Electronic Journal of Statistics
Vol. 11 (2017) 3368–3406
ISSN: 1935-7524
DOI: 10.1214/17-EJS1342

Learning vs earning trade-off with missing or censored observations: The two-armed Bayesian nonparametric beta-Stacy bandit problem

Stefano Peluso

Università Cattolica del Sacro Cuore
Department of Statistical Sciences,
Università della Svizzera Italiana
Institute of Computational Science
e-mail: stefano.peluso@unicatt.it

Antonietta Mira

Università della Svizzera Italiana
Institute of Computational Science,
Università degli Studi dell'Insubria
Department of Sciences and High Technology
e-mail: antonietta.mira@usi.ch

and

Pietro Muliere

Università Commerciale Luigi Bocconi
Department of Decision Sciences,
Bocconi Institute for Data Science and Analytics
e-mail: pietro.muliere@unibocconi.it

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 first show some properties of the expected advantage of choosing one arm over the other, namely the monotonicity in the arm response and, limited to the case of continuous state space, the continuity in the right-censored arm response. We partially characterize optimal strategies by proving the existence of stay-with-a-winner and stay-with-a-winner/switch-on-a-loser break-even points, under non-restrictive conditions that include the special cases of the simple 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 framework.

MSC 2010 subject classifications: Primary 62C10; secondary 62N01.

Received December 2016.

Contents

1. Introduction

1.1. Problem description

In a discrete-stage two-armed bandit problem, there are two stochastic processes (the two arms), and a sequential decision process (a strategy) selects, at each stage, 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 a better known arm 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.

More formally, let X_k, $Y_k \in [0, \infty) =: \mathbb{R}^+$ be random variables (responses) generated from, respectively, arm 1 and 2 at stage k, for $k = 1, 2, \ldots, n$, where $n \in \mathbb{N} \setminus \{0\}$, $\mathbb{N} := \{0, 1, 2, 3, \ldots\}$, is the (possibly infinite) bandit horizon. If X_k and Y_k are the responses of the k-th patient to treatment 1 and 2, the k-th *stage* is interpreted as the moment at which it has to be decided the treatment to assign to the k-th patient, given past responses of the previous $k - 1$ patients to the treatments. Generally speaking, the k-th stage of the bandit problem is the phase when one of the two arms is chosen to be observed, on the basis of the responses of past $k - 1$ subjects. We assume that $X_1, X_2, \ldots, X_n$ given the probability law F^1 are i.i.d with law F^1, and that $Y_1, Y_2, \ldots, Y_n$ given the probability law F^2 are i.i.d with law F^2, with F^1 and F^2 independent. Then, we assume *exchangeable* responses within treatments and independent responses between treatments.

A strategy is interpreted, following Berry and Fristedt (1985), 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, or, equivalently, to which arm to assign the next subject. With the exception of the simplest cases, explicit specifications of optimal strategies are hindered by computational issues. As a consequence, as in Chattopadhyay (1994), optimal strategies can only be partially characterized in terms of *break-even* observations. We will consider two kinds of strategies: stay-with-a-winner and stay-with-a-winner/switch-on-a-loser strategies, assuming, without loss of generality, that a higher realized value of a random variable gives a higher payoff. Intuitively, if at the current stage arm 1 is optimal, according to the stay-with-a-winner strategy, arm 1 is optimally chosen to be observed at the next stage if the observation from arm 1 at the current stage is higher than a break-even point. The stay-with-a-winner break-even point is defined such that the expected advantage of arm 1 over arm 2 is at least as high as it was before the observation at the current stage. On the other hand, in a stay-with-a-winner/switch-on-a-loser strategy, if arm 1 is currently observed, optimally or not, arm 1 will be optimally chosen at the next stage if the current observation is higher than a break-even point (different from the break-even point of the previous strategy), otherwise arm 2 is optimally chosen. The stay-with-a-winner/switch-on-a-loser break-even point is defined such that the expected advantage remains positive after the observation from the arm at the current stage. We will define more formally the two strategies in Section 3.3.

1.2. *Related literature*

Early examples of bandit problems are treated in Robbins (1952), Bellman (1956) and Bradt, Johnson and Karlin (1956). Among later works, Chernoff

(1968) focuses on two Gaussian arms $F^i = N(\mu^i, \sigma^2), i = 1, 2$, with unknown drifts μ^1 and μ^2 of the first and second arm respectively, and known constant and common variance σ^2; Berry (1972) gives sufficient conditions for optimal selection and the existence of a stay-with-a-winner strategy in a Bernoulli two-armed bandit, $F^i = Bern(p^i), i = 1, 2$, where p^i is the unknown probability of observing a realized value of 1 from arm i; 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 $F^1 \sim DP(\alpha)$, the Dirichlet process introduced in Ferguson (1973), with bounded nonnull measure α on $\mathbb{R}$, 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. Dirichlet bandits are generalized to two-armed problems $F^i \sim DP(\alpha^i)$, α^i probability measure on $\mathbb{R}$, $i = 1, 2$, in Chattopadhyay (1994), where the existence of stay-with-a-winner and stay-with-a-winner/switch-on-a-loser optimal strategies is proven. Some other properties of Dirichlet bandits are studied in Yu (2011).

1.3. Our contribution

In 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 (BS) 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, extensively discussed in Section 2, characterize the two beta-Stacy processes. As specified in Phadia (2013), the beta-Stacy process 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. We will discuss in more details in Section 2 the properties of the beta-Stacy process.

