Beta-Stacy Bandit Problems with Censored Data

Stefano Peluso * Antonietta Mira † Pietro Muliere ‡

December 14, 2016

Abstract

Existing Bayesian nonparametric methodologies for bandit problems focus on exact observations, leaving a gap in those bandit applications where censored observations are crucial. We address this gap by extending a Bayesian nonparametric two-armed bandit problem to right-censored data, where each arm is generated from a beta-Stacy process as defined by Walker and Muliere (1997). We prove the existence of optimal stay-with-a-winner and switch-on-a-loser strategies, by imposing non-restrictive conditions on the parameters of the beta-Stacy processes, including the special cases of the homogeneous process and the Dirichlet process. Numerical estimations and simulations for a variety of discrete and continuous state space settings are presented to illustrate the performance and flexibility of our method.

1 Introduction

In a discrete-time two-armed bandit problem, there are two stochastic processes (the two arms) and a sequential decision process (a strategy) that selects, at each time, which one of the two processes to observe. This selection is made on the basis of the previous observations, and it balances two conflicting benefits: the immediate payoff coming from the exploitation of an arm (so far) known to be better and the information concerning future payoffs coming from the exploration of a less known arm. A strategy is said to be *optimal* if it yields the maximal expected payoff, and an arm is said to be optimal if it is selected at the beginning of an optimal strategy.

A strategy can be seen as a function that assigns, to each partial history of observations, the integer 1 or 2 indicating the arm to be observed at the next stage (Berry and Fristedt 1985). With the exception of the simplest cases, explicit specifications of optimal strategies are hindered by computational issues. As a consequence, following Chattopadhyay (1994), optimal strategies can only be partially characterized in terms of *break-even* observations. We will consider two kinds of optimal strategies: stay-with-a-winner and switch-on-a-loser strategies. Assuming, without loss of generality, that a higher realized value of a random variable gives a higher payoff in the bandit problem, the break-even observation in a *stay-with-a-winner* strategy is that value at which the

*Corresponding author. Università Cattolica del Sacro Cuore, Department of Statistical Sciences, E-mail: stefano.peluso@unicatt.it. Largo Gemelli, 1 20136 Milan, Italy. Università della Svizzera Italiana, InterDisciplinary Institute of Data Science.

†Università della Svizzera Italiana, InterDisciplinary Institute of Data Science. Università degli Studi dell'Insubria, Department of Sciences and High Technology.

‡Università Commerciale Luigi Bocconi, Department of Decision Sciences. Bocconi Institute for Data Science and Analytics (BIDSA).

expected advantage of choosing arm 1 over arm 2 is null. Then the expected advantage is positive (negative) for values observed from arm 1 greater (lower) than the break-even, and at these values arm 1 (arm 2) is chosen at the next stage a optimal. On the other hand, the break-even observation in a *switch-on-a-loser* strategy is that realized value at which the expected advantage remains unaltered at the next stage. For values observed from arm 1 higher (lower) than the break-even, the expected advantage increases (decreases) relative to its initial value, and arm 1 (arm 2) is optimally chosen at the next stage.

More formally, let X_j and Y_j be random variables generated from, respectively, arm 1 and 2 at stage j. For any positive integer k, $X_1, X_2, \ldots, X_k$ given F_1 are i.i.d with probability measure F_1, and $Y_1, Y_2, \ldots, Y_k$ given F_2 are i.i.d with probability measure F_2. The (possibly infinite) bandit horizon is n and $A_n = (a_1, a_2, \ldots, a_n)$ is a nonincreasing sequence of discount factors.

Early examples of bandit problems are treated in Robbins (1952), Bellman (1956) and Bradt et al. (1956). Among later works, Chernoff (1968) focuses on two Gaussian arms $F_i = N(\mu_i, \sigma^2), i = 1, 2$, with unknown drifts and known constant variance; Berry (1972) gives sufficient conditions for optimal selection in a Bernoulli two-armed bandit, $F_i = Bern(p_i), i = 1, 2$, proving a stay-with-a-winner strategy; Berry and Fristedt (1979) characterize optimal strategies for Bernoulli one-armed bandits ($F_1 = Bern(p)$ and F_2 known) with regular discount sequences; Gittins (1979) introduces dynamic allocation indices for optimal strategies in multi-armed bandits. Clayton and Berry (1985) is the first paper that extends the bandit problem to a Bayesian nonparametric framework, considering a random $F_1 \sim DP(\alpha)$ and known F_2: the probability measure associated to the random variables in one of the two arms is random and extracted from the Dirichlet process introduced in Ferguson (1973). Dirichlet bandits are generalized to two-armed problems $F_i \sim DP(\alpha_i), i = 1, 2$, in Chattopadhyay (1994), where the existence of stay-with-a-winner and switch-on-a-loser optimal strategies is proven. Some other properties of Dirichlet bandits are studied in Yu (2011).

In the this paper we extend Bayesian nonparametric bandits to problems where each arm generates an infinite sequence of exchangeable random variables (de Finetti 1937) having, as de Finetti measure, the beta-Stacy process of Walker and Muliere (1997). In our framework the two arms are random, with $F_i \sim BS(\alpha^i, \beta^i)$, $i = 1, 2$, where α^i and β^i, discussed in Section 2, characterize the two beta-Stacy processes. The Dirichlet bandit of Clayton and Berry (1985) and Chattopadhyay (1994) is an important special case of our setting, as is the bandit arms with the homogeneous process of Susarla and Van Ryzin (1976) and Ferguson and Phadia (1979). Our main result is that, under constraints on the parameters of the beta-Stacy processes (constraints that include the cases of the homogeneous process and the Dirichlet process), optimal stay-with-a-winner and switch-on-a-loser strategies exist and can be used for dealing with right censored or exact observations. As specified in Phadia (2013), the beta-Stacy process belongs to the class of neutral to the right (NTR) processes introduced by Doksum (1974), and it generalizes the Dirichlet process in two respects: more flexible prior information may be represented and, unlike the Dirichlet process, it is conjugate to right censored data. Also, when the prior process is assumed to be Dirichlet, the posterior distribution given right censored observations is a beta-Stacy process. Beta-Stacy bandit problems are motivated by the importance of dealing with censored observations in typical bandit applications: the two arms can be two treatments available for a certain disease (Berry and Fristedt 1985); patients arrive one at a time and a treatment is assigned. The patient returns information on the effectiveness of the treatment, but this response can be censored, e.g. if the patient interrupts the treatment. Another classical example of a setting where censored

observations may arise is managing a team of industrial scientists working on several research projects, and deciding which sequence of project executions would maximize the expected total value (Nash 1973): the observed project value is censored if the project is interrupted due, for instance, to reduced financial support. A final example of a bandit problem with censored data is that of an industrial processor choosing which jobs to process to minimize the processing time (Gittins et al. 2011 and references therein): jobs may return censored observations if the task is unfinished for system breakdowns. The arms introduced in the present paper have random distribution functions generated by beta-Stacy processes, permitting the analysis of bandit problems with censored observations, while retaining mathematical tractability.

In Section 2 we introduce the beta-Stacy process and the bandit problem. Then, we show the existence of stay-with-a-winner and switch-on-a-loser strategies, for $\alpha^1(\cdot)$ and $\alpha^2(\cdot)$ having discrete (Section 3) and continuous support (Section 4). We apply our methods to simulated studies in Section 5, and conclude with examples of potential further applications and research directions in Section 6.

2 Preliminaries

2.1 Beta-Stacy process

Let $\mathfrak{F}$ be the space of cumulative distribution functions (cdf's) on $[0,\infty)$. A probability distribution is placed on $\mathfrak{F}$ through the definition of a stochastic process F on $([0,\infty),\mathcal{A})$, where $\mathcal{A}$ is the Borel σ-field of subsets, such that, with probability 1, the sample paths of F are cdf's. Let the right continuous measure $\alpha(\cdot)$, with $\alpha(0)=0$, and the positive function $\beta(\cdot)$ both be defined on $[0,\infty)$, and let $\{t_k\}$ be the countable set of discontinuity points of $\alpha(\cdot)$, such that $\alpha\{t_k\}=\alpha(t_k)-\alpha(t_{k^-})=S_k>0$ for all k, and S_k is the jump in t_k. Let $\alpha_c(t)=\alpha(t)-\sum_{t_k\leq t}\alpha\{t_k\}$, so that $\alpha_c(\cdot)$ is a continuous measure. Note that in the rest of the paper, whenever clear, the explicit dependence on $\alpha(\cdot)$ and on $\beta(\cdot)$ of all quantities of interest has been omitted for notational clarity.

Definition 2.1. *F is a beta-Stacy process on $([0,\infty),\mathcal{A})$ with parameters $\alpha(\cdot)$ and $\beta(\cdot)$, that is $F\sim BS(\alpha,\beta)$, if for all $t\geq 0$, $F(t)=1-\exp\{-Z(t)\}$, where Z is a Lévy process with Lévy measure for $Z(t)$ given, for $v>0$, by*

$$dN_t(v)=\frac{dv}{1-\exp(-v)}\int_0^t \exp(-v(\beta(s)+\alpha\{s\})d\alpha_c(s))$$

and with moment generating function given by

$$\log E\left[-\phi Z(t)\right]=\sum_{t_k\leq t} logE\left[\exp(-\phi S_k)\right]+\int_0^t(\exp(-\phi v)-1)dN_t(v),$$

where $1-\exp(S_k)\sim Beta(\alpha\{t_k\},\beta(t_k))$.

We now state the two theorems from Walker and Muliere (1997) we will use in the sequel, on the conjugacy of the beta-Stacy process, distinguishing between discrete and continuous beta-Stacy parameters.

Theorem 2.2. *(Walker and Muliere 1997) The posterior distribution of a beta-Stacy process with discrete parameters $\{\alpha_j,\beta_j\}$, $j\in\mathbb{N}$, is also a beta-Stacy process, with parameters α_j+n_j and*

$\beta_j + m_j$, *where* n_j *is the number of exact observations at* j *and* m_j *is the sum of the number of exact observations in* $\{l:\ l > j\}$ *and censored observations in* $\{l:\ l \geq j\}$.

The theorem above clarifies an important property of the beta-Stacy process: its conjugacy under sampling, possibly with right censoring. Furthermore, to have a.s. a cdf, the discrete parameters α and β in Walker and Muliere (1997) are required to satisfy the condition

$$\prod_j \left(1 - \frac{\alpha_j}{\beta_j + \alpha_j}\right) = 0, \quad j \in \mathbb{N}. \tag{1}$$

In Theorem 4 of Walker and Muliere (1997) the conjugacy under sampling with possible right censorship is formalized in the continuous case:

Theorem 2.3. *(Walker and Muliere 1997) The posterior distribution of a beta-Stacy process with continuous parameters* $\{\alpha(\cdot), \beta(\cdot)\}$ *is also a beta-Stacy process, with parameters* $\alpha(t) + N\{t\}$ *and* $\beta(t) + M(t)$*, where* $N\{s\}$ *is the counting process for uncensored observations in* s*, and* $M(s)$ *is the sum of the number of exact observations in* $\{t:\ t > s\}$ *and of censored observations in* $\{t:\ t \geq s\}$.

