, d^{(n)}, A_n^N)$.

In the case of truncated geometric discounting, the next main result shows a monotonicity property of the index values for the same state but changing time horizon. The intuitive interpretation is that the more selections we have to make under the same conditions, the more sacrifice in the immediate expected reward we can take to select the unknown arm, so the gain from understanding the unknown distribution can be benefited more at later selections and hence a potentially higher overall value may be reached.

Necessarily to maintain the same conditions for the first selection, we assume a truncated geometric discounting sequence.

Theorem 2 *Let* $A_1^N = (1, \rho, \rho^2, \ldots, \rho^{N-1}, 0, \ldots)$ *be a truncated geometric discounting sequence. Given the same initial state* $s = (x, d)$ *(where* x *is the observed covariate for the first selection and* d *is the given prior distribution) for all but different horizons* $N = 1, 2, \ldots$*, let* $B^*(x, d, A_1^N)$ *be the index value such that*

$$\Delta\left((x, d), B^*(x, d, A_1^N), A_1^N\right) = 0, N = 1, 2, \ldots.$$

Then

$$\beta^{(0)}x = B^*(x, d, A_1^1) \leq B^*(x, d, A_1^2) \leq \cdots \leq B^*(x, d, A_1^N) \leq \cdots.$$

The limit $B^*(x, d) = \lim_{N\to\infty} B^*(x, d, A_1^N)$ *exists such that* $\beta^{(0)}x < B^*(x, d) < \infty$ *and satisfies the equation* $\Delta\left((x, d), B^*(x, d), A\right) = 0$ *for the infinite horizon model, where* $A = (1, \rho, \rho^2, \ldots)$ *is the geometric discounting sequence.*

There is an interesting corollary saying that the myopic strategy is not optimal in general. In the case of $\beta^{(n)}x_n < B$, the immediate expected payoff is smaller from the unknown arm. But when the difference is not significant large, it may still be optimal to select the unknown arm because doing so may provide useful information and higher expected payoffs in the future. This information gathering may compensate for the loss of immediate expected payoff and give higher value of the whole bandit process. The difference between the index value and the immediate expected payoff is termed the learning component of the index by Gittins and Wang (1992).

Corollary 1 *The myopic strategy is not optimal in general.*

Proof Consider the case of $N = 2$. Let B be such that $\beta^{(0)}x < B < B^*(x, d, A_1^2)$ for any given initial state $s = (x, d)$. For the bandit model $V\left((x, d), B, A_1^2\right)$, the unknown arm is uniquely optimal for the first selection, however the myopic strategy selects the known arm. □

4 Conclusion

Bandit models with covariate variables are motivated by practical applications in clinical trials and other fields. Essentially information from covariates can be useful for making better informed selections among statistical populations. Results in this paper represent significant extensions of current one-armed bandit models with a covariate. For example, the results from both Woodroofe (1979) and Sarkar (1991) show that the asymptotically optimal strategy depends on the value of the covariate. We have obtained similar results in both cases of finite and infinite horizon models. The index value obtained in this paper which characterizes the optimal initial selection of arms depends on the observed covariate and may be regarded as an extension of the celebrated Gittins index to the case of non-stationary, finite horizon models.

Bandit problems are related to response adaptive design of clinical trials (Hu and Rosenberger 2006, Page 7), although there are significant difference in philosophy and methodology. Nevertheless, results for bandit processes with covariates may provide useful insight for investigating covariate adjusted response adaptive design of clinical trials (Hu and Rosenberger 2006, Page 6). Response adaptive designs are becoming important and popular because they use information so far accumulated from the trial to modify the randomization procedure and deliberately bias treatment allocation in order to assign more patients to the potentially better treatment without undermining the validity and integrity of the clinical research (Yi and Wang 2009, Li and Wang 2012). It is possible to extend the results obtained in this paper to the framework in Li and Wang (2012), with the possibility of mis-measured covariate X.

The results in this paper may also be extended to finance and economics. For example, we may assume that the profit from sales or investments is described by a regression model. Then results obtained in this paper may be extended to the dynamic pricing model introduced in Wang (2007) and the optimal investment and consumption model in Wang and Wang (2010).