The Dirichlet bandit of Clayton and Berry (1985) and Chattopadhyay (1994) is therefore an important special case of our setting, as is the bandit problem with the simple 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 simple homogeneous process and the Dirichlet process), stay-with-a-winner and stay-with-a-winner/switch-on-a-loser break-even points characterizing optimal strategies exist and can be used for dealing with right-censored or exact observations. A right-censored observation is a realized value that is capped by a known censoring level: the observed response from arm 1 at stage k is $x_k = \min\{x_k^*, c_k^x\}$, the minimum between a true exact unobserved x_k^* and a known censoring level c_k^x, and equivalently for arm 2. We stress that we know if each observation has been censored or not, so that in the sequel we will simply denote by $X_1, \dots, X_n$ the exact random variables from arm 1, and by $Y_1, \dots, Y_n$ those from arm 2, with realized values $x_1, \dots, x_n$ and $y_1, \dots, y_n$ each known to be right-censored or not. For every k, c_k^x and c_k^y become known at stage $k+1$, after the response of the k-th subject is observed. We assume that subjects's responses are immediate, in the sense that the (realized) response of subject k is observed together with the information that it is censored or not and with the value of c_k^x (or c_k^y), but in our setting a censored observation will never become exact. A typical example of censored observation in our setting would be the survival time returned at the current stage by a patient who dies for causes not related to the treatment or who abandons the study, with no possibility of having in the future the exact response of the patient.

Arm responses with missing values can be seen as a special case of arms with right-censored observations: since X_k and $Y_k \in \mathbb{R}^+$, a missing observation can be treated as a censored observation with censorship level equal to zero. Coherently to Hardwick, Oehmke and Stout (1998), we then consider a missing observation as an observation subject to the hardest case of censorship: the one giving no information whatsoever on the true response value. On the other hand, in commonly right-censored observations (as in the motivating examples in Introduction) the censorship level obviously provides information on the minimum value of the true exact but unobserved response.

1.4. Some examples motivating bandits with censorship

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: this response can be censored if the patient returns its survival time after the treatment, but she interrupts the treatment or she dies for unrelated causes, or the obervational period ends before her death. The responses are therefore patients' survival times after the treatment (which may be censored) and the objective is to maximize the total discounted expected survival times. Another classical example of a bandit application with censored observations may arise when a manager of several teams of chemical scientists has to decide on the allocation of resources among the teams, with the aim of minimizing the expected time up to the creation of new successful products (Nash 1973): the two arms are the two teams of scientists, and a fixed budget for the creation of the new product is assigned only to one team. The arm response is the time invested by the team to create the successful product, a response that can be right-censored if the project is interrupted due to reduced financial support. A final example of a bandit problem with censored data is that of a batch job scheduling of an industrial processor, choosing which jobs to process at each stage, in the aim of minimizing the whole expected processing time (Gittins, Glazebrook and Weber 2011 and references therein): the arms correspond to the different jobs that the machine has to execute, and the arm response is the time needed to execute a specific task componing the job. The response may be censored if the task has failed and is unfinished for system breakdowns, or if it lasts more than the maximum amount of time allocated to the job. We then know that the task lasts at least up to the breakdown or up to the maximum time, but we do not have the precise task duration.

1.5. Outline of the paper

In Section 2 we introduce and define the beta-Stacy process (Section 2.1), relating it to other known stochastic processes (2.2) and recalling its posterior properties (2.3). The beta-Stacy bandit problem with two discrete-stage arms is detailed in Section 3: we first describe the mechanism of the problem in Section 3.1, then we introduce some further notation in Section 3.2, with particular emphasis on the expected bandit payoff and the expected advantage of choosing arm 1 over arm 2; finally in Section 3.3 we characterize the stay-with-a-winner and stay-with-a-winner/switch-on-a-loser strategies we study. In Sections 4 and 5 we show, respectively for discrete and continuous beta-Stacy arms, some properties of the expected advantage, namely its monotonicity (Sections 4.2 and 5.2) and continuity (5.3) in the arm response, and we then show the existence of break-even points of stay-with-a-winner and stay-with-a-winner/switch-on-a-loser strategies (4.3 and 5.4). We apply our methods to simulated problem instances in Section 6, and conclude with examples of potential further applications and research directions in Section 7.

2. Beta-Stacy process preliminaries

2.1. Introduction and definition

Under the assumption of *exchangeability* of the sequence of random variables $X_1, \dots, X_k, \dots$, with $k \in \mathbb{N} \backslash \{0\}$ and each $X_i \in \mathbb{R}^+$, from de Finetti Representation theorem (de Finetti 1937) there exist a random probability measure P and a corresponding random cumulative distribution function (cdf) F, conditionally on which $X_1, \dots, X_k, \dots$, are i.i.d. from F. That is, there exists a unique probability (or de Finetti) measure Q, defined on the space of probability measures on $(\mathbb{R}^+, \mathcal{A})$, $\mathcal{A}$ the Borel σ-field of subsets of $\mathbb{R}^+$, such that the joint distribution of $X_1, \dots, X_n$, for any $n \in \mathbb{N}$ and events $A_1, \dots, A_k$ in $\mathcal{A}$, can be written as

$$\mathbb{P}\left(X_1 \in A_1, \dots, X_k \in A_k\right) = \int \left\{ \prod_{i=1}^{k} P(A_i) \right\} Q(dP).$$

In our framework F is fixed to be the beta-Stacy process defined below. In the rest of the paper we denote with $\mathbb{E}$ the expected value with respect to the probability measure $\mathbb{P}$. The expected value of F, $\mathbb{E}[F(t)]$ for all $t \in \mathbb{R}^+$, is called the *base measure* of F. Furthermore, the assumption of exchangeability implies, for any event $A \in \mathcal{A}$, that

$$\mathbb{P}\left(X_{k+1} \in A | X_1, \ldots, X_k\right) = \mathbb{E}\left[P(A) | X_1, \ldots, X_n \in A_k\right],$$

with the special case, for any $t \in \mathbb{R}^+$,

$$\mathbb{P}\left(X_{k+1} \leq t | X_1, \ldots, X_k\right) = \mathbb{E}\left[F(t) | X_1, \ldots, X_k\right]. \tag{1}$$

Let the right continuous measure α and the positive function β both be defined on $\mathbb{R}^+$, with $\alpha(0) = 0$. For $t \in \mathbb{R}^+$, we write $\alpha(t)$ for the value of the measure α over the region $[0, t]$, that is $\alpha(t) := \alpha([0, t])$. Let $\alpha\{t\} := \alpha(t) - \alpha(t^-) \geq 0$ for all $t \in \mathbb{R}^+$, where $\alpha(t^-) := \lim_{s \uparrow t} \alpha(s)$. Let $\{t_k\}, k \in \mathbb{N}$, be the countable set of discontinuity points of α, corresponding to jumps $\alpha\{t\} > 0$. Let $\alpha_c(t) = \alpha(t) - \sum_{t_k \leq t} \alpha\{t_k\}$, so that α_c is a continuous measure.