A posteriori the beta-Stacy process with continuous $\alpha(\cdot)$ and $\beta(\cdot)$ can be represented as $F(t) = 1 - e^{-(Z_c(t) + Z_f(t))}$ (Ferguson 1974; Walker and Damien 1998), where the component Z_f incorporates the fixed points of discontinuity in the locations of the exact observations, and Z_c is the residual part of the Lévy process. In particular, the corresponding jump S_i at location t_i, where the exact observation occurred, is such that

$$1 - \exp(-S_i) \sim Beta(N\{t_i\}, \beta(t_i) + M(t_i)).$$

Furthermore, from Walker and Muliere (1997), continuous $\alpha(\cdot)$ and $\beta(\cdot)$ satisfy the condition

$$\int_0^\infty d\alpha(s)/\beta(s) = \infty.$$

In the rest of the paper, we only consider beta-Stacy processes, but it is reasonable to assume that the results can be generalized to the class of NTR processes. The NTR process of Doksum (1974) may be viewed in terms of a process with independent non-negative increments, via the parameterization $F(t) = 1 - e^{-Z(t)}$, $t \in \mathbb{R}^+$, where Z is a process with independent nonnegative increments. The beta-Stacy process is a NTR process where Z is a log-beta process, that keeps the conjugacy property under sampling exact or right censored observations. When $\beta(t) = \alpha([t, \infty))$ for all $t \geq 0$, we obtain the Dirichlet process of Ferguson (1973). Another important special case is the homogenous process of Susarla and Van Ryzin (1976) and Ferguson and Phadia (1979), arising when $\beta(t) = \beta$ constant for all t.

2.2 Bandit problem

In the proposed framework $(\{\alpha^1, \beta^1\}, \{\alpha^2, \beta^2\}; A_n)$ denote the two-armed bandit problem where arm i has a beta-Stacy prior with parameters (α^i, β^i), for $i = 1, 2$, and $A_n = (a_1, a_2, \ldots, a_n)$ is a nonincreasing discount sequence. Therefore $F_1 \sim BS(\alpha^1, \beta^1)$ and the observation at stage 1 from arm 1 is a realization of the random variable $X_1 | F_1 \sim F_1$. At the generic k-th stage, if the sample $\underline{x}$ has been collected in the past stages from the observation of arm 1, the new observation

will be the realized value of $X_k|F_1, \underline{x} \sim F_1|\underline{x}$, where $F_1|\underline{x}$ is still a random distribution from a beta-Stacy process with updated parameters, thanks to the conjugacy theorems provided in the previous subsection. Note that the sample $\underline{x}$ can be partitioned in censored and exact observations: $\underline{x} = [\underline{x}_{exact}, \underline{x}_{cens}]$. Equivalent definitions hold for $Y_1, Y_2, \ldots, Y_k$ and the observed sample $\underline{y}$ from arm 2.

By censored observation we mean that the observed value at stage k from arm 1 is the realized value of $X_k = \min\{X_k^*, C_k\}$, the minimum between a true exact value at stage k and a censoring time C_k, and equivalently for Y_k from arm 2.

A special choice of β^i, detailed below, and absence of censored observations reduce our setting to the Dirichlet bandit problem $(\{\alpha^1\}, \{\alpha^2\}; A_n)$ of Chattopadhyay (1994).

Similarly to Berry and Fristedt (1985), we let $W(\{\alpha^1, \beta^1\}, \{\alpha^2, \beta^2\}; A_n)$ be the expected payoff under an optimal strategy; $W^i(\{\alpha^1, \beta^1\}, \{\alpha^2, \beta^2\}; A_n)$ the expected payoff of a strategy starting from arm i and proceeding optimally; $\Delta(\{\alpha^1, \beta^1\}, \{\alpha^2, \beta^2\}; A_n)$ the expected advantage of initially choosing arm 1 over arm 2 assuming optimal continuation in both stay-with-a-winner and switch-on-a-loser strategies; finally

$$\Delta^+(\{\alpha^1, \beta^1\}, \{\alpha^2, \beta^2\}; A_n) = \max(0, \Delta(\{\alpha^1, \beta^1\}, \{\alpha^2, \beta^2\}; A_n))$$

and

$$\Delta^-(\{\alpha^1, \beta^1\}, \{\alpha^2, \beta^2\}; A_n) = \min(0, \Delta(\{\alpha^1, \beta^1\}, \{\alpha^2, \beta^2\}; A_n)).$$

More generally, we define the discount sequence $A_n^k = (a_{k+1}, a_{k+2}, \ldots, a_n)$, and we now introduce the two strategies that we study: stay-with-a-winner and switch-on-a-loser strategies.

Definition 2.4. *For $X_k|F_1 \sim F_1$, $Y_k|F_2 \sim F_2$, $F_1 \sim BS(\alpha^1, \beta^1)$ and $F_2 \sim BS(\alpha^2, \beta^2)$, at the generic stage $k \geq 1$, the stay-with-the winner strategy selects arm 1 at stage $k+1$ if one of the following two conditions holds:*

- *at stage k, $X_k = x$ exact or censored from arm 1 is observed and*
$$\Delta\left(\{\alpha^{1,x}, \beta^{1,x}\}, \{\alpha^2, \beta^2\}; A_n^1\right) \geq \Delta\left(\{\alpha^1, \beta^1\}, \{\alpha^2, \beta^2\}; A_n\right),$$
- *at stage k, $Y_k = y$ exact or censored from arm 2 is observed and*
$$\Delta\left(\{\alpha^1, \beta^1\}, \{\alpha^{2,y}, \beta^{2,y}\}; A_n^1\right) \geq \Delta\left(\{\alpha^1, \beta^1\}, \{\alpha^2, \beta^2\}; A_n\right),$$

and selects arm 2 otherwise.
The switch-on-a-loser strategy selects arm 1 at stage $k+1$ if one of the following two conditions holds:

- *at stage k, $X_k = x$ exact or censored from arm 1 is observed and*
$$\Delta\left(\{\alpha^{1,x}, \beta^{1,x}\}, \{\alpha^2, \beta^2\}; A_n^1\right) \geq 0,$$
- *at stage k, $Y_k = y$ exact or censored from arm 2 is observed and*
$$\Delta\left(\{\alpha^1, \beta^1\}, \{\alpha^{2,y}, \beta^{2,y}\}; A_n^1\right) \geq 0,$$

  and selects arm 2 otherwise.

Both strategies at stage $k = 1$ select arm 1 if $\Delta\left(\{\alpha^1, \beta^1\}, \{\alpha^2, \beta^2\}; A_n\right) > 0$ and arm 2 otherwise.

The break-even point of a strategy is that realized value of X from the first arm (or Y from the second arm) at which the strategy is indifferent in the choice of the two arms, and a strategy is said to exist if its break-even point exists.

3 Break-Even Observations: Discrete Case

The optimal strategy is given in terms of a break-even observation: the first (second) arm is observed until it yields a value higher (lower) than the break-even one. In this section we prove that a break-even observation for a stay-with-a-winner and a switch-on-a-loser optimal strategies exist for the discrete beta-Stacy two-armed bandit problem, under certain conditions on the parameters of the processes governing the arms.

Let F_i the random distribution function corresponding to arm i, with X_k and Y_k having supports in $\mathbb{N}$ for all $k \geq 1$. Let $\alpha^i = (\alpha_1^i, \alpha_2^i, \dots)$ and $\beta^i = (\beta_1^i, \beta_2^i, \dots)$ for $i = 1, 2$.

From the construction of the discrete time beta-Stacy process, Walker and Muliere (1997) show that $\mathbb{P}(X_1 = j|\{\alpha^1, \beta^1\})$, $j \in \mathbb{N}$, is

$$\mathbb{P}(X_1 = j|\{\alpha^1, \beta^1\}) = \frac{\alpha_j^1}{\alpha_j^1 + \beta_j^1} \prod_{i=1}^{j-1} \left(1 - \frac{\alpha_i^1}{\alpha_i^1 + \beta_i^1}\right)$$

and similarly for Y_1. Then the prior means of the two arms are respectively

$$\begin{aligned} E_X[X|\{\alpha^1, \beta^1\}] &= \sum_{j=1}^{+\infty} \mathbb{P}(X \geq j|\{\alpha^1, \beta^1\}) \\ &= \sum_{j=1}^{+\infty} \prod_{i<j} \left(1 - \frac{\alpha_i^1}{\alpha_i^1 + \beta_i^1}\right) =: \mu_1 \end{aligned}$$

and

$$E_Y[Y|\{\alpha^2, \beta^2\}] = \sum_{j=1}^{+\infty} \prod_{i<j} \left(1 - \frac{\alpha_i^2}{\alpha_i^2 + \beta_i^2}\right) =: \mu_2.$$

Given observations

$$\begin{aligned} \underline{x} &= (x_1, \dots, x_k) \text{ from arm 1,} \\ \underline{y} &= (y_1, \dots, y_k) \text{ from arm 2,} \end{aligned}$$

the conditional expectation of any function $h(X)$ can be computed using Theorem 2.2, and it is denoted with $E_X[h(X)|\underline{x}]$, where we remind that we omit in all quantities of interest the dependence from $\{\alpha^1, \beta^1\}$ and $\{\alpha^2, \beta^2\}$. Then $\mathbb{P}(X = j|\underline{x})$ is therefore, for $j \in \mathbb{N}$,

$$P(X = j|\underline{x}) = \frac{\alpha_j^{1,\underline{x}}}{\alpha_j^{1,\underline{x}} + \beta_j^{1,\underline{x}}} \prod_{i=1}^{j-1} \left(1 - \frac{\alpha_i^{1,\underline{x}}}{\alpha_i^{1,\underline{x}} + \beta_i^{1,\underline{x}}}\right)$$

where

$$\begin{aligned} \alpha_j^{1,\underline{x}} &:= \alpha_j^1 + n_j, \\ \beta_j^{1,\underline{x}} &:= \beta_j^1 + m_j, \end{aligned}$$

with

$$\begin{aligned} n_j &: \text{number of exact observations equal to } j, \\ m_j &: \text{number of exact observations in } \{l : \ l > j\} \text{ and} \\ &\quad \text{censored observations in } \{l : \ l \geq j\}, \end{aligned}$$

and with dependence of n_j and m_j from $\underline{x}$ neglected for ease of notation. The posterior mean is therefore

$$E_X[X|\underline{x}] = \sum_{j=1}^{+\infty} \mathbb{P}(X \geq j|\underline{x}) = \sum_{j=1}^{+\infty} \prod_{i<j} \left(1 - \frac{\alpha_i^{1,\underline{x}}}{\alpha_i^{1,\underline{x}} + \beta_i^{1,\underline{x}}}\right) =: \mu_{1,\underline{x}}$$