The models constructed, methods used and results obtained can be generalized in various directions. If the horizon N is random, the technique used in Wang and Yi (2009, Theorem 2), may be applied to reformulate the random horizon bandit model as an infinite horizon bandit model. The case of random horizon N is also dealt with in Wang and Gittins (1992) in the framework of classic bandit models. The performance of dynamic allocation indices and their calculations were considered. The unknown arm can also be directly extended to the form of $Y_{1,n} = \beta X_n - \theta + \epsilon$ after assuming independence. This further generalizes Woodroofe's model by taking $\beta \equiv 1$. The model can also be extended to multiple linear regression with several independent covariates, for which results similar to the Gittins index strategy may be derived. In this extension, we can assume zero intercepts because we can always apply correlation adjusted transformations of the variables and focus on an equivalent but standardized regression model without intercepts.

In practice, the method of dynamic programming is implemented in the usual manner. For any given initial state consisting of the observed covariate and the prior distribution, we derive the optimal value characterized by Eq. (1) by means of backward induction. The optimal value starting at stage 2 and onward is determined by the newly observed covariate and the updated posterior distribution (after observing

the response from the first selection) by the Eqs. (2) and (3). Recursively the optimal value starting at any stage is characterized by the optimal value starting at the later stage, via Eq. (1). Continuing recursively, when there is only selection to make, the optimal value is given by the maximum of the two immediate expected payoffs from the two arms, which are easily calculated for any updated posterior distribution and observed covariate. We then substitute this last stage optimal value into the second last stage Eq. (1), and continue until we reach the initial state. Although the dynamic programming is a powerful method for multi-stage optimization problems, it faces the curse of dimensionality because we will have to keep tract of all paths whose dimension increases dramatically. The curse of dimensionality is also a difficulty with the Bayesian method because of the many paths from the prior distribution to its posterior distribution. It was the curse of dimensionality that motivated the development of the Gittins index for bandit problems, which was based on the idea of looking forward optimally (with a random number of steps) instead of looking backward. The index value $B^*(s_n, A_n^N)$ in this paper is regarded as a finite horizon extension of the infinite horizon Gittins index. In principle, the calculation of $B^*(s_n, A_n^N)$ is based on setting Eq. (4) to 0 and solving for B. However in practice, this calculation is tedious if not impossible.

Partial results for two-armed bandit models with a covariate have been derived and will be presented in a separate paper. We are also working on bandit models with dependent arms which are characterized by multiple linear regressions. Applications of the techniques and results in this papers to finance and economics problems mentioned above are also under consideration.

Appendix: Proofs of results

For the existence of the index value $B^*(x_n, d^{(n)}, A_n^N)$ which characterizes the optimal initial selection of the arm at the given state $s_n = (x_n, d^{(n)})$, we show that the advantage function $\Delta\left((x_n, d^{(n)}), B, A_n^N\right)$ is a continuous and monotonic function of B and takes both positive and negative values.

Lemma 1 *For any given state $s_n = (x_n, d^{(n)}), n = 1, \ldots, N$, the functions $V\left((x_n, d^{(n)}), B, A_n^N\right)$ and $V^{(i)}\left((x_n, d^{(n)}), B, A_n^N\right), i = 0, 1$, are continuous in B.*

Therefore the advantage function $\Delta\left((x_n, d^{(n)}), B, A_n^N\right)$ is also continuous in B.

Proof The result is easily proved by backward induction on $n = N, N-1, \ldots, 1$, and the Dominated Convergence Theorem, by applying the Eq. (2) for $V^{(1)}\left(s_n, B, A_n^N\right)$, Eq. (3) for $V^{(0)}\left(s_n, B, A_n^N\right)$, Eq. (1) for $V\left(s_n, B, A_n^N\right)$, and Eq. (4) for $\Delta\left(s_n, B, A_n^N\right)$. □

Lemma 2 *For any given state $s_n = (x_n, d^{(n)}), n = 1, \ldots, N$, the function $\Delta\left((x_n, d^{(n)}), B, A_n^N\right)$ is nonincreasing in B.*

Proof We prove by backward induction on $n = N, N-1, \ldots, 1$. The lemma is clearly true when $n = N$ (i.e., for A_N^N) because there is only one selection. For the function Δ, define $\Delta^+ = \max\{\Delta, 0\}$ and $\Delta^- = \max\{-\Delta, 0\}$. Then the function V can be written as $V = V^{(0)} + \Delta^+$ or $V = V^{(1)} + \Delta^-$.