Definition 2.1. *F is a beta-Stacy process on $(\mathbb{R}^+, \mathcal{A})$ with parameters $\alpha(t)$ and $\beta(t)$, $t \in \mathbb{R}^+$ or $t \in \mathbb{N}$, that is $F \sim BS(\alpha, \beta)$, if $F(t) = 1 - \exp\{-Z(t)\}$ for all 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 log moment generating function given by

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

where $1 - \exp(-S_{t_k}) \sim Beta(\alpha\{t_k\}, \beta(t_k))$, and t_k for some $k \in \mathbb{N}$ are the discontinuity points of α. If α is purely atomic on $\mathbb{N}$ (it is strictly positive only at t_k, for some $k \in \mathbb{N}$), we denote the beta-Stacy process to be discrete, otherwise the beta-Stacy process is said to be continuous.

In the previous definition we can intepret $dN_t(v)$ as the rate of arrival (intensity) of a Poisson process with jump of size v, whilst ϕ denotes the argument of the characteristic function $\mathbb{E}\left[-\phi Z(t)\right]$. Note also that $d\alpha(t)$ and $\beta(t)$, $t \in \mathbb{R}^+$ can be respectively thought intuitively as the measure that a priori the Beta-Stacy process assigns to the infinitesimal interval around t, and to the interval (t, ∞). Finally note that it is not relevant to include the point 0 in the domain of the beta-Stacy parameters, since from the assumption $\alpha(0) = 0$ the point zero has always null mass. For $X|F \sim F$, $F \sim BS\left(\alpha, \beta\right)$ discrete[1], from (1) we can write, for all $t \in \mathbb{R}^+$,

$$\mathbb{P}\left(X \leq t\right) = \prod_{j=1}^{t} \left(1 - \frac{\alpha\{j\}}{\alpha\{j\} + \beta(j)}\right),$$

[1] The argument t of F is an element of $\mathbb{R}^+$, and F a random cdf of a random variable with discrete or continuous support. Therefore the domain of F has to be distinguished from the discrete or continuous support of the random variables following the law F, and from the stages of the bandit problem, which in the current paper are always assumed to be discrete.

whilst if $F \sim BS(\alpha, \beta)$ continuous, with α having discontinuity points $\{t_k\}$:

$$\mathbb{P}(X \leq t) = 1 - \exp\left\{-\int_0^t \frac{d\alpha_c(s)}{\alpha\{s\} + \beta(s)}\right\} \prod_{t_k \leq t} \left(1 - \frac{\alpha\{t_k\}}{\alpha\{t_k\} + \beta(t_k)}\right).$$

In order for Q to have a cdf almost surely (a.s.), the parameters α and β of a discrete beta-Stacy process are required to satisfy the condition

$$\prod_{t \in \mathbb{N}} \left(1 - \frac{\alpha\{t\}}{\beta(t) + \alpha\{t\}}\right) = 0. \tag{2}$$

When α has no discontinuity points, the analogue of condition (2) is the requirement of α and β satisfying

$$\int_0^\infty d\alpha(t)/\beta(t) = \infty. \tag{3}$$

When α has both continuous and discrete parts, condition (2) has to hold for all t which are discontinuity points of α, and condition (3) has to hold for α_c, the continuous part of α defined above.

2.2. A first example and relation to other processes

In the current subsection we illustrate an example of beta-Stacy process with one prior discontinuity point, and we clarify under which conditions on the parameters of the process it reduces to the Dirichlet process and to the simple homogeneous process.

As a first instance of a (continuous) beta-Stacy process, we fix, for all $t \in \mathbb{R}^+$ and some $\lambda, l \in \mathbb{R}^+$

$$\alpha(t) = 1 - e^{-\lambda t} + \mathbb{1}_{\{l\}}(t), \qquad \beta(t) = e^{-\lambda t},$$

where, for some event A, $\mathbb{1}_A$ denotes the indicator function of A, equal to 1 if its argument belongs to A and 0 otherwise, and l is the prior discontinuity point. Then it is clear that the Lévy measure is given, for all $t, v \in \mathbb{R}^+$, by

$$dN_t(v) = \frac{dv}{1 - e^{-v}} \left(\frac{e^{-ve^{-\lambda t}} - e^{-v}}{v}\right), \tag{4}$$

and the log moment generating function, for all $\phi \in \mathbb{R}$, can be shown to be equal to

$$\begin{aligned}\log \mathbb{E}\left[e^{-\phi Z(t)}\right] &= \frac{1}{v(1 - e^{-v})}\left(\left(\phi + e^{-\lambda t}\right)^{-1} - e^{\lambda t} + 1 - (\phi + 1)^{-1}\right) \\ &+ \log\left(\sum_{j=0}^{\infty} \binom{\phi + j - 1}{\phi - 1} e^{-\lambda j} B\left(j + 1, e^{-\lambda j}\right)\right) \mathbb{1}_{[-\infty, 0)}(\phi)\end{aligned}$$

$$+\log\left(\frac{e^{-\lambda k}}{e^{-\lambda k}+\phi}\right)\mathbb{1}_{[0,\infty)}(\phi), \tag{5}$$

where $B(\cdot,\cdot)$ is the usual beta function. Therefore, for all $t \in \mathbb{R}^+$,

$$\mathbb{E}[F(t)] = 1 - e^{-t}\frac{1+e^{-\lambda l}-\mathbb{1}_{[l,+\infty)}(t)}{1+e^{-\lambda l}}.$$