Following the notation introduced above,

$$\begin{aligned}
&W(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_n)\\
&\quad = \max\left\{W^1(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_n), W^2(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_n)\right\},\\
&\Delta(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_n)\\
&\quad = W^1(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_n) - W^2(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_n),\\
&W^1(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_n)\\
&\quad = W(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_n) + \Delta^-(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_n),\\
&W^2(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_n)\\
&\quad = W(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_n) - \Delta^+(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_n).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&W^1(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_n)\\
&\quad = a_1\mu_1 + E_X\left[W(\{\alpha^{1,X},\beta^{1,X}\},\{\alpha^2,\beta^2\};A_n^1)\right]\\
&\quad = a_1\mu_1 + E_X\left[W^2(\{\alpha^{1,X},\beta^{1,X}\},\{\alpha^2,\beta^2\};A_n^1)\right]\\
&\qquad + E_X\left[\Delta^+(\{\alpha^{1,X},\beta^{1,X}\},\{\alpha^2,\beta^2\};A_n^1)\right],\\
&W^2(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_n)\\
&\quad = a_1\mu_2 + E_Y\left[W(\{\alpha^1,\beta^1\},\{\alpha^{2,Y},\beta^{2,Y}\};A_n^1)\right]\\
&\quad = a_1\mu_2 + E_Y\left[W^1(\{\alpha^1,\beta^1\},\{\alpha^{2,Y},\beta^{2,Y}\};A_n^1)\right]\\
&\qquad - E_Y\left[\Delta^-(\{\alpha^1,\beta^1\},\{\alpha^{2,Y},\beta^{2,Y}\};A_n^1)\right],
\end{aligned}$$

and

$$\begin{aligned}
&\Delta(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_n)\\
&\quad = W^1(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_n) - W^2(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_n)\\
&\quad = a_1\mu_1 + E_X\left[W^2(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_n^1)\right]\\
&\qquad - a_1\mu_2 - E_Y\left[W^1(\{\alpha^1,\beta^1\},\{\alpha^{2,Y},\beta^{2,Y}\};A_n^1)\right]\\
&\qquad + E_X\left[\Delta^+(\{\alpha^{1,X},\beta^{1,X}\},\{\alpha^2,\beta^2\};A_n^1)\right]\\
&\qquad + E_Y\left[\Delta^-(\{\alpha^1,\beta^1\},\{\alpha^{2,Y},\beta^{2,Y}\};A_n^1)\right].
\end{aligned}$$

Using arguments similar to those in Berry and Fristedt (1985) and Chattopadhyay (1994),

$$a_1\mu_1 + E_X\left[W^2(\{\alpha^{1,X},\beta^{1,X}\},\{\alpha^2,\beta^2\};A_n^1)\right]$$

is the expected payoff of first selecting arm 1, followed by arm 2 and then continuing optimally. Similarly,

$$a_1\mu_2 + E_Y\left[W^1(\{\alpha^1,\beta^1\},\{\alpha^{2,Y},\beta^{2,Y}\};A_n^1)\right]$$

is the expected payoff of selecting arm 2 first and arm 1 second and then continuing optimally. Subtracting the second payoff from the first one we obtain $(a_1 - a_2)(\mu_1 - \mu_2)$. From this fact,

$$\begin{aligned}\Delta(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_n) &= (a_1-a_2)(\mu_1-\mu_2)\\ &\quad +E_X\left[\Delta^+(\{\alpha^{1,X},\beta^{1,X}\},\{\alpha^2,\beta^2\};A_n^1)\right]\\ &\quad +E_Y\left[\Delta^-(\{\alpha^1,\beta^1\},\{\alpha^{2,Y},\beta^{2,Y}\};A_n^1)\right] \qquad (2)\end{aligned}$$

In the next proposition we show that, given an exact or a right censored observation $X_1 = x$ from arm 1, the advantage of choosing arm 1 over arm 2 increases as x increases.

Proposition 3.1. *For all $\{\alpha^1,\beta^1\}$ and $\{\alpha^2,\beta^2\}$ such that $\beta_j^1 \le \beta_{j+1}^1 + \alpha_{j+1}^1$, for all $j \in \mathbb{N}$, and for all nonincreasing discount sequences A_n,*

$$\Delta\left(\{\alpha^{1,x},\beta^{1,x}\},\{\alpha^2,\beta^2\};A_n\right)$$

is nondecreasing in x.

Proof. By induction, for $n = 1$, we have

$$\begin{aligned}\Delta\left(\{\alpha^{1,x},\beta^{1,x}\},\{\alpha^2,\beta^2\};A_1\right) &= a_1\left(\sum_{j=1}^{+\infty}\prod_{i<j}\left(1-\frac{\alpha_i^{1,x}}{\alpha_i^{1,x}+\beta_i^{1,x}}\right)-\mu_2\right)\\ &= a_1(\mu_{1,x}-\mu_2). \qquad (3)\end{aligned}$$

Fix $x^* = x + 1$. We first prove that $\mu_{1,x^*} - \mu_{1,x} \ge 0$. For this purpose, we study separately the j-terms in the sum of $\mu_{1,x}$ and μ_{1,x^*} when $j \le x$, $j = x^*$ and $j > x^*$. When x is an exact observation,

- The j-terms with $j \le x$ are the same in $\mu_{1,x}$ and μ_{1,x^*}.
- For $j = x^*$, in $\mu_{1,x}$ we have

  $$\prod_{i<j}\left(\frac{\beta_i^1+1}{\alpha_i^1+\beta_i^1+1}\right)\frac{\beta_x^1}{\alpha_x^1+\beta_x^1+1},$$

  whilst in μ_{1,x^*},

  $$\prod_{i<j}\left(\frac{\beta_i^1+1}{\alpha_i^1+\beta_i^1+1}\right)\frac{\beta_x^1+1}{\alpha_x^1+\beta_x^1+1},$$

  and the x^*-term of μ_{1,x^*} is weakly higher.
- For $j > x^*$, the j-term of $\mu_{1,x}$ is

  $$\prod_{i<x}\left(\frac{\beta_i^1+1}{\alpha_i^1+\beta_i^1+1}\right)\frac{\beta_x^1}{\alpha_x^1+\beta_x^1+1}\frac{\beta_{x^*}^1}{\alpha_{x^*}^1+\beta_{x^*}^1}\prod_{x^*<i<j}\frac{\beta_i^1}{\alpha_i^1+\beta_i^1},$$

  whilst the j-term of μ_{1,x^*} is

  $$\prod_{i<x}\left(\frac{\beta_i^1+1}{\alpha_i^1+\beta_i^1+1}\right)\frac{\beta_x^1+1}{\alpha_x^1+\beta_x^1+1}\frac{\beta_{x^*}^1}{\alpha_{x^*}^1+\beta_{x^*}^1+1}\prod_{x^*<i<j}\frac{\beta_i^1}{\alpha_i^1+\beta_i^1};$$

the j-term of μ_{1,x^*} is weakly higher if

$$\frac{\beta^1_x+1}{\alpha^1_{x^*}+\beta^1_{x^*}+1} \geq \frac{\beta^1_x}{\alpha^1_{x^*}+\beta^1_{x^*}},$$

equivalent to $\beta^1_x \leq \alpha^1_{x^*}+\beta^1_{x^*}$, for all x and for all $x^* > x$.

Similarly, the monotonicity of $\mu_{1,x}$ can be proved when x is a right censored observation: the j-terms with $k \leq x^*$ are the same in $\mu_{1,x}$ and μ_{1,x^*}, whilst for $j > x^*$ the two terms in, respectively, $\mu_{1,x}$ and μ_{1,x^*} are

$$\prod_{i<x}\left(\frac{\beta^1_i+1}{\alpha^1_i+\beta^1_i+1}\right)\frac{\beta^1_x+1}{\alpha^1_x+\beta^1_x+1}\frac{\beta^1_{x^*}}{\alpha^1_{x^*}+\beta^1_{x^*}}\prod_{x^*<i<j}\frac{\beta^1_i}{\alpha^1_i+\beta^1_i},$$
$$\prod_{i<x}\left(\frac{\beta^1_i+1}{\alpha^1_i+\beta^1_i+1}\right)\frac{\beta^1_x+1}{\alpha^1_x+\beta^1_x+1}\frac{\beta^1_{x^*}+1}{\alpha^1_{x^*}+\beta^1_{x^*}+1}\prod_{x^*<i<j}\frac{\beta^1_i}{\alpha^1_i+\beta^1_i},$$

where the term in μ_{1,x^*} is weakly higher. Then, for $n = 1$ the statement is true since $\mu_{1,x}$ is nondecreasing in x and $a_1 \geq 0$. From the induction hypothesis, we assume the monotonic property for $n = m-1$. By (2),

$$\begin{aligned}\Delta(\{\alpha^{1,x},\beta^{1,x}\},\{\alpha^2,\beta^2\};A_m) &= (a_1-a_2)(\mu_{1,x}-\mu_2)\\ &+E_X\left[\Delta^+(\{\alpha^{1,(x,X)},\beta^{1,(x,X)}\},\{\alpha^2,\beta^2\};A^1_m)\right]\\ &+E_Y\left[\Delta^-(\{\alpha^{1,x},\beta^{1,x}\},\{\alpha^{2,Y},\beta^{2,Y}\};A^1_m)\right]. \end{aligned} \tag{4}$$

The first term in the right hand side of (4) is nondecreasing in x since $\mu_{1,x}$ is nondecreasing in x and $a_1 - a_2 \geq 0$. The second and third term are nondecreasing in x from the induction hypothesis. □

Remark 3.2. The constraints $\beta^1_j \leq \beta^1_{j+1}+\alpha^1_{j+1}$ are needed for the monotonicity of $\mu_{1,x}$. The condition is not required if all observations are censored, but is necessary if some observations are exact. The constraint is naturally verified in the Dirichlet two-armed problem, obtained from the beta-Stacy in the special case of $\beta^1_j = \beta^1_{j+1}+\alpha^1_{j+1}$, $j \in \mathbb{N}$. Also, a bandit problem with simple homogeneous processes (Susarla and Van Ryzin 1976; Ferguson and Phadia 1979) for each arm, corresponding to the case $\beta^1_{j+1} = \beta^1_j$ for all $j \in \mathbb{N}$, satisfies the constraints.

The following propositions will be used in Theorems 3.5 and 3.6.

Proposition 3.3. *For all $\{\alpha^1,\beta^1\}$ and $\{\alpha^2,\beta^2\}$ as in Proposition 3.1 and for all nonincreasing discount sequences A_n, if the condition*

$$\prod_{i<+\infty}\left(1-\frac{\alpha^1_i}{\alpha^1_i+\beta^1_i+1}\right) > 0 \tag{5}$$

is verified, then

$$\lim_{x\to+\infty}\Delta\left(\{\alpha^{1,x},\beta^{1,x}\},\{\alpha^2,\beta^2\};A_n\right) = \infty$$

and

$$\lim_{x\to 0}\Delta\left(\{\alpha^{1,x},\beta^{1,x}\},\{\alpha^2,\beta^2\};A_n\right) = \min_x \Delta\left(\{\alpha^{1,x},\beta^{1,x}\},\{\alpha^2,\beta^2\};A_n\right).$$

Proof. The result for the $x \to 0$ is a direct consequence of the monotonicity property shown in Proposition 3.1. We are left to prove the limit to $+\infty$. Consider x increasing to ∞. By induction, for $n = 1$, $\mu_{1,x}$ diverges to $+\infty$ as $x \to +\infty$ since

$$\begin{aligned}\lim_{x\to+\infty} \mu_{1,x} &\geq \sum_{j=1}^{+\infty} \prod_{i<+\infty} \frac{\beta_i^1 + 1}{\alpha_i^1 + \beta_i^1 + 1} \\ &= \prod_{i<+\infty} \left(1 - \frac{\alpha_i^1}{\alpha_i^1 + \beta_i^1 + 1}\right) \cdot \sum_{j=1}^{+\infty} 1 = +\infty. \end{aligned} \tag{6}$$