Suppose that the lemma is true for A_n^N. For A_{n-1}^N, we have

$$
\begin{aligned}
&\Delta\left((x_{n-1}, d^{(n-1)}), B, A_{n-1}^N\right) \\
&= \rho_{n-1}\beta^{(n-1)}x_{n-1} + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} V\left((x, d^{(n)}(y)), B, A_n^N\right) h(y|s_{n-1}) f(x)\mathrm{d}x\mathrm{d}y \\
&\quad -\rho_{n-1}B - \int_{-\infty}^{\infty} V\left((x, d^{(n-1)}), B, A_n^N\right) f(x)\mathrm{d}x \\
&= \rho_{n-1}\beta^{(n-1)}x_{n-1} \\
&\quad + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} V^{(0)}\left((x, d^{(n)}(y)), B, A_n^N\right) h(y|s_{n-1}) f(x)\mathrm{d}x\mathrm{d}y \\
&\quad + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Delta^{+}\left((x, d^{(n)}(y)), B, A_n^N\right) h(y|s_{n-1}) f(x)\mathrm{d}x\mathrm{d}y \\
&\quad -\rho_{n-1}B - \int_{-\infty}^{\infty} V^{(1)}\left((x, d^{(n-1)}), B, A_n^N\right) f(x)\mathrm{d}x \\
&\quad - \int_{-\infty}^{\infty} \Delta^{-}\left((x, d^{(n-1)}), B, A_n^N\right) f(x)\mathrm{d}x \\
&= \rho_{n-1}\beta^{(n-1)}x_{n-1} \\
&\quad +\rho_n B + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} V\left((z, d^{(n+1)}(x, d^{(n)}(y))), B, A_{n+1}^N\right) \\
&\quad \times h(y|s_{n-1}) f(x)\mathrm{d}x\mathrm{d}y f(z)\mathrm{d}z \\
&\quad + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Delta^{+}\left((x, d^{(n)}(y)), B, A_n^N\right) h(y|s_{n-1}) f(x)\mathrm{d}x\mathrm{d}y \\
&\quad -\rho_{n-1}B - \rho_n\beta^{(n-1)}E(X_n) \\
&\quad - \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} V\left((z, d^{(n+1)}(x, d^{(n)}(y))), B, A_{n+1}^N\right) \\
&\quad \times h(y|s_{n-1}) f(x)\mathrm{d}x\mathrm{d}y f(z)\mathrm{d}z \\
&\quad - \int_{-\infty}^{\infty} \Delta^{-}\left((x, d^{(n-1)}), B, A_n^N\right) f(x)\mathrm{d}x \\
&= (\rho_n - \rho_{n-1})B + \rho_{n-1}\beta^{(n-1)}x_{n-1} - \rho_n\beta^{(n-1)}E(X_n) \\
&\quad + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Delta^{+}\left((x, d^{(n)}(y)), B, A_n^N\right) h(y|s_{n-1}) f(x)\mathrm{d}x\mathrm{d}y \\
&\quad - \int_{-\infty}^{\infty} \Delta^{-}\left((x, d^{(n-1)}), B, A_n^N\right) f(x)\mathrm{d}x.
\end{aligned}
$$

The first term is non-increasing in B because the discounting sequence is non-increasing, and is decreasing in B if the discounting sequence is strictly decreasing. Furthermore the backward induction assumption implies that $\Delta^{+}\left((x, d^{(n)}(y)), B, A_n^N\right)$ is non-increasing in B and $\Delta^{-}\left((x, d^{(n-1)}), B, A_n^N\right)$ is non-decreasing in B. So the function $\Delta\left((x_{n-1}, d^{(n-1)}), B, A_{n-1}^N\right)$ is non-increasing in B. □

Proof (of theorem 1) The existence of the index value $B^*(x_n, d^{(n)}, A_n^N)$ satisfying the equation $\Delta\left((x_n, d^{(n)}), B^*(x_n, d^{(n)}, A_n^N), A_n^N\right) = 0$ is obvious from the continuity and monotonicity of $\Delta\left((x_n, d^{(n)}), B, A_n^N\right)$ in B, and the facts that

$$\lim_{B\to-\infty} \Delta\left((x_n, d^{(n)}), B, A_n^N\right) > 0, \quad \lim_{B\to\infty} \Delta\left((x_n, d^{(n)}), B, A_n^N\right) < 0.$$