When $\beta(t) = \alpha((t,\infty))$ for all $t \in \mathbb{R}^*$ or $t \in \mathbb{N}$, and α any measure on $\mathbb{R}^+$, we obtain the Dirichlet process prior of Ferguson (1973), as in this case $\alpha(t) = 1-e^{-\lambda t}$ and $\beta(t) = e^{-\lambda t}$, with Lévy measure as in (4), log moment generating function equal to the first term in the right hand side of (5), and a base measure with exponential density of parameter λ. Note that if $\beta(t) = \alpha((t,\infty))$ (then a priori a Dirichlet process) and then there are right-censored observations, a posteriori the process is not a Dirichlet process but a more general beta-Stacy, since the relation between the α and β, both updated after the right-censored observations, changes. Another important special case is the homogenous process of Susarla and Van Ryzin (1976) and Ferguson and Phadia (1979), arising when $\beta(t) = \beta \in \mathbb{R}^+$ constant for all $t \in \mathbb{R}^+$. As an example of simple homogeneous process, if we fix $\alpha(t) = 1 - e^{-\lambda t}$ and $\beta(t) = \beta$, it can be shown that the Lévy measure and the corresponding log moment generating function of the process are, for $\phi \in \mathbb{R}$, the following:

$$\begin{aligned} dN_t(v) &= \frac{1-e^{-\lambda t}}{1-e^{-v}}e^{-v\beta}dv \\ \log\mathbb{E}\left[e^{-\phi Z(t)}\right] &= \left(e^{-\lambda t}-1\right)\left(H_{\beta+\phi-1}-H_{\beta-1}\right), \end{aligned}$$

where $H_x = \int_0^1 (1-y^x)(1-y)\,dy$ is the harmonic number for $x \in \mathbb{R}\setminus\{-1,-2,\dots\}$. The base measure of the simple homogeneous process is

$$\mathbb{E}[F(t)] = 1 - \exp\left\{-\frac{1}{\beta\lambda}\left(e^{-\lambda t-1}\right)\right\},\ t \in \mathbb{R}^+$$

In the rest of the paper, we only consider beta-Stacy processes in the general formulation of Definition 2.1, but it is reasonable to conjecture that the results can be generalized to the class of Neutral to the Right (NTR) processes (Doksum 1974). The NTR process 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 so-called log-beta process (Walker and Muliere 1997), that keeps the conjugacy property under sampling exact or right-censored observations.

2.3. Posterior properties

We now state the theorem of Walker and Muliere (1997) on the conjugacy of the beta-Stacy process.

Theorem 2.2. *(Walker and Muliere 1997) Assume we observe $X_k = x_k$, for $k = 1, \dots, n$, $n \in \mathbb{N}$ denoting the sample size, and such that $X_k|F \sim F$, where $F \sim BS(\alpha, \beta)$ is a discrete (continuous) beta-Stacy process. We partition $\mathbf{x} = (x_1, \dots, x_n)$ as $[\mathbf{x}^{exact}, \mathbf{x}^{cens}]$ for respectively exact and right-censored observations. Then the posterior $F|(X_1 = x_1, \dots, X_n = x_n)$ is also a discrete (continuous) beta-Stacy process $BS(\alpha + N_{\mathbf{x}}, \beta + M_{\mathbf{x}})$, where, for all $t \in \mathbb{R}^+$, $N_{\mathbf{x}}\{t\} = \sum_{j:x_j \in \mathbf{x}^{exact}} \mathbb{1}_{\{x_j\}}(t)$ is the number of exact observations equal to t and $M_{\mathbf{x}}(t) = \sum_{i:x_i \in \mathbf{x}^{exact}} \mathbb{1}_{[0,x_i)}(t) + \sum_{i:x_i \in \mathbf{x}^{cens}} \mathbb{1}_{[0,x_i]}(t)$ is the sum of the number of exact observations greater than t and censored observations greater or equal to t.*

The theorem above clarifies an important property of the (continuous or discrete) beta-Stacy process: its conjugacy under sampling, possibly with right censoring. A posteriori (after the observation of $\mathbf{x}$) the corresponding jumps S_t, for all $t \in \mathbf{x}$, are such that

$$1 - \exp(-S_t) \sim Beta(\alpha\{t\} + N_{\mathbf{x}}\{t\}, \beta(t) + M_{\mathbf{x}}(t)).$$

From the conjugacy property and equation (1), for $X_1, \dots, X_n$ exchangeable from $F \sim BS(\alpha, \beta)$ continuous and α with discontinuity points $\{t_k\}$ each in $\mathbb{R}^+$ or in $\mathbb{N}$, we can write, for any $n \in \mathbb{N}$ and $t \in \mathbb{R}^+$

$$\begin{aligned}\mathbb{P}(X_{n+1} \le t | X_1, \dots, X_n) \quad = \quad & 1 - \exp\left\{ - \int_0^t \frac{d\alpha_c(s)}{\alpha\{s\} + \beta(s) + N_{\mathbf{x}}\{s\} + M_{\mathbf{x}}(s)} \right\} \\ & \cdot \prod_{t_k \le t} \left(1 - \frac{\alpha\{t_k\} + N_{\mathbf{x}}\{t_k\}}{\alpha\{t_k\} + \beta(t_k) + N_{\mathbf{x}}\{t_k\} + M_{\mathbf{x}}(t_k)} \right),\end{aligned}$$

which specializes, for α with no discontinuity points, to

$$\mathbb{P}(X_{n+1} \le t | X_1, \dots, X_n) = 1 - \exp\left\{ - \int_0^t \frac{d\alpha(s)}{\beta(s) + N_{\mathbf{x}}\{s\} + M_{\mathbf{x}}(s)} \right\}$$

and, for the discrete beta-Stacy process and for all $t \in \mathbb{N}$ to

$$\mathbb{P}(X_{n+1} \le t | X_1, \dots, X_n) = \prod_{j=1}^{t} \left(1 - \frac{\alpha\{j\} + N_{\mathbf{x}}\{j\}}{\alpha\{j\} + \beta(j) + N_{\mathbf{x}}\{j\} + M_{\mathbf{x}}(j)} \right).$$

Note that the update of the discrete or continuous beta-Stacy parameters keeps track of not only the number of observations (censored or not), but also of their values.