Then, $\Delta\left(\{\alpha^{1,x}, \beta^{1,x}\}, \{\alpha^2, \beta^2\}; A_1\right) = a_1(\mu_{1,x} - \mu_2)$ goes to $+\infty$ since $\mu_{1,x}$ is divergent and $a_1 > 0$. Assume now that the statement is true for $n = m - 1$. By (4),

$$\begin{aligned}\Delta(\{\alpha^{1,x}, \beta^{1,x}\}, \{\alpha^2, \beta^2\}; A_m) &= (a_1 - a_2)(\mu_{1,x} - \mu_2) \\ &+ E_X\left[\Delta^+(\{\alpha^{1,(x,X)}, \beta^{1,(x,X)}\}, \{\alpha^2, \beta^2\}; A_m^1)\right] \\ &+ E_Y\left[\Delta^-(\{\alpha^{1,x}, \beta^{1,x}\}, \{\alpha^{2,Y}, \beta^{2,Y}\}; A_m^1)\right].\end{aligned}$$

For the first term $(a_1 - a_2)(\mu_{1,x} - \mu_2)$ on the right hand side of the formula above there are two possible cases: $a_1 - a_2 > 0$ or $a_1 - a_2 = 0$. In the latter case the term is zero, while when $a_1 - a_2 > 0$ it diverges to $+\infty$. For the second term, note that $\Delta^+(\{\alpha^{1,(x,X)}, \beta^{1,(x,X)}\}, \{\alpha^2, \beta^2\}; A_m^1)$ is a nondecreasing sequence in x (by Proposition 3.1), bounded below by 0 (by definition) and divergent to $+\infty$ (by the induction hypothesis). We can then apply the monotone convergence theorem and obtain

$$\begin{aligned}&\lim_{x\to+\infty} E_X\left[\Delta^+(\{\alpha^{1,(x,X)}, \beta^{1,(x,X)}\}, \{\alpha^2, \beta^2\}; A_m^1)\right] \\ &= E_X\left[\lim_{x\to+\infty} \Delta^+(\{\alpha^{1,(x,X)}, \beta^{1,(x,X)}\}, \{\alpha^2, \beta^2\}; A_m^1)\right] = +\infty.\end{aligned}$$

For the third term, notice that

$$\Delta(\{\alpha^{1,x}, \beta^{1,x}\}, \{\alpha^{2,y}, \beta^{2,y}\}; A_m^1) = -\Delta(\{\alpha^{2,y}, \beta^{2,y}\}, \{\alpha^{1,x}, \beta^{1,x}\}; A_m^1).$$

Furthermore, $\Delta^+(\{\alpha^{2,y}, \beta^{2,y}\}, \{\alpha^{1,x}, \beta^{1,x}\}; A_m^1)$ converges to 0 as l diverges, and it is bounded above by $\left|\Delta^+(\{\alpha^{2,y}, \beta^{2,y}\}, \{\alpha^{1,x=0}, \beta^{1,x=0}\}; A_m^1)\right|$. By the dominated convergence theorem we have

$$\begin{aligned}&\lim_{x\to+\infty} E_Y\left[\Delta^-(\{\alpha^{1,x}, \beta^{1,x}\}, \{\alpha^{2,Y}, \beta^{2,Y}\}; A_m^1)\right] \\ &= -\lim_{x\to+\infty} E_Y\left[\Delta^+(\{\alpha^{2,Y}, \beta^{2,Y}\}, \{\alpha^{1,x}, \beta^{1,x}\}; A_m^1)\right] \\ &= -E_Y\left[\lim_{x\to+\infty} \Delta^+(\{\alpha^{2,Y}, \beta^{2,Y}\}, \{\alpha^{1,x}, \beta^{1,x}\}; A_m^1)\right] = 0.\end{aligned}$$

□

Remark 3.4. In Proposition 3.3 condition (5) is required, and the beta-Stacy process in the discrete case is defined such that condition (1) is verified. Both conditions are satisfied when their ratio diverges, that is when

$$\lim_{j\to\infty} \prod_{i<j} \left(1 + \frac{1}{\beta_i^1}\right) \frac{\alpha_i^1 + \beta_i^1}{\alpha_i^1 + \beta_i^1 + 1} = +\infty.$$

This constraint does not pose restrictions, and it is satisfied, as expected, in the special cases of the homogeneous process and the Dirichlet process.

We finally state the following theorems, showing that there exist break-even points determining, respectively, a stay-with-a-winner and a switch-on-a-loser optimal strategy. The theorems generalize Theorem 2.1 and Theorem 2.2 of Chattopadhyay (1994), proving the existence of the break-even observations in a context more general than the Dirichlet arms, at the cost of some restrictions on the choice of the parameters of the beta-Stacy process.

Theorem 3.5. *For all $\{\alpha^1, \beta^1\}$ and $\{\alpha^2, \beta^2\}$ as in Proposition 3.1, for all nonincreasing discount sequences A_n and $n \geq 2$, if the condition*

$$\lim_{x \to 0} \Delta\left(\{\alpha^{1,x}, \beta^{1,x}\}, \{\alpha^2, \beta^2\}; A_n^1\right) \leq \Delta\left(\{\alpha^1, \beta^1\}, \{\alpha^2, \beta^2\}; A_n\right) \tag{7}$$

holds, there exists a break-even point $b\left(\{\alpha^1, \beta^1\}, \{\alpha^2, \beta^2\}; A_n\right)$ such that

$$\Delta\left(\{\alpha^{1,x}, \beta^{1,x}\}, \{\alpha^2, \beta^2\}; A_n^1\right) \geq \Delta\left(\{\alpha^1, \beta^1\}, \{\alpha^2, \beta^2\}; A_n\right)$$

if $x \geq b\left(\{\alpha^1, \beta^1\}, \{\alpha^2, \beta^2\}; A_n\right)$ and

$$\Delta\left(\{\alpha^{1,x}, \beta^{1,x}\}, \{\alpha^2, \beta^2\}; A_n^1\right) \leq \Delta\left(\{\alpha^1, \beta^1\}, \{\alpha^2, \beta^2\}; A_n\right)$$

if $x \leq b\left(\{\alpha^1, \beta^1\}, \{\alpha^2, \beta^2\}; A_n\right)$.

Proof. From Proposition 3.1, $\Delta\left(\{\alpha^{1,x}, \beta^{1,x}\}, \{\alpha^2, \beta^2\}; A_n^1\right)$ is non decreasing in x, starting from a value lower than and $\Delta\left(\{\alpha^1, \beta^1\}, \{\alpha^2, \beta^2\}; A_n\right)$ and going to infinity (Proposition 3.3). This is enough to claim that there exists a break-even point b which satisfies the properties in the theorem. □

Theorem 3.6. *For all $\{\alpha^1, \beta^1\}$ and $\{\alpha^2, \beta^2\}$ as in Proposition 3.1, for all nonincreasing discount sequences A_n and $n \geq 2$, if the condition*

$$\lim_{x \to 0} \Delta\left(\{\alpha^{1,x}, \beta^{1,x}\}, \{\alpha^2, \beta^2\}; A_n^1\right) \leq 0 \tag{8}$$

holds, there exists a break-even point $d\left(\{\alpha^1, \beta^1\}, \{\alpha^2, \beta^2\}; A_n\right)$ such that

$$\Delta\left(\{\alpha^{1,x}, \beta^{1,x}\}, \{\alpha^2, \beta^2\}; A_n^1\right) \geq 0 \quad \text{if} \quad x \geq d\left(\{\alpha^1, \beta^1\}, \{\alpha^2, \beta^2\}; A_n\right)$$

and

$$\Delta\left(\{\alpha^{1,x}, \beta^{1,x}\}, \{\alpha^2, \beta^2\}; A_n^1\right) \leq 0 \quad \text{if} \quad x \leq d\left(\{\alpha^1, \beta^1\}, \{\alpha^2, \beta^2\}; A_n\right).$$

Proof. As in the proof of Theorem 3.5, there exists a point d satisfying the properties. □

Remark 3.7. The domain of $\Delta\left(\{\alpha^{1,x}, \beta^{1,x}\}, \{\alpha^2, \beta^2\}; A_n\right)$, as a function of x, is $\mathbb{N}$, equipped with a discrete topology for which every subset is open, and all functions from a discrete topological space to any topological space are continuous. In the two theorems above, it is not possible to apply the intermediate value theorem for continuous functions since $\Delta\left(\{\alpha^{1,x}, \beta^{1,x}\}, \{\alpha^2, \beta^2\}; A_n\right)$, seen as a function of x, has a domain that is not a connected topological space. As a consequence it is not guaranteed that an observation exactly equal to the break-even can be observed since it could be that $b\left(\{\alpha^1, \beta^1\}, \{\alpha^2, \beta^2\}; A_n\right) \notin \mathbb{N}$ or $d\left(\{\alpha^1, \beta^1\}, \{\alpha^2, \beta^2\}; A_n\right) \notin \mathbb{N}$. Still, the break-evens give two reference points for optimal choices between the arms.

Remark 3.8. Conditions (7) and (8) are needed because the support of the base measure of the beta-Stacy process in limited to $(0, +\infty)$. Both Clayton and Berry (1985) and Chattopadhyay (1994) notice that when the support is bounded, the existence of break-even observations requires additional conditions at the boundaries. In particular, both conditions intuitively mean that if a very bad observation from arm 1 is extracted (x close to 0), the expected advantage of choosing that arm in the next time instant reduces and the alternative arm is preferred under the current strategy.

We now study the two-armed bandit problem when the base measures of the beta-Stacy processes are continuous measures.

4 Break-Even Observations: Continuous Case

In the present section we treat the continuous beta-Stacy two-armed problem. X_k and Y_k, respectively from arm 1 and arm 2 at stage k, can assume values in $\mathbb{R}^+ = (0, +\infty)$ and $\alpha(\cdot)$ and $\beta(\cdot)$ are, respectively, a continuous measure and a positive function, both defined on $\mathbb{R}^+$. $\alpha(\cdot)$ is assumed, without loss of generality, to have no discontinuity points.

Equation (8) in Walker and Muliere (1997), $t \in \mathbb{R}^+$, says that

$$\begin{aligned} \mathbb{P}(X \leq t) &= 1 - \exp\left\{ -\int_0^t \frac{d\alpha^1(s)}{\beta^1(s)} \right\} \\ &=: 1 - \prod_{[0,t]} \left(1 - \frac{d\alpha^1(s)}{\beta^1(s) + \alpha^1\{s\}} \right) \end{aligned}$$

where $\prod_{[0,t]}$ denotes the *product integral*, an operator commonly used in the survival analysis literature. For any partition $a_1 = z_0 < z_1 < \cdots < z_m = a_2$, if $l_m = \max_{i=1,\ldots,m} |x_i - x_{i-1}|$, the product integral is defined as

$$\prod_{[a_1,a_2]} \{1 + d\Gamma(z)\} := \lim_{l_m \to 0} \prod_j \{1 + \Gamma(z_j) - \Gamma(z_{j-1})\}.$$

See Gill and Johansen (1990) for a survey of applications of product integrals to survival analysis. Then, we can compute, in analogy with the discrete case,

$$E_X[X] = \int_0^{+\infty} \mathbb{P}(X > t) dt = \int_0^{+\infty} \prod_{[0,t]} \left(1 - \frac{d\alpha^1(s)}{\beta^1(s)} \right) dt =: \mu_1$$

and, similarly,

$$E_Y[Y] = \int_0^{+\infty} \prod_{[0,t]} \left(1 - \frac{d\alpha^2(s)}{\beta^2(s)} \right) dt =: \mu_2$$

assuming, without loss of generality, that $\mu_1 \leq \mu_2$.