Since $\Delta\left((x_n, d^{(n)}), B, A_n^N\right)$ is non-increasing in B and hence takes value 0 possibly over an interval, the set of all index values may form an interval. Finally, the optimal selection is made based on $B^*(x_n, d^{(n)}, A_n^N)$ according to the advantage function Δ, so that the unknown arm is optimal initially if and only if $\Delta\left((x_n, d^{(n)}), B, A_n^N\right) \geq 0$ which is true if and only if $B \leq B^*(x_n, d^{(n)}, A_n^N)$. □

To prove the monotonicity of the index values $B^*(x, d, A_1^N)$, $N = 1, 2, \ldots$, for a fixed initial state $s = (x, d)$, we need a lemma on the recursive relationship of the Δ functions, assuming a truncated geometric discounting sequence.

Lemma 3 *Assume that* $A_1^N = (1, \rho, \rho^2, \ldots, \rho^{N-1}, 0, \ldots), 0 < \rho < 1$, *is the truncated, geometric discounting sequence. For any given observed covariate* x *and prior distribution* d *(and hence for any given initial state* $s = (x, d)$*), if* $\Delta\left((x, d), B, A_1^N\right) = 0$, *then* $\Delta\left((x, d), B, A_1^{N+1}\right) \geq 0$.

Proof After the initial selection, the discounting sequence becomes $(\rho, \rho^2, \ldots, \rho^{N-1}, 0, \ldots) = \rho A_1^{N-1}$. The equation $\Delta\left((x, d), B, A_1^N\right) = 0$ implies both

$$\beta^{(0)}x - B = \rho \int_{-\infty}^{\infty} V\left((z, d), B, A_1^{N-1}\right) f(z)\mathrm{d}z - \rho \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} V\left((z, d^{(1)}(y)), B, A_1^{N-1}\right) h(y|s) f(z)\mathrm{d}z\mathrm{d}y$$

and

$$V\left((x, d), B, A_1^N\right) = B + \rho \int_{-\infty}^{\infty} V\left((z, d), B, A_1^{N-1}\right) f(z)\mathrm{d}z.$$

Hence

$$\begin{aligned}
&\Delta\left((x, d), B, A_1^{N+1}\right) \\
&= \beta^{(0)}x + \rho \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} V\left((z, d^{(1)}(y)), B, A_1^N\right) h(y|s) f(z)\mathrm{d}z\mathrm{d}y \\
&\quad - B - \rho \int_{-\infty}^{\infty} V\left((z, d), B, A_1^N\right) f(z)\mathrm{d}z \\
&= \rho \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left\{V\left((z, d^{(1)}(y)), B, A_1^N\right) - V\left((z, d^{(1)}(y)), B, A_1^{N-1}\right)\right\}
\end{aligned}$$

$$\times h(y|s) f(z)\mathrm{d}z\mathrm{d}y$$
$$-\rho \int_{-\infty}^{\infty} \left\{ V\left((z,d), B, A_1^N\right) - V\left((z,d), B, A_1^{N-1}\right) \right\} f(z)\mathrm{d}z.$$

Let π^* be an optimal strategy for $V\left((z,d), B, A_1^N\right)$. For $V\left((z,d), B, A_1^{N-1}\right)$, we follow π^* for the $N-1$ selections. Then $V\left((z,d), B, A_1^N\right) - V\left((z,d), B, A_1^{N-1}\right) \leq \rho^{N-2} E_{\pi^*}(Z_N)$ where Z_N is the random reward from the N^{th} selection.

Let π^{**} be an optimal strategy for $V\left((z, d^{(1)}(y)), B, A_1^{N-1}\right)$. For $V\left((z, d^{(1)}(y)), B, A_1^N\right)$, we follow π^{**} for the first $N-1$ selections and then follow π^* for the N^{th} selection. Hence

$$V\left((z, d^{(1)}(y)), B, A_1^N\right) - V\left((z, d^{(1)}(y)), B, A_1^{N-1}\right) \geq \rho^{N-2} E_{\pi^{**},\pi^*}(Z_N)$$

where Z_N is the random reward from the N^{th} selection.