3. The discrete-stage two-armed bandit problem

3.1. Beta-Stacy bandit problem formulation

In the proposed framework, $\left(\{\alpha^1, \beta^1\}, \{\alpha^2, \beta^2\}; \mathbf{A}_n\right)$ denotes the two-armed beta-Stacy bandit problem, with arm i having a beta-Stacy prior $BS(\alpha^i, \beta^i)$, for

$i = 1, 2$, and with $\mathbf{A}_n = (a_1, a_2, \ldots, a_n)$ being a nonincreasing discount sequence. Therefore the special choice of β^i that follows the discussion after Equation (5), together with the absence of censored observations, reduce our setting to the Dirichlet bandit problem $(\{\alpha^1\}, \{\alpha^2\}; \mathbf{A}_n)$ of Chattopadhyay (1994). The objective in the bandit problem is to maximize the expected payoff, more precisely defined in the next subsection. Assuming without loss of generality that a higher response from the arms corresponds to a higher payoff, we want to choose at each stage which arm to observe with the aim of maximizing in expectation the sum of all observations from the arms.

We only consider bandit problems with discrete stages: at the beginning of stage 1 (in the disease motivating example, before the assigment of the first patient to treatment 1 or 2) it is chosen which arm to observe, only on the basis of α^1, β^1 and α^2, β^2: if the first arm is chosen, we will observe some realized value of X_1, possibly right-censored, where $X_1|F^1 \sim F^1$ and $F^1 \sim BS\left(\alpha^1, \beta^1\right)$; if the second arm is chosen, we will observe some realized value of Y_1, possibly right-censored, where $Y_1|F^2 \sim F^2$ and $F^2 \sim BS\left(\alpha^2, \beta^2\right)$. Therefore at stage 1 the best arm is chosen for the first subject, optimally only on the basis of the prior choices of α^1, β^1 and α^2, β^2, since no previous observations is available yet. Taking into account the additional information coming from the observation at stage 1 of X_1 or Y_1, we will choose the arm to observe at stage 2, in a way that maximizes the expected payoff. Intuitively, if for instance a high realized value of X_1 is observed, it will be more likely to observe X_2 instead of Y_2, that is to observe again from the same arm at the next stage, in a trade-off between the exploitation of an arm that is better known to return high observations, and the exploration of the potentially better but less known arm. At stage $k > 1$, $k \in \mathbb{N}$, we decide to observe X_k from arm 1 or Y_k from arm 2 on the basis of the obervations $[\mathbf{x}_{k-1}, \mathbf{y}_{k-1}] = [\mathbf{x}_{k-1}^{exact}, \mathbf{x}_{k-1}^{cens}, \mathbf{y}_{k-1}^{exact}, \mathbf{y}_{k-1}^{cens}]$ from the past $k-1$ stages, where $\mathbf{x}_{k-1}$ are defined to be the observations from arm 1 at stages from 1 to $k-1$, partitioned in $[\mathbf{x}_{k-1}^{exact}, \mathbf{x}_{k-1}^{cens}]$ for exact and right-censored observations, and similarly for $\mathbf{y}_{k-1}$ and $[\mathbf{y}_{k-1}^{exact}, \mathbf{y}_{k-1}^{cens}]$ from arm 2.

It is important to highlight that we assume throughout that we know if an observation is right-censored or not. Furthermore, the elements of

$$[\mathbf{x}_{k-1}^{exact}, \mathbf{x}_{k-1}^{cens}, \mathbf{y}_{k-1}^{exact}, \mathbf{y}_{k-1}^{cens}]$$

can be empty: for instance if all the observations from arm 1 are exact up to stage $k-1$, $\mathbf{x}_{k-1}^{cens} = \emptyset$, or if arm 2 has never been observed up to stage $k-1$, $\mathbf{y}_{k-1}^{exact} = \mathbf{y}_{k-1}^{cens} = \emptyset$

For all $k \in \mathbb{N}$, $X_k|(F^1, [\mathbf{x}_{k-1}, \mathbf{y}_{k-1}]) \overset{d}{=} X_k|F^1 \sim F^1$, where $F^1 \sim BS\left(\alpha^1, \beta^1\right)$ (and similarly for Y_k), but with no conditioning on F^1, there is a dependence between X_k and previous observations from arm 1:

$$X_k|[\mathbf{x}_{k-1}, \mathbf{y}_{k-1}] = X_k|\mathbf{x}_{k-1} = \int X_k|F^1 \; dF_{k-1}^1, \tag{6}$$

and similarly for Y_k, with

$$F_{k-1}^1 \quad \sim \quad BS(\alpha_{\mathbf{x}_{k-1}}^1, \beta_{\mathbf{x}_{k-1}}^1),$$

$$F^2_{k-1} \quad \sim \quad BS(\alpha^2_{\mathbf{y}_{k-1}}, \beta^2_{\mathbf{y}_{k-1}}),$$

beta-Stacy processes with parameters updated in accordance to Theorem 2.2 above, defined as:

$$\alpha^1_{\mathbf{x}_{k-1}} \quad := \quad \alpha^1 + N_{\mathbf{x}_{k-1}}, \qquad \beta^1_{\mathbf{x}_{k-1}} := \beta^1 + M_{\mathbf{x}_{k-1}}, \tag{7}$$

$$\alpha^2_{\mathbf{y}_{k-1}} \quad := \quad \alpha^2 + N_{\mathbf{y}_{k-1}}, \qquad \beta^2_{\mathbf{y}_{k-1}} := \beta^2 + M_{\mathbf{x}_{k-1}}, \tag{8}$$

where, coherently with the notation set up in Section 2.3, for all $t \in \mathbb{R}^+$ (or in $\mathbb{N}$ for the discrete beta-Stacy process),

$$\begin{aligned} N_{\mathbf{x}_{k-1}}\{t\} &= \sum_{j: x_j \in \mathbf{x}^{exact}_{k-1}} \mathbb{1}_{\{x_j\}}(t), \\ M_{\mathbf{x}_{k-1}}(t) &= \sum_{j: x_j \in \mathbf{x}^{exact}_{k-1}} \mathbb{1}_{[0,x_j)}(t) + \sum_{j: x_j \in \mathbf{x}^{cens}_{k-1}} \mathbb{1}_{[0,x_j]}(t), \end{aligned}$$

and similarly for $N_{\mathbf{y}_{k-1}}$ and $M_{\mathbf{y}_{k-1}}$. For notational convenience, we also define quantities prior to any information as $F^i_0 := F^i$, $i = 1, 2$, $\alpha^1_{\mathbf{x}_0} := \alpha^1$, $\alpha^2_{\mathbf{y}_0} := \alpha^2$, $\beta^1_{\mathbf{x}_0} := \beta^1$ and $\beta^2_{\mathbf{y}_0} := \beta^2$.

3.2. Expected payoff and advantage

As detailed in the previous section, in a bandit problem $(\{\alpha^1, \beta^1\}, \{\alpha^2, \beta^2\}; \mathbf{A}_n)$, a strategy selects at each stage $k = 1, \dots, n$ which arm to observe, on the basis of past observations from the two arms. Then, a strategy can be characterized by a n-dimensional binary vector $\mathbf{\Gamma} = (\gamma_1, \dots, \gamma_n)$, where for $k = 1, \dots, n$,

$$\gamma_k := \gamma_k([\mathbf{x}_{k-1}, \mathbf{y}_{k-1}]) = \begin{cases} 1, & \text{if observe arm 1} \\ 0, & \text{if observe arm 2} \end{cases},$$

with γ_k dependent on past observations from both arms. Without loss of generality, we assume that higher observations are *better*, so that, for the discount sequence $\mathbf{A}_n = (a_1, \dots, a_n)$, we can write the payoff as

$$\sum_{k=1}^{n} a_k \left(\gamma_k X_k + (1 - \gamma_k) Y_k\right).$$

An optimal strategy maximizes the expected payoff, that is the expected discounted sum of arms responses. With the exception of the simplest cases, explicit characterizations of optimal strategies are hindered by computational difficulties, imposing the need of partial characterizations of optimal strategies via break-even observations (Chattopadhyay 1994; Clayton and Berry 1985). In particular, we will prove the existence of stay-with-a-winner and stay-with-a-winner/switch-on-a-loser break-even points. Similarly to Chattopadhyay (1994), we let

$$W(\{\alpha^1, \beta^1\}, \{\alpha^2, \beta^2\}; \mathbf{A}_n) := \sup_{\mathbf{\Gamma}} \mathbb{E} \left\{ \sum_{k=1}^{n} a_k \left(\gamma_k X_k + (1 - \gamma_k) Y_k\right) \right\}$$

be the expected payoff under an optimal strategy, whilst $W^i(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};\mathbf{A}_n)$ is defined to be the expected payoff of a strategy starting from arm i and proceeding optimally. We define $\Delta(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};\mathbf{A}_n)$ to be the expected advantage of initially choosing arm 1 over arm 2 assuming optimal continuation, that is

$$\Delta(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};\mathbf{A}_n) = \\ W^1(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};\mathbf{A}_n) - W^2(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};\mathbf{A}_n).$$

Furthermore, we use the notation

$$\Delta^+(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};\mathbf{A}_n) := \max(0,\Delta(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};\mathbf{A}_n))$$

and

$$\Delta^-(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};\mathbf{A}_n) := \min(0,\Delta(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};\mathbf{A}_n)).$$

All the quantities defined above can be written more generally, substituting to the prior beta-Stacy parameters, the correspondent posterior parameters. For instance,

$$\Delta(\{\alpha^1_{\mathbf{x}_{k-1}},\beta^1_{\mathbf{x}_{k-1}}\},\{\alpha^2_{\mathbf{y}_{k-1}},\beta^2_{\mathbf{y}_{k-1}}\};\mathbf{A}_n^{k-1})$$

is the expected advantage of choosing arm 1 over arm 2 at stage $k \in \mathbb{N}$, after the observation of $[\mathbf{x}_{k-1},\mathbf{y}_{k-1}]$ from the arms in the preceding $k-1$ stages, and where $\mathbf{A}_n^{k-1} := (a_k, a_{k+1},\dots,a_n)$.