The conditional expectation of a function $h_1(X)$, given observations from the first arm, $\underline{x} = (x_1, \ldots, x_k)$, is denoted $E[h(X)|\underline{x}]$, removing the explicit dependence on α^1 and β^1. We further define

$$\alpha^{1,\underline{x}}(s) = \alpha^1(s) + N\{s\} \text{ and } \beta^{1,\underline{x}}(s) = \beta^1(s) + M(s)$$

for all $s \in [0, \infty)$. Analogous notation is used for $h_2(Y)$ given $\underline{y} = (y_1, \ldots, y_k)$ from the second arm. Therefore, from Theorem 2.3,

$$\begin{aligned}\mathbb{P}(X \leq t|\underline{x}) &= 1 - \exp\left\{-\int_0^t \frac{d\alpha^1(s)}{\beta^1(s) + N\{s\} + M(s)}\right\} \\ &\quad \cdot \left(1 - \frac{N\{t\}}{\beta(t) + N\{t\} + M(t)}\right) \\ &=: 1 - \prod_{[0,t]} \left(1 - \frac{d\alpha^{1,\underline{x}}(s)}{\beta^{1,\underline{x}}(s) + \alpha^{1,\underline{x}}\{s\}}\right)\end{aligned}$$

and, partitioning $\underline{x} = [\underline{x}_{exact}, \underline{x}_{cens}]$ for respectively exact and censored observations, the posterior mean is

$$\begin{aligned}E_X[X|\underline{x}] &= \mathbb{P}(X \notin \underline{x}_{exact}|\underline{x}) \cdot \int_0^{+\infty} \mathbb{P}(X > t|X \notin \underline{x}_{exact}, \underline{x})dt + \\ &\quad \mathbb{P}(X \in \underline{x}_{exact}|\underline{x}) \cdot \sum_{x \in \underline{x}_{exact}} x\mathbb{P}(X = x|X \in \underline{x}_{exact}, \underline{x}) \\ &= \mathbb{P}(X \notin \underline{x}_{exact}|\underline{x}) \int_0^{+\infty} \prod_{[0,t]} \left(1 - \frac{d\alpha^{1,\underline{x}_{cens}}(s)}{\beta^{1,\underline{x}_{cens}}(s) + \alpha^{1,\underline{x}_{cens}}\{s\}}\right) dt + \\ &\quad \mathbb{P}(X \in \underline{x}_{exact}|\underline{x}) \sum_{x \in \underline{x}_{exact}} x\mathbb{P}(X = x|X \in \underline{x}_{exact}, \underline{x}) =: \mu_{1,\underline{x}}\end{aligned}$$

The function $\Delta(\{\alpha^1, \beta^1\}, \{\alpha^2, \beta^2\}; A_n)$ can be expressed as in (2). In the following propositions we study the properties of Δ, with the aim of proving the existence of break-even observations determining optimal stay-with-a-winner and switch-on-a-loser strategies.

Proposition 4.1. *For all $\{\alpha^1, \beta^1\}$ and $\{\alpha^2, \beta^2\}$ such that $-\frac{\partial}{\partial x}\beta^1(x) \geq \alpha^1(x)$ and for all non-increasing discount sequences A_n, $\Delta\left(\{\alpha^{1,x}, \beta^{1,x}\}, \{\alpha^2, \beta^2\}; A_n\right)$ is nondecreasing in x, for all $x \in [0, \infty)$.*

Proof. By induction, for $n = 1$, and x censored to the right,

$$\begin{aligned}\Delta\left(\{\alpha^{1,x}, \beta^{1,x}\}, \{\alpha^2, \beta^2\}; A_1\right) &= a_1(\mu_{1,x} - \mu_2) \\ &= a_1\left(\int_0^{+\infty} \mathbb{P}(X > t|x)dt - \mu_2\right) \\ &= a_1\left(\int_0^{+\infty} \prod_{[0,t]} \left(1 - \frac{d\alpha^1(s) + dN(s)}{\beta^1(s) + N\{s\} + M(s)}\right) dt - \mu_2\right)\end{aligned}$$

where $N(s) = N\{[0, s]\}$. We first show that $\mu_{1,x}$ is nondecreasing in x, for x being censored to the

right. Notice that $\mu_{1,x}$ can be written as

$$\begin{aligned}\mu_{1,x} &= \int_0^{+\infty} \exp\left\{-\int_0^t \frac{d\alpha^1(s)}{\beta^1(s)+N\{s\}+M(s)}\right\} \\ &\quad \left(1-\frac{N\{t\}}{\beta(t)+N\{t\}+M(t)}\right)dt \\ &= \int_0^x \exp\left\{-\int_0^t \frac{d\alpha^1(s)}{\beta^1(s)+1}\right\}dt \\ &\quad + \int_x^{+\infty} \exp\left\{-\left(\int_0^x \frac{d\alpha^1(s)}{\beta^1(s)+1}+\int_x^t \frac{d\alpha^1(s)}{\beta^1(s)}\right)\right\}dt.\end{aligned}$$

The integrand in $\mu_{1,x}$ as a function of t has a discontinuity point when $t=x$, but its value at this point is ignored since it does not contribute to the evaluation of $\mu_{1,x}$. Take now any $x^*>x$, and separate the cases $t\le x$, $t\in(x,x^*)$ and $t\ge x^*$:

- When $t\le x$, the integrands in $\mu_{1,x}$ and μ_{1,x^*} are the same.
- When $t\ge x^*$, the integrand in $\mu_{1,x}$ is

  $$\exp\left\{-\left(\int_0^x \frac{d\alpha^1(s)}{\beta^1(s)+1}+\int_x^t \frac{d\alpha^1(s)}{\beta^1(s)}\right)\right\},$$

  whilst the integral in μ_{1,x^*} is

  $$\exp\left\{-\left(\int_0^{x^*} \frac{d\alpha^1(s)}{\beta^1(s)+1}+\int_{x^*}^t \frac{d\alpha^1(s)}{\beta^1(s)}\right)\right\},$$

  with the integrand in μ_{1,x^*} always greater than or equal to the one in $\mu_{1,x}$.
- Finally, when $t\in(x,x^*)$, the integrands in $\mu_{1,x}$ and μ_{1,x^*} are, respectively,

  $$\exp\left\{-\left(\int_0^x \frac{d\alpha^1(s)}{\beta^1(s)+1}+\int_x^t \frac{d\alpha^1(s)}{\beta^1(s)}\right)\right\}$$

  and

  $$\exp\left\{-\int_0^t \frac{d\alpha^1(s)}{\beta^1(s)+1}\right\},$$

proving that $\mu_{1,x^*}\ge\mu_{1,x}$ and that the statement is true for $n=1$. On the other hand, when x is not censored

$$\begin{aligned}\mu_{1,x} &= \mathbb{P}(X\neq x|x)\mu_1+\mathbb{P}(X=x|x)x \\ &= \left(1-\exp\left\{-\int_0^x \frac{d\alpha^1(s)}{\beta^1(s)+1}\right\}\frac{1}{\beta^1(x)+1}\right)\mu_1+\exp\left\{-\int_0^x \frac{d\alpha^1(s)}{\beta^1(s)+1}\right\}\frac{x}{\beta^1(x)+1},\end{aligned}$$

and $\mu_{1,x*}\ge\mu_{1,x}$ for all $x*>x$, if and only if $\mathbb{P}(X=x|x)$, the probability of X from arm 1 being equal to the previous exact observation, is nondecreasing in x. This condition is equivalent to $-\frac{\partial}{\partial x}\beta^1(x)\ge\alpha^1(x)$, as required in the proposition.

By induction, assuming the monotonicity property for $n=m-1$, the proof is completed along the lines of Proposition 3.1. □

Remark 4.2. As in the discrete case, monotonicity of the posterior mean is recovered under a condition on the parameters of the beta-Stacy process. The condition $\beta_j^1 \leq \beta_{j+1}^1 + \alpha_{j+1}^1$, $j \in \mathbb{N}$, in Proposition 3.1 finds its continuous analogue $-\frac{\partial}{\partial x}\beta^1(x) \geq \alpha^1(x)$, $x \in \mathbb{R}^+$, in Proposition 4.1. It is important to note that the condition is not required if only censored observations are extracted from the arms, but it is necessary in case of exact observations. As in the discrete section, the special cases of Dirichlet and homogeneous processes are included, and they correspond, respectively, to $-\frac{\partial}{\partial x}\beta^1(x) = \alpha^1(x)$ and to $\frac{\partial}{\partial x}\beta^1(x) = 0$.

Proposition 4.3. *For all $\{\alpha^1, \beta^1\}$ and $\{\alpha^2, \beta^2\}$ as in Proposition 4.1 and such that $\int_0^\infty d\alpha^1(s)/(\beta^1(s)+1) < \infty$, for all $x \in \mathbb{R}^+$ and all nonincreasing discount sequences A_n,*

$$\lim_{x\to+\infty} \Delta\left(\{\alpha^{1,x}, \beta^{1,x}\}, \{\alpha^2, \beta^2\}; A_n\right) = \infty$$

and

$$\lim_{x\to 0} \Delta\left(\{\alpha^{1,x}, \beta^{1,x}\}, \{\alpha^2, \beta^2\}; A_n\right) = \min_x \Delta\left(\{\alpha^{1,x}, \beta^{1,x}\}, \{\alpha^2, \beta^2\}; A_n\right).$$

Proof. The case $x \to 0$ is an immediate consequence of Proposition 4.1. To study the case where x diverges, we proceed by induction. Note that for $n = 1$, $\Delta\left(\{\alpha^{1,x}, \beta^{1,x}\}, \{\alpha^2, \beta^2\}; A_1\right) = a_1(\mu_{1,x} - \mu_2)$ and

$$\begin{aligned}\lim_{x\to+\infty} \mu_{1,x} &= \int_0^{+\infty} \exp\left\{-\int_0^t \frac{d\alpha^1(s)}{\beta^1(s)+1}\right\} dt \\ &\geq \exp\left\{-\int_0^{+\infty} \frac{d\alpha^1(s)}{\beta^1(s)+1}\right\} \int_0^{+\infty} 1dt = +\infty,\end{aligned}$$

where the last equality is true since $\int_0^\infty d\alpha^1(s)/(\beta^1(s)+1) < \infty$ is equivalent to

$$\exp\left\{-\int_0^{+\infty} \frac{d\alpha^1(s)}{\beta^1(s)+1}\right\} > 0.$$

This proves that $\lim_{x\to+\infty} \Delta\left(\{\alpha^{1,x}, \beta^{1,x}\}, \{\alpha^2, \beta^2\}; A_1\right) = \infty$. The rest of the proof follows the same lines as the proof of Proposition 3.3. □

Remark 4.4. In the above proposition the additional condition $\int_0^\infty d\alpha^1(s)/(\beta^1(s)+1) < \infty$ is required. Note that the beta-Stacy process is defined such that $\int_0^\infty d\alpha^1(s)/\beta^1(s) = \infty$. These two improper integrals should have a different asymptotic behavior, a condition that is verified when, from the limit comparison test for integrals, the limit of the ratio of the two integrands is different from 1, that is when

$$\lim_{s\to\infty} \left(1 + \frac{1}{\beta^1(s)}\right) \neq 1.$$

For finite $\beta^1(s)$, this is satisfied, and, as expected, includes the special cases of the homogeneous process and the Dirichlet process. In short, the additional constraint rules out cases of exploding $\beta^1(s)$. Usually, for some base distribution F_0, $\beta^1(s) = M \cdot F_0[s, \infty)$, converging to 0 (see Walker and Muliere 1997).