If π^* selects the known arm at the N^{th} selection, then $E_{\pi^*}(Z_N) = E_{\pi^{**},\pi^*}(Z_N) = B$ and hence

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left\{ V\left((z, d^{(1)}(y)), B, A_1^N\right) - V\left((z, d^{(1)}(y)), B, A_1^{N-1}\right) \right\}$$
$$\times h(y|s) f(z)\mathrm{d}z\mathrm{d}y$$
$$\geq \rho^{N-2} B \geq \int_{-\infty}^{\infty} \left\{ V\left((z,d), B, A_1^N\right) - V\left((z,d), B, A_1^{N-1}\right) \right\} f(z)\mathrm{d}z.$$

Suppose that π^* selects the unknown arm at the N^{th} selection. Since the sequence of posterior distributions of β forms a martingale, we have $E(\beta|d) = E(E(\beta|d^{(N)})|d)$. Therefore if the unknown arm is selected, we have

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left\{ V\left((z, d^{(1)}(y)), B, A_1^N\right) - V\left((z, d^{(1)}(y)), B, A_1^{N-1}\right) \right\} h(y|s) f(z)\mathrm{d}z\mathrm{d}y$$
$$\geq \rho^{N-1} E(\beta|d) E_{\pi^*}(Z_N) \geq \int_{-\infty}^{\infty} \left\{ V\left((z,d), B, A_1^N\right) - V\left((z,d), B, A_1^{N-1}\right) \right\} f(z)\mathrm{d}z.$$

Suppose finally that π^* selects the unknown arm at the N^{th} selection for the set G of z values. Let G^C be the compliment of G and I_G be the indicator function of G. Then

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left\{V\left((z,d^{(1)}(y)),B,A_1^N\right)-V\left((z,d^{(1)}(y)),B,A_1^{N-1}\right)\right\}h(y|s)f(z)\mathrm{d}z\mathrm{d}y$$
$$\geq \rho^{N-1}\int_{-\infty}^{\infty}\left[I_G E(\beta|d)E_{\pi^*}(Z_N)+I_{G^c}B\right]f(z)\mathrm{d}z$$
$$\geq \int_{-\infty}^{\infty}\left\{V\left((z,d),B,A_1^N\right)-V\left((z,d),B,A_1^{N-1}\right)\right\}f(z)\mathrm{d}z.$$

Here we have considered only deterministic strategies because there is a deterministic strategy which is optimal. In any case, we have shown that $\Delta\left((x,d),B,A_1^{N+1}\right)\geq 0$. □

Proof (of theorem 2) The inequality $B^*(x,d,A_1^N)\leq B^*(x,d,A_1^{N+1})$ follows from the above Lemma 3 and monotonicity of $\Delta\left((x,d),B,A_1^N\right)$ in B. The limit $B^*(x,d)=\lim_{N\to\infty}B^*(x,d,A_1^N)$ of a non-decreasing sequence $B^*(x,d,A_1^N)$, $N=1,2,\ldots,$ exists and satisfies the equation

$$\Delta\left(x,d,B^*(x,d),A\right)=\lim_{N\to\infty}\Delta\left(x,d,B^*(x,d,A_1^N),A_1^N\right)=0$$

for $A=(1,\rho,\rho^2,\ldots)$, due to uniformly bounded values of the expected payoffs.

If $B^*(x,d)=\infty$, then we select the known arm initially. Hence $V\left(x,d,B^*(x,d),A\right)=\infty$, contradicting the finiteness of the optimal value function. We prove $\beta^{(0)}x<B^*(x,d,A_1^2)$ and hence $\beta^{(0)}x<B^*(x,d)$. Suppose that $\beta^{(0)}x=B^*(x,d,A_1^2)$. Then the immediate expected payoff from the two arms are identical and cancel out. The expected payoff from the second, also the last, selection is the maximum of the expected payoffs from the two arms given updated posterior distributions. Therefore, assuming $\beta^{(0)}>0$, we have