3.3. Bandit strategies with optimal properties

Let $X_k|F^1 \sim F^1$, $Y_k|F^2 \sim F^2$, with $F^1 \sim BS(\alpha^1,\beta^1)$, $F^2 \sim BS(\alpha^2,\beta^2)$, F^1 and F^2 independent, and bandit stages $k \in \{1,2,\dots,n\}$, $n \in \mathbb{N}$. We study two strategies: stay-with-a-winner and stay-with-a-winner/switch-on-a-loser strategies, following the nomenclature of Chattopadhyay (1994). As anticipated in Section 1, in a stay-with-a-winner strategy, the arm currently chosen to be observed is again observed at the next stage if its expected advantage, relative to the alternative arm, is higher than the expected advantage computed before the current observation of the arm. Therefore, an optimal arm chosen at the current stage will again be optimal at the next stage if chosen by the stay-with-a-winner strategy. We now characterize this first strategy:

Definition 3.1. The *stay-with-a winner* strategy at stage $k = 1$ selects arm 1 if

$$\Delta\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};\mathbf{A}_n\right) > 0$$

and arm 2 otherwise. At stage $k > 1$, after the observation of $[\mathbf{x}_{k-1},\mathbf{y}_{k-1}]$, the strategy chooses to observe arm 1 if

$$\Delta\left(\{\alpha^1_{\mathbf{x}_{k-1}},\beta^1_{\mathbf{x}_{k-1}}\},\{\alpha^2_{\mathbf{y}_{k-1}},\beta^2_{\mathbf{y}_{k-1}}\};\mathbf{A}_n^{k-1}\right) \geq \\ \Delta\left(\{\alpha^1_{\mathbf{x}_{k-2}},\beta^1_{\mathbf{x}_{k-2}}\},\{\alpha^2_{\mathbf{y}_{k-2}},\beta^2_{\mathbf{y}_{k-2}}\};\mathbf{A}_n^{k-2}\right),$$

and selects arm 2 otherwise. The stay-with-a-winner break-even point at stage $k-1$, $k>1$, is the realized value of X_{k-1} from the first arm (or Y_{k-1} from the second arm) for which the inequality above becomes an equality, making the strategy indifferent in the choice of the two arms at the next stage k.

On the other hand, in a stay-with-a-winner/switch-on-a-loser strategy, the currently observed arm is optimal at the next stage and it will be chosen if the observation is higher than the break-even point at the current stage; if not, the optimal arm to observe at the next stage is the other one.

Definition 3.2. The *stay-with-a-winner/switch-on-a-loser* strategy at stage $k \geq 1$, after the observation of $[\mathbf{x}_{k-1}, \mathbf{y}_{k-1}]$, chooses to observe arm 1 if

$$\Delta\left(\{\alpha^1_{\mathbf{x}_{k-1}}, \beta^1_{\mathbf{x}_{k-1}}\}, \{\alpha^2_{\mathbf{y}_{k-1}}, \beta^2_{\mathbf{y}_{k-1}}\}; \mathbf{A}_n^{k-1}\right) \geq 0$$

and selects arm 2 otherwise. The stay-with-a-winner/switch-on-a-loser break-even point at stage $k-1$, $k<n$ is the realized value of X_{k-1} from the first arm (or Y_{k-1} from the second arm) for which the inequality above becomes an equality, making the strategy indifferent in the choice of the two arms at the following stage k.

4. Bandit solution with discrete beta-Stacy processes

4.1. Framework setting

The first (second) bandit arm is observed as long as it yields a value higher (lower) than the break-even point. Let F^i be the random distribution function corresponding to arm i, $F^i \sim BS(\alpha^i, \beta^i)$ discrete, with X_k and Y_k having supports in $\mathbb{N}$ for all $k \geq 1$, for $i = 1, 2$. Omitting for notational convenience from now on the dependence of $\mathbb{P}$ and $\mathbb{E}$ on $\alpha^1, \beta^1, \alpha^2, \beta^2$, at stage 1 and for all $t \in \mathbb{N}$, we have

$$\mathbb{P}(X_1 = t) = \frac{\alpha^1\{t\}}{\alpha^1\{t\} + \beta^1(t)} \prod_{j=0}^{t-1} \left(1 - \frac{\alpha^1\{j\}}{\alpha^1\{j\} + \beta^1(j)}\right)$$

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

$$\mathbb{E}[X_1] = \sum_{t=1}^{+\infty} \mathbb{P}(X_1 \geq t) = \sum_{t=1}^{+\infty} \prod_{j=0}^{t-1} \left(1 - \frac{\alpha^1\{j\}}{\alpha^1\{j\} + \beta^1(j)}\right) =: \mu^1$$

and

$$\mathbb{E}[Y_1] = \sum_{t=1}^{+\infty} \prod_{j=0}^{t-1} \left(1 - \frac{\alpha^2\{j\}}{\alpha^2\{j\} + \beta^2(j)}\right) =: \mu^2.$$

Given observations $[\mathbf{x}_{k-1}, \mathbf{y}_{k-1}] = [\mathbf{x}^{exact}_{k-1}, \mathbf{x}^{cens}_{k-1}, \mathbf{y}^{exact}_{k-1}, \mathbf{y}^{cens}_{k-1}]$ from arms 1 and 2 up to stage $k-1$, the conditional expectation of any function $h(X)$ can be computed using Theorem 2.2, and it is denoted by $\mathbb{E}[h(X_k)|\mathbf{x}_{k-1}]$. In the sequel, the updates a posteriori of α^i and β^i, $i = 1, 2$, follow the notation introduced in formulae (7) and (8): for instance, $\alpha^1_{\mathbf{x}_{k-1}}$ is the update of α^1 after having observed $\mathbf{x}_{k-1}$, and it is a fixed measure since $\mathbf{x}_{k-1}$ is fully observed; similarly, $\alpha^1_{X_1}$ is the random measure (since X_1 is random) that updates α^1 by taking into account the randomness of the first arm at the first stage. Then $\mathbb{P}(X_k = t|\mathbf{x}_{k-1})$ is, for $t \in \mathbb{N}$,

$$\mathbb{P}(X_k = t|\mathbf{x}_{k-1}) = \frac{\alpha^1_{\mathbf{x}_{k-1}}\{t\}}{\alpha^1_{\mathbf{x}_{k-1}}\{t\} + \beta^1_{\mathbf{x}_{k-1}}(t)} \prod_{j=0}^{t-1} \left(1 - \frac{\alpha^1_{\mathbf{x}_{k-1}}\{j\}}{\alpha^1_{\mathbf{x}_{k-1}}\{j\} + \beta^1_{\mathbf{x}_{k-1}}(j)}\right),$$

and the posterior mean is

$$\mathbb{E}[X_k|\mathbf{x}_{k-1}] = \sum_{t=1}^{+\infty} \prod_{j=0}^{t-1} \left(1 - \frac{\alpha^1_{\mathbf{x}_{k-1}}\{j\}}{\alpha^1_{\mathbf{x}_{k-1}}\{j\} + \beta^1_{\mathbf{x}_{k-1}}(j)}\right) =: \mu^1_{\mathbf{x}_{k-1}}.$$