Proposition 4.5. *For all $\{\alpha^1, \beta^1\}$ and $\{\alpha^2, \beta^2\}$ and all nonincreasing discount sequences A_n, $\Delta\left(\{\alpha^{1,x}, \beta^{1,x}\}, \{\alpha^2, \beta^2\}; A_n\right)$ is a continuous function of x, for $x \in [0, \infty)$ censored to the right.*

Proof. It is enough to show that, for any increasing or decreasing sequence $\{x\}$ converging to x_0,

$$\Delta\left(\{\alpha^{1,x},\beta^{1,x}\},\{\alpha^2,\beta^2\};A_n\right)\to\Delta\left(\{\alpha^{1,x_0},\beta^{1,x_0}\},\{\alpha^2,\beta^2\};A_n\right).$$

We provide the proof only for an increasing sequence $\{x\}$, since the decreasing sequence case is similar.

By induction, first fix $n=1$, so that $\Delta\left(\{\alpha^{1,x},\beta^{1,x}\},\{\alpha^2,\beta^2\};A_1\right)=a_1(\mu_{1,x}-\mu_2)$. The continuity of Δ in x is shown through the continuity of $\mu_{1,x}$. Taking any increasing sequence $\{x\}$ converging to x_0, then

$$\begin{aligned}\lim_{x\to x_0}\mu_{1,x} &= \lim_{x\to x_0}\left(\int_0^x\exp\left\{-\int_0^t\frac{d\alpha^1(s)}{\beta^1(s)+1}\right\}dt\right.\\ &+ \left.\int_x^{+\infty}\exp\left\{-\left(\int_0^x\frac{d\alpha^1(s)}{\beta^1(s)+1}+\int_x^t\frac{d\alpha^1(s)}{\beta^1(s)}\right)\right\}dt\right)\\ &= \int_0^{x_0}\exp\left\{-\int_0^t\frac{d\alpha^1(s)}{\beta^1(s)+1}\right\}dt\\ &+ \exp\left\{-\int_0^{x_0}\frac{d\alpha^1(s)}{\beta^1(s)+1}\right\}\cdot\lim_{x\to x_0}\int_x^{+\infty}\exp\left\{-\int_x^t\frac{d\alpha^1(s)}{\beta^1(s)}\right\}dt,\end{aligned}$$

where the last equality is justified by the continuity in x of

$$\int_0^x\exp\left\{-\int_0^t\frac{d\alpha^1(s)}{\beta^1(s)+1}\right\}dt\qquad\text{and}\qquad\exp\left\{-\int_0^x\frac{d\alpha^1(s)}{\beta^1(s)+1}\right\}.$$

To finally see that $\mu_{1,x}$ is continuous, we need to prove the continuity in x of the function

$$H(x):=\int_x^{+\infty}\exp\left\{-\int_x^t\frac{d\alpha^1(s)}{\beta^1(s)}\right\}dt.$$

Note that the function H is a parameterized Riemann integral, whose integration extremes are also dependent on the parameter. H is given by the composition of two functions:

$$H_2(h,x):=\int_h^{+\infty}\exp\left\{-\int_x^t\frac{d\alpha^1(s)}{\beta^1(s)}\right\}dt$$

and $h(x)=x$. The latter is obviously continuous. For the continuity of H_2, note that

$$\left|\exp\left\{-\int_x^t\frac{d\alpha^1(s)}{\beta^1(s)}\right\}\right|\le 1,$$

and we can apply the dominated convergence theorem to the sequence of functions in x

$$\exp\left\{-\int_x^t\frac{d\alpha^1(s)}{\beta^1(s)}\right\}$$

for any given value of $h\in(0,\infty)$. Then

$$\begin{aligned}\lim_{x\to x_0}H_2(h,x) &= \lim_{x\to x_0}\int_h^{+\infty}\exp\left\{-\int_x^t\frac{d\alpha^1(s)}{\beta^1(s)}\right\}dt\\ &= \int_h^{+\infty}\exp\left\{-\int_{x_0}^t\frac{d\alpha^1(s)}{\beta^1(s)}\right\}dt=H_2(h,x_0).\end{aligned}$$

Assume now that the statement is true for $n = m - 1$. By (4),

$$\begin{aligned}\Delta(\{\alpha^{1,x},\beta^{1,x}\},\{\alpha^2,\beta^2\};A_m) &= (a_1-a_2)(\mu_{1,x}-\mu_2)\\ &+E_X\left[\Delta^+(\{\alpha^{1,(x,X)},\beta^{1,(x,X)}\},\{\alpha^2,\beta^2\};A^1_m)\right]\\ &+E_Y\left[\Delta^-(\{\alpha^{1,x},\beta^{1,x}\},\{\alpha^{2,Y},\beta^{2,Y}\};A^1_m)\right].\end{aligned}$$

The first term $(a_1-a_2)(\mu_{1,x}-\mu_2)$ on the right hand side of the formula is continuous in x (this follows from the continuity of $\mu_{1,x}$). For the second term, note that $\Delta^+(\{\alpha^{1,(x,X)},\beta^{1,(x,X)}\},\{\alpha^2,\beta^2\};A^1_m)$ is a nondecreasing sequence in x (by Proposition 4.1), bounded below by 0 (by its definition) and convergent to $\Delta^+(\{\alpha^{1,(x_0,X)},\beta^{1,(x_0,X)}\},\{\alpha^2,\beta^2\};A^1_m)$ (by the induction hypothesis). We can then apply the monotone convergence theorem:

$$\begin{aligned}&\lim_{x\to x_0} E_X\left[\Delta^+(\{\alpha^{1,(x,X)},\beta^{1,(x,X)}\},\{\alpha^2,\beta^2\};A^1_m)\right]\\ &= E_X\left[\lim_{x\to x_0}\Delta^+(\{\alpha^{1,(x_0,X)},\beta^{1,(x_0,X)}\},\{\alpha^2,\beta^2\};A^1_m)\right].\end{aligned}$$

For the third term, notice that

$$\Delta(\{\alpha^{1,x},\beta^{1,x}\},\{\alpha^{2,y},\beta^{2,y}\};A^1_m) = -\Delta(\{\alpha^{2,y},\beta^{2,y}\},\{\alpha^{1,x},\beta^{1,x}\};A^1_m).$$

Furthermore, $\Delta^+(\{\alpha^{2,y},\beta^{2,y}\},\{\alpha^{1,x},\beta^{1,x}\};A^1_m)$ converges to

$$\Delta^+(\{\alpha^{2,y},\beta^{2,y}\},\{\alpha^{1,x_0},\beta^{1,x_0}\};A^1_m)$$

as x converges (by the induction hypothesis), and it is bounded above by

$$\left|\Delta^+(\{\alpha^{2,y},\beta^{2,y}\},\{\alpha^{1,x=0},\beta^{1,x=0}\};A^1_m)\right|.$$

By the dominated convergence theorem,

$$\begin{aligned}&\lim_{x\to x_0} E_Y\left[\Delta^-(\{\alpha^{1,x},\beta^{1,x}\},\{\alpha^{2,Y},\beta^{2,Y}\};A^1_m)\right]\\ &= -\lim_{x\to x_0} E_Y\left[\Delta^+(\{\alpha^{2,Y},\beta^{2,Y}\},\{\alpha^{1,x},\beta^{1,x}\};A^1_m)\right]\\ &= -E_Y\left[\lim_{x\to x_0}\Delta^+(\{\alpha^{2,Y},\beta^{2,Y}\},\{\alpha^{1,x},\beta^{1,x}\};A^1_m)\right]\\ &= E_Y\left[\Delta^-(\{\alpha^{1,x_0},\beta^{1,x_0}\},\{\alpha^{2,Y},\beta^{2,Y}\};A^1_m)\right],\end{aligned}$$

proving continuity for the generic bandit horizon n. □

Theorem 4.6. *For all $\{\alpha^1,\beta^1\}$ and $\{\alpha^2,\beta^2\}$ as in Proposition 4.3, for all nonincreasing discount sequences A_n and $n \geq 2$, if condition (7) holds, there exists a break-even point $b\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_n\right)$ such that*

$$\Delta\left(\{\alpha^{1,x},\beta^{1,x}\},\{\alpha^2,\beta^2\};A^1_n\right) \geq \Delta\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_n\right)$$

if $x \geq b\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_n\right)$ and

$$\Delta\left(\{\alpha^{1,x},\beta^{1,x}\},\{\alpha^2,\beta^2\};A^1_n\right) \leq \Delta\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_n\right)$$

if $x \leq b\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_n\right)$.

Proof. From Propositions 4.1 and 4.5, $\Delta\left(\{\alpha^{1,x},\beta^{1,x}\},\{\alpha^2,\beta^2\};A_n^1\right)$ is nondecreasing in x and continuous (the latter only for x censored), starting from a value lower than

$$\Delta\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_n\right)$$

and growing to infinity (Proposition 4.3). Then there exists a point $b\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_n\right)$ which satisfies the properties of the theorem. □

Theorem 4.7. *For all $\{\alpha^1,\beta^1\}$ and $\{\alpha^2,\beta^2\}$ as in Proposition 4.3, for all nonincreasing discount sequences A_n and $n\geq 2$, if condition (8) holds, there exists a break-even point $d\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_n\right)$ such that*

$$\Delta\left(\{\alpha^{1,x},\beta^{1,x}\},\{\alpha^2,\beta^2\};A_n^1\right)\geq 0 \quad \text{if } x\geq d\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_n\right)$$

and

$$\Delta\left(\{\alpha^{1,x},\beta^{1,x}\},\{\alpha^2,\beta^2\};A_n^1\right)\leq 0 \quad \text{if } x\leq d\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_n\right).$$

Proof. As in Theorem 4.6, there exists a point d satisfying the properties. □

Remark 4.8. For Theorems 4.6 and 4.7, Remark 3.8 on the boundary conditions is still valid. Remark 3.7 can be applied if there are exact observations, whilst the intermediate value theorem for continuous functions guarantees that an observation exactly equal to the break-even can be observed in case of censored data.