$$\begin{aligned}
0&=\Delta\left((x,d),B^*(x,d,A_1^2),A_1^2\right)\\
&=\rho\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty}V((z,d^{(1)}(y)),B,A_1^1)h(y|s)\mathrm{d}y-V((z,d),B,A_1^1)\right]f(z)\mathrm{d}z\\
&=\rho\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty}\max\left\{\frac{M\beta^{(0)}+xy}{M+x^2}z,\beta^{(0)}x\right\}h(y|s)\mathrm{d}y-\max\left\{\beta^{(0)}z,\beta^{(0)}x\right\}\right]f(z)\mathrm{d}z\\
&=\rho\int_{-\infty}^{x}\left[\int_{-\infty}^{\infty}\max\left\{\frac{M\beta^{(0)}+xy}{M+x^2}z,\beta^{(0)}x\right\}h(y|s)\mathrm{d}y-\beta^{(0)}x\right]f(z)\mathrm{d}z\\
&\quad+\rho\int_{x}^{\infty}\left[\int_{-\infty}^{\infty}\max\left\{\frac{M\beta^{(0)}+xy}{M+x^2}z,\beta^{(0)}x\right\}h(y|s)\mathrm{d}y-\beta^{(0)}z\right]f(z)\mathrm{d}z.
\end{aligned}$$

Both integrands are clearly non-negative and strictly positive for certain values of z, so the right-hand side is strictly positive. Similar result is true if $\beta^{(0)}<0$. This is a contradiction. □

Acknowledgments The authors thank an anonymous Associate Editor and two anonymous referees for constructive comments which have significantly improved the presentation of the paper. Both Xikui

Wang and Yanqing Yi acknowledge research supports from the Natural Sciences and Engineering Research Council (NSERC) of Canada. Yanqing Yi acknowledges the IRIF Start-up Fund from the Government of Newfoundland and Labrador, through the Department of Innovation, Trade and Rural Development.

References

Bansal, A. K. (2007). *Bayesian parametric inference*. Oxford, UK: Alpha Science.
Berry, D. A., Fristedt, B. (1985). *Bandit problems*. London: Chapman and Hall.
Eick, S. G. (1988). The two-armed bandit with delayed responses. *Annals of Statistics*, *16*, 254–265.
Gittins, J. C. (1989). *Multi-armed bandit allocation indices*. Chichester: Wiley.
Gittins, J., Wang, Y.-G. (1992). The learning component of dynamic allocation indices. *Annals of Statistics*, *20*, 1625–1636.
Goldenshluger, A., Zeevi, A. (2009). Woodroofe's one-armed bandit problem revisited. *Annals of Applied Probability*, *19*, 1603–1633.
Hu, F., Rosenberger, W. F. (2006). *The theory of response-adaptive randomization in clinical trials*. Hoboken: Wiley.
Li, X., Wang, X. (2012). Variance-penalized response-adaptive randomization with mismeasurement. *Journal of Statistical Planning and Inference*, *142*, 2128–2135.
Sarkar, J. (1991). One-armed bandit problems with covariates. *Annals of Statistics*, *19*, 1978–2002.
Wang, X. (2000). A bandit process with delayed responses. *Statistics and Probability Letters*, *48*, 303–307.
Wang, X. (2007). Dynamic pricing with a Poisson bandit model. *Sequential Analysis*, *26*, 355–365.
Wang, X., Bickis, M. G. (2003). One-armed bandit models with continuous and delayed responses. *Mathematical Methods in Operations Research*, *58*, 209–219.
Wang, X., Wang, Y. (2010). Optimal investment and consumption with stochastic dividends. *Applied Stochastic Models in Business and Industry*, *26*, 792–808.
Wang, X., Yi, Y. (2009). An optimal investment and consumption model with stochastic returns. *Applied Stochastic Models in Business and Industry, 25*, 45–55.
Wang, Y.-G., Gittins, J. (1992). Bayesian bandits in clinical trials. *Sequential Analysis*, *11*, 313–325.
Woodroofe, M. (1979). One-armed bandit problem with a concomitant variable. *Journal of American Statistical Association*, *74*, 799–806.
Yang, Y., Zhu, D. (2002). Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates. *Annals of Statistics*, *30*, 100–121.
Yi, Y., Wang, X. (2009). Response adaptive designs with a variance-penalized criterion. *Biometrical Journal*, *5*, 763–773.