5 Applications

5.1 Geometric beta-Stacy parameters

Fix, for all $j\in\mathbb{N}$, $\beta_j^1=M_1\cdot 0.9^{j-1}$ and $\alpha_j^1=0.1M_1\cdot 0.9^{j-1}$ for the first arm, and $\beta_j^2=M_2\cdot 0.92^{j-1}$ and $\alpha_j^2=0.08M_2\cdot 0.92^{j-1}$ for the second arm, for $j=1,2,\ldots$, with $\mu_1<\mu_2$ a priori. Let $M_i=\alpha^i(\mathbb{N})$, $i=1,2$ be the total masses of the measures α^1 and α^2. Note that different values of M_1 and M_2 do not affect prior means, μ_1 and μ_2, but only posterior means. We observe censored to the right observations, fix the bandit horizon to $n=3$ and the discount sequence is $A_3=(1,0.9,0.8)$. Choosing higher values for n is feasible and to higher values correspond higher processing times. We generate $X_1^{(l)}$ and $Y_1^{(l)}$ from the two arms, for $l=1,\ldots,T=100$. Then for each $X_1^{(l)}$ we generate T copies of X_2 from the first arm, and for each $Y_1^{(l)}$ we generate T copies of Y_2 from the second arm. Sampling from prior and posterior beta-Stacy processes is done, respectively, with Algorithm A and B in Al Labadi and Zarepour (2013). See also De Blasi (2007) for an alternative way of simulating from the beta-Stacy process. For each scenario, we evaluate μ_{1,x_1}, $\mu_{1,(x_1,x_2)}$, μ_{2,y_1} and $\mu_{2,(y_1,y_2)}$; we then evaluate $\Delta\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_3\right)$, reported in Table 1 for different values of M_1 and M_2. Holding everything else constant, there is a tendency for $\Delta\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_3\right)$ to increase in M_2 and decrease in M_1. This result is coherent with the exploitation-exploration trade-off mentioned in the Introduction, and suggests that the less is known about the arm, the more appealing is to select the arm, since higher information can be gained from its exploration. Furthermore, as both M_1 and M_2 increase, $\Delta\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_3\right)$ approaches $\mu_1-\mu_2=-2.5$. When $\Delta\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_3\right)$ is positive, the optimal arm is the first one, and viceversa when is negative. Most of the times, the difference in the prior means makes the second arm the optimal one, except in cases with $M_1<<M_2$: there are situations where the higher uncertainty (lower M_1) around the base distribution of the first arm, relative to the high confidence in the base distribution of the second arm (larger M_2), makes the first arm preferable to be explored.

Table 1: Estimated $\Delta\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_3\right)$, with $\beta_j^1 = M_1 \cdot 0.9^{j-1}$ and $\alpha_j^1 = 0.1M_1 \cdot 0.9^{j-1}$ for the first arm, and $\beta_j^2 = M_2 \cdot 0.92^{j-1}$ and $\alpha_j^2 = 0.08M_2 \cdot 0.92^{j-1}$ for the second arm, for $j = 1, 2, \ldots$ and for different values of $M_i = \alpha^i(\mathbb{N})$, $i = 1, 2$.

	M_2				
M_1	0.1	1	5	10	100
0.1	-11.0406	0.4198	10.4029	12.1847	11.9691
1	-19.6250	-9.4420	1.2262	3.0780	5.5041
5	-21.7856	-13.2648	-5.3174	-3.6489	-1.8068
10	-21.8101	-13.6430	-6.1519	-4.3691	-2.6009
100	-22.1149	-13.2984	-6.0458	-4.5251	-2.7353

5.2 Discrete uniform beta-Stacy parameters

Consider the beta-Stacy two-armed bandit problem with, for $k = 1, 2$,

$$\alpha_i^k = M_k \frac{1}{2h_k+1} \mathbb{1}_{i\in\{\mu_k-h_k,\ldots,\mu_k+h_k\}}, \tag{9}$$

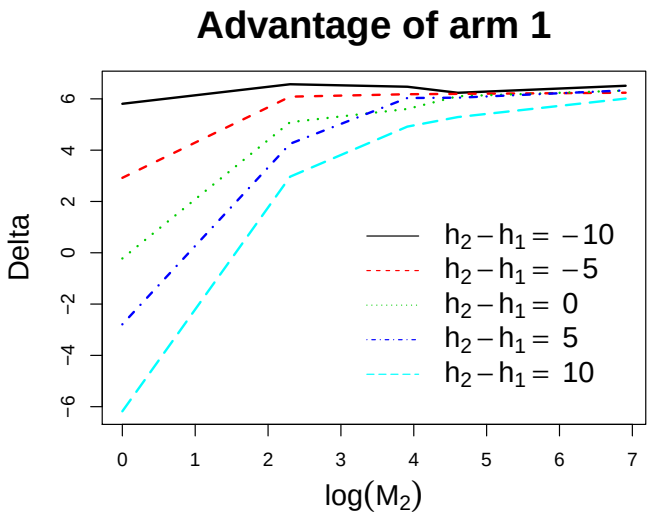
$$\beta_i^k = M_k \left(\frac{h_k+\mu_k-i}{2h_k+1} \mathbb{1}_{i\in\{\mu_k-h_k,\ldots,\mu_k+h_k\}} + \mathbb{1}_{i<\mu_k-h_k} \right), \tag{10}$$

where $h_k \in \mathbb{N}$ and $h_k < \mu_k$, and with $M_k = \alpha^k(\mathbb{N})$. The parameter h_k is positively related to the variability of the base distribution of the beta-Stacy process of arm k, whilst M_k is negatively related to the variability around the base measure. We fix $\mu_1 = \mu_2 = 20$, and $M_1 = h_1 = 1$, to see how the expected advantage of one arm over the other is affected by a change in h_2 and in M_2, without being affected by different prior means. With μ_1 and μ_2 equal, the choice od the optimal arm is entirely driven by the chance of exploring less-known (more volatile) arms. The rest of the setting is fixed as in the previous example. In Figure 1 we report the value of $\Delta\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_3\right)$ for different h_2 and M_2. The green dotted line, corresponding to the case $h_1 = h_2 = 1$, shows how a lower variability (higher M_2) around the base measure of arm 2, makes this arm less interesting to explore, in favor of arm 1. The same effect is caused by a change in the variability of the base measure of arm 2: for $h_2 - h_1 < 0$ and for $M_1 = M_2 = 1$, arm 1 is preferred, up to a $\Delta\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_3\right) \approx 6$ for $h_2 - h_1 = -10$. Viceversa, higher positive values of $h_2 - h_1$ correspond to higher preference for arm 2. Furthermore, the effect of a change in M_2 seems to dominate: as we increase M_2, the distances among the scenarios with different h_2 decrease and concentrate on positive expected advantages of the first arm over the second one.

5.3 Exponential beta-Stacy parameters

We extend to the two-armed bandit problem a numerical example in Ferguson and Phadia (1979) and Walker and Muliere (1997). For the first arm fix $\beta^1(s) = exp(-s/10)$ and $d\alpha^1(s) = exp(-s/10)/10ds$, whist for the second arm $\beta^2(s) = exp(-s/12)$ and $d\alpha^2(s) = exp(-s/12)/12ds$, for $s \in \mathbb{R}^+$, so that $\mu_1 < \mu_2$ a priori. We fix the bandit horizon to $n = 3$ and the discount sequence $A_3 = (1, 0.9, 0.8)$. As in the discrete example, we generate $X_1^{(l)}$ and $Y_1^{(l)}$ from the two arms, for $l = 1, \ldots, T = 150$. Then for each $X_1^{(l)}$ we generate T copies of X_2 from the first arm, and for each $Y_1^{(l)}$ we gener-

Figure 1: *Estimated* $\Delta\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_3\right)$*, with parameters as specified in equations* (9) *and* (10)*, for different variability around the base measure* (M_2) *and different base measure variability* (h) *of the beta-Stacy process from the second arm. The prior means are both equal to* $\mu_1=\mu_2=20$*, and* $M_1=h_1=1$*.*



ate T copies of Y_2 from the second arm, also using Algorithm A and B in Al Labadi and Zarepour (2013). In the top-left plot of Figure 2, two randomly extracted prior distributions for the two arms are reported. For each scenario, we evaluate μ_{1,x_1}, $\mu_{1,(x_1,x_2)}$, μ_{2,y_1} and $\mu_{2,(y_1,y_2)}$; we then evaluate $\Delta\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_3\right)$ and $\Delta\left(\{\alpha^{1,x},\beta^{1,x}\},\{\alpha^2,\beta^2\};A_3^1\right)$, for $x\in[0,\infty)$. In the top-right plot of Figure 2 these Δs are shown, when the observation in the first and second periods are respectively exact and censored. Monotonicity in x of $\Delta\left(\{\alpha^{1,x},\beta^{1,x}\},\{\alpha^2,\beta^2\};A_3^1\right)$ is numerically verified. Note also that conditions (7) and (8) are satisfied, so that the two break-even points exist. Since $\Delta\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_3\right)=-2.47<0$, $b\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_3\right)\leq d\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_3\right)$. In particular, the break-even observation for the stay-with-a-winner strategy is $b\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_3\right)=15.42$, whilst the break-even for the switch-on-a-loser strategy is $d\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_3\right)=19.37$. Optimal strategies can be completely determined. For instance, arm 2 is optimally selected at the beginning since $\Delta\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_3\right)<0$. If an exact observation from arm 2 is extracted equal, say, to $y_1=4$, for the stay-with-a-winner strategy arm 1 is optimal at this stage since

$$\Delta\left(\{\alpha^1,\beta^1\},\{\alpha^{2,y_1=4},\beta^{2,y_1=4}\};A_3^1\right)=-1.60,$$

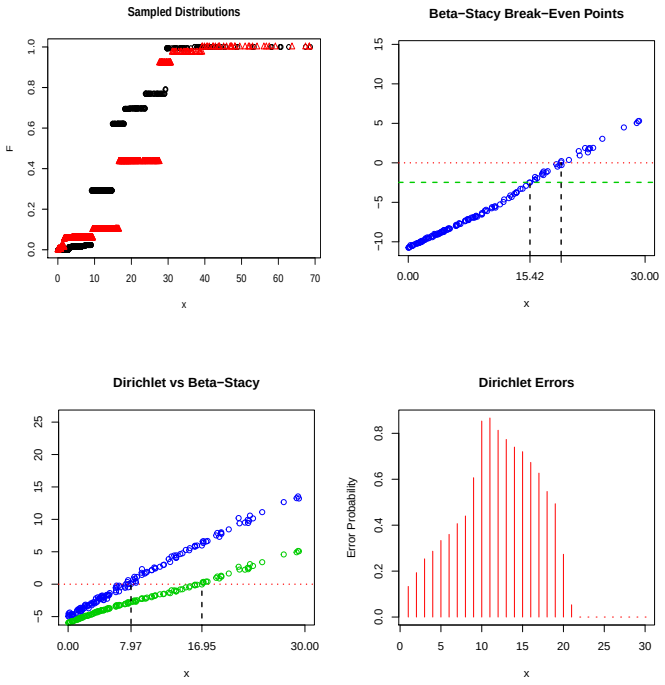
greater than $\Delta\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};A_3\right)=-2.47$. At this stage arm 1 would not be optimal for the switch-on-a-loser strategy, since $\Delta\left(\{\alpha^1,\beta^1\},\{\alpha^{2,y=4},\beta^{2,y=4}\};A_3^1\right)<0$. Suppose now a censored observation from arm 1 equal, say, to $x_2=15.5$ is observed, for which

$$\Delta\left(\{\alpha^{1,x_2=15.5},\beta^{1,x_2=15.5}\},\{\alpha^{2,y_1=4},\beta^{2,y_1=4}\};A_3^2\right)=-1,$$

greater than $\Delta\left(\{\alpha^1,\beta^1\},\{\alpha^{2,y_1=4},\beta^{2,y_1=4}\};A_3^1\right)$. Therefore in the last stage arm 1 is again optimal for the stay-with-a-winner strategy. Furthermore, in the bottom-left plot we report

$$\Delta\left(\{\alpha^{1,x},\beta^{1,x}\},\{\alpha^2,\beta^2\};A_3^1\right)$$

Figure 2: *Top-left: Two distributions sampled from the beta-Stacy process of Walker and Muliere (1997), with algorithm A in Al Labadi and Zarepour (2013). The black line is from arm 1, the red one from arm 2, with parameters* $\beta^1(s) = exp(-s/10)$, $d\alpha^1(s) = exp(-s/10)/10ds$, $\beta^2(s) = exp(-s/12)$ *and* $d\alpha^2(s) = exp(-s/12)/12ds$, *for* $s \in [0, \infty)$. *Top-right: in blue* $\Delta\left(\{\alpha^{1,x}, \beta^{1,x}\}, \{\alpha^2, \beta^2\}; A_3^1\right)$, *in green* $\Delta\left(\{\alpha^1, \beta^1\}, \{\alpha^2, \beta^2\}; A_3\right)$. *The 0 value is also highlighted in red. The intersections determine the break-even observations of the two strategies outlined in the text. Bottom-left: in blue* $\Delta\left(\{\alpha^{1,x}, \beta^{1,x}\}, \{\alpha^2, \beta^2\}; A_3^1\right)$ *for the beta-Stacy bandit problem. The green line represents the corresponding quantity for the Dirichlet bandit that ignores the censorship. Bottom-right: Error probability of the stay-with-the winner strategy implied by the Dirichlet bandit problem, when censorship is neglected, as function of the first observation from arm 1.*



when $x \in [0, \infty)$ is right-censored, and the expected advantage of arm 1 if the data were incorrectly supposed to be exact. In other words, we compare the beta-Stacy bandit problem with the corresponding Dirichlet bandit problem that ignores the censorship, to quantify the difference between the two in the simulation setting and highlight the relevance of properly accounting for right-censored data. There is a range of values from 7.97 to 16.95 at which the Dirichlet bandit problem would take the wrong strategy, since $\Delta\left(\{\alpha^{1,x}, \beta^{1,x}\}, \{\alpha^2, \beta^2\}; A_3^1\right)$ would be of opposite sign, relative to the corresponding beta-Stacy quantity. The break-even for the Dirichlet bandit is too low, since it judges the observations to be exact and therefore does not account for the increased chance of observing higher values in the future. If we repeat the experiment 150 times for each value of x from 1 to 30, we can compute the probability for the Dirichlet bandit being in error in the choice of the optimal arm, after the observation of x from arm 1. The probability is reported in the bottom-right plot of Figure 2.

6 Conclusions and further directions

We have studied Bayesian nonparametric bandit problems with right-censored data, where two independent arms are generated by beta-Stacy processes. The problem extends the one-armed and two-armed Dirichlet bandit problem of Clayton and Berry (1985) and Chattopadhyay (1994) since the beta-Stacy process simplifies to the Dirichlet process for a special choice of the process parameters and in the absence of censored observations. We have shown some properties of the expected advantage of the first arm over the second arm, and the existence of stay-with-a-winner and switch-on-a-loser optimal strategies, under non-restrictive constraints on the process parameters.

Our framework can be further extended in several directions to different bandit strategies and multi-armed problems. First, the common formulation of the Bernoulli bandit can be replicated through the choice of Bernoulli base measures, centered on success probabilities that are learnt as observations are collected. Second, semi-uniform strategies with greedy behaviour can be addressed: epsilon-greedy and epsilon-first strategies (Watkins 1989; Sutton and Barto 1998) that dedicate a proportion of phases to, respectively, random and purely exploratory phases, can be derived by randomizing the reinforcement learning mechanism of the arms' parameters (Muliere et al. 2006); epsilon-decreasing and VBDE strategies (Cesa-Bianchi and Fisher 1998; Tokic 2010) would require a beta-Stacy parameter update mechanism depedent on the number of steps or on the values extracted from the arms. Third, probability matching strategies such as Thompson sampling (Thompson 1933) are easy to implement in our framework: in multi-armed bandit problems, beta-Stacy random distributions can be sampled from their posterior distributions at each stage, and, conditional to the sample, some reward related to the single arm may be computed and all rewards compared. Fourth, the extension to multi-armed contextual bandits (Langford and Zhang 2008) can be implemented by introducing dependence of the arm parameters on external regressors, or introducing dependence between Bayesian nonparametric arms through partial exchangeability (de Finetti 1938, 1959), for instance with the mixture of Dirichlet processes of Antoniak (1974), the Bivariate Dirichlet process of Walker and Muliere (2003) or the Bivariate beta-Stacy process of Muliere et al. (2007). In this direction, Battiston et al. (2016) adopt hierarchical Poisson-Dirichlet processes in multi-armed bandit problems. Fifth, the sequential nature of the Bayesian framework and the flexibility of nonparametric priors permit to handle more general cases of non-stationary bandit problems (Garivier and Moulines 2008), where the underlying base distribution of the beta-Stacy processes can change during play: in this context all past observations simply affect the modified priors of the new models. Finally, it could be of interest to apply the proposed framework to study the results in Gittins (1979) and Gittins et al. (2011) in multi-armed Bayesian nonparametric problems: in this case, Dynamic Allocation Indices may be computed from the simulations of the stochastic processes related to each arm.

Acknoledgements

We thank two anonymous referees, the Editor and the Associate Editor, for their detailed and insightful comments that contributed significanty to improve the paper.

References

Al Labadi, L. and Zarepour, M. (2013). A Bayesian nonparametric goodness of fit test for right censored data based on approximate samples from the beta-Stacy process. *The Canadian Journal of Statistics*, 41:466–487.

Antoniak, C. (1974). Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *Annals of Statistics*, 2:1152–1174.

Battiston, M., Favaro, S., and Teh, Y. (2016). Multi-armed bandit for species discovery: a Bayesian nonparametric approach. *Journal of the American Statistical Association*. Forthcoming.

Bellman, R. (1956). A Problem in the Sequential Design of Experiments. *Sankhya*, 16:221–229.

Berry, D. (1972). A Bernoulli Two-Armed Bandit. *The Annals of Mathematical Statistics*, 43:871–897.

Berry, D. and Fristedt, B. (1979). Bernoulli One-Armed Bandits - Arbitrary Discount Sequences. *The Annals of Statistics*, 7:1086–1105.

Berry, D. and Fristedt, B. (1985). *Bandit Problems: Sequential Allocation of Experiments*. Chapman and Hall, New York.

Bradt, R., Johnson, S., and Karlin, S. (1956). On Sequential Designs for Maximizing the Sum of n Observations. *The Annals of Mathematical Statistics*, 33:847–856.

Cesa-Bianchi, N. and Fisher, P. (1998). Finite-time regret bounds for the multiarmed bandit problem. In *Proceedings of the 15th International Conference on Machine Learning*, pages 100–108, Morgan Kaufmann, San Francisco, CA.

Chattopadhyay, M. (1994). Two-Armed Dirichlet Bandits With Discounting. *The Annals of Statistics*, 22:1212–1221.

Chernoff, H. (1968). Optimal Stochastic Control. *Sankhya*, 30:221–252.

Clayton, M. and Berry, D. (1985). Bayesian Nonparametric Bandits. *The Annals of Statistics*, 13:1523–1534.

De Blasi, P. (2007). Simulation of the Beta-Stacy Process with Application to Analysis of Censored Data. In *Encyclopedia of Statistics in Quality and Reliability, F. Ruggeri, R.S. Kennt and F. Faltin*, pages 1814–1819, John Wiley & Sons Ltd., Chichester, UK.

de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Annales de l'institut Henri Poincaré*, 7:1–68.

de Finetti, B. (1938). Sur la condition d'equivalence partielle, VI Colloque Geneve. *Acta. Sci. Ind. Paris*, 739:5–18.

de Finetti, B. (1959). La probabilitá e la statistica nei rapporti con l'induzione, secondo i diversi punti di vista. Atti corso CIME su Induzione e Statistica, Varenna.

Doksum, K. (1974). Tailfree and neutral random probabilities and their posterior distributions. *Annals of Probability*, 2:183–201.

Ferguson, T. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1:209–230.

Ferguson, T. (1974). Prior Distributions on Spaces of Probability Measures. *The Annals of Statistics*, 2:615–629.

Ferguson, T. and Phadia, E. (1979). Bayesian Nonparametric Estimation Based on Censored Data. *The Annals of Statistics*, 7:163–186.

Garivier, A. and Moulines, E. (2008). On upper-confidence bound policies for non-stationary bandit problems. Available at https://hal.archives-ouvertes.fr/hal-00281392.

Gill, R. and Johansen, S. (1990). A Survey of Product Integration with a View Toward Application in Survival Analysis. *The Annals of Statistics*, 18:1501–1555.

Gittins, J. (1979). Bandit Processes and Dynamic Allocation Indices (with discussion). *Journal of the Royal Statistical Society, Series B*, 41:148–177.

Gittins, J., Glazebrook, and Weber, R. (2011). *Multi-armed Bandit Allocation Indices*. John Wiley & Sons, Ltd, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom.

Langford, J. and Zhang, T. (2008). The epoch-greedy algorithm for contextual multi-armed bandits. In *Advances in Neural Information Processing Systems 20*, pages 817–284, Curran Associates, Inc.

Muliere, P., Bulla, P., and Walker, S. (2007). Bayesian Nonparametric Estimation of Bivariate Survival Function. *Statistica Sinica*, 17:427–444.

Muliere, P., Paganoni, A., and Secchi, P. (2006). A randomly reinforced urn. *Journal of Statistical Planning and Inference*, 136:1853–1874.

Nash, P. (1973). Optimal Allocation of Resources Between Research Projects. Ph.D. thesis, Cambridge Univ., England.

Phadia, E. (2013). *Prior Processes and Their Applications*. Springer-Verlag, Berlin.

Robbins, H. (1952). Some Aspects of the Sequential Design of Experiments. *Bullettin of American Mathematical Society*, 58:527–535.

Susarla, V. and Van Ryzin, J. (1976). Nonparametric Bayesian estimation of survival curves from incomplete observations. *Journal of the American Statistical Association*, 71:897–902.

Sutton, R. and Barto, A. (1998). *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, Massachusetts.

Thompson, W. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:285–294.

Tokic, M. (2010). Adaptive ϵ-greedy exploration in reinforcement learning based on value differences. In *KI 2010: Advances in Artificial Intelligence, Lecture Notes in Computer Science*, pages 203–210, Springer-Verlag.

Walker, S. and Damien, P. (1998). A full Bayesian Non-Parametric Analysis Involving a Neutral to the Right Process. *Scandinavian Journal of Statistics*, 25:669–680.

Walker, S. and Muliere, P. (1997). Beta-Stacy Processes and a Generalization of the Pólya-Urn Scheme. *The Annals of Statistics*, 25:1762–1780.

Walker, S. and Muliere, P. (2003). A Bivariate Dirichlet Process. *Statistics and Probability Letters*, 64:1–7.

Watkins, C. (1989). Learning from Delayed Rewards. Ph.D. thesis, Cambridge Univ., England.

Yu, Y. (2011). Prior Ordering and Monotonicity in Dirichlet Bandits. Working paper.