A useful result we will often use below and that we show in the Appendix is that the expected advantage of arm 1 over arm 2 can be written as

$$\begin{aligned} \Delta(\{\alpha^1, \beta^1\}, \{\alpha^2, \beta^2\}; \mathbf{A}_n) \quad = \quad & (a_1 - a_2)(\mu^1 - \mu^2) \\ & + \mathbb{E}\left[\Delta^+(\{\alpha^1_{X_1}, \beta^1_{X_1}\}, \{\alpha^2, \beta^2\}; \mathbf{A}^1_n)\right] \\ & + \mathbb{E}\left[\Delta^-(\{\alpha^1, \beta^1\}, \{\alpha^2_{Y_1}, \beta^2_{Y_1}\}; \mathbf{A}^1_n)\right] \qquad (9) \end{aligned}$$

where, coherently with the notation introduced, for $t \in \mathbb{N}$

$$\begin{aligned} \alpha^1_{X_1} \quad &= \quad \alpha^1 + N_{X_1}, \quad N_{X_1}\{t\} = \begin{cases} \mathbb{1}_{\{X_1\}}(t), & X_1 \text{ exact} \\ 0, & X_1 \text{ right-censored} \end{cases}, \\ \beta^1_{X_1} \quad &= \quad \beta^1 + M_{X_1}, \quad M_{X_1}(t) = \begin{cases} \mathbb{1}_{[0,X_1)}(t), & X_1 \text{ exact} \\ \mathbb{1}_{[0,X_1]}(t), & X_1 \text{ right-censored} \end{cases}, \end{aligned}$$

with similar construction for $\alpha^2_{X_1}$ and $\beta^2_{X_1}$. In general, for $k \in \{1, 2, \ldots, n\}$:

$$\begin{aligned} \Delta(\{\alpha^1_{\mathbf{x}_{k-1}}, \beta^1_{\mathbf{x}_{k-1}}\}, \{\alpha^2_{\mathbf{y}_{k-1}}, \beta^2_{\mathbf{y}_{k-1}}\}; \mathbf{A}^{k-1}_n) = & (a_k - a_{k+1})(\mu^1_{\mathbf{x}_{k-1}} - \mu^2_{\mathbf{y}_{k-1}}) \\ & + \mathbb{E}\left[\Delta^+(\{\alpha^1_{[\mathbf{x}_{k-1}, X_k]}, \beta^1_{[\mathbf{x}_{k-1}, X_k]}\}, \{\alpha^2, \beta^2\}; \mathbf{A}^k_n)\right] \\ & + \mathbb{E}\left[\Delta^-(\{\alpha^1, \beta^1\}, \{\alpha^2_{[\mathbf{y}_{k-1}, Y_k]}, \beta^2_{[\mathbf{y}_{k-1}, Y_k]}\}; \mathbf{A}^k_n)\right], \qquad (10) \end{aligned}$$

where $\alpha^1_{[\mathbf{x}_{k-1}, X_k]} = \alpha^1_{\mathbf{x}_{k-1}} + N_{X_k}$, and similarly for all other quantities of interest in (10).

4.2. Monotonicity of the expected advantage

In the next proposition we show that, given an exact or a right-censored observation $X_1 = x$ from arm 1, the expected advantage of choosing arm 1 over arm 2 at stage 2 increases as x increases. We remark that we prove this and all subsequent results for the expected advatage at stage 2, and after the observation of arm 1 at stage 1, but identical statements can proved in the same way for the expected advantage at stage $k \geq 1$ after the observation of $[\mathbf{x}_{k-1}, \mathbf{y}_{k-1}]$ up to stage $k-1$, at the cost of the slight increase in the notational burden of substituting in the next propositions and theorems α^1_x and β^1_x with $\alpha^1_{[\mathbf{x}_{k-1},x]}$ and $\beta^1_{[\mathbf{x}_{k-1},x]}$ and α^2 and β^2 with $\alpha^2_{\mathbf{x}_{k-1}}$ and $\beta^2_{\mathbf{x}_{k-1}}$. Furthermore, all the results are stated assuming that the arm to observe at stage 1 is the first one, but they can all similarly be stated in the case when arm 2 is observed at stage 1.

Proposition 4.1. *For all α^1, β^1 and α^2, β^2 such that $\beta^1(t) \leq \beta^1(t+1) + \alpha^1\{t+1\}$, for all $t \in \mathbb{N}$, and for all nonincreasing discount sequences $\mathbf{A}_n$,*

$$\Delta\left(\{\alpha^1_x, \beta^1_x\}, \{\alpha^2, \beta^2\}; \mathbf{A}_n\right)$$

is nondecreasing in x, for all $x \in \mathbb{N}$.

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

$$\begin{aligned}\Delta\left(\{\alpha^1_x, \beta^1_x\}, \{\alpha^2, \beta^2\}; \mathbf{A}_1\right) &= a_1\left(\sum_{t=1}^{+\infty}\prod_{j=0}^{t-1}\left(1 - \frac{\alpha^1_x\{j\}}{\alpha^1_x\{j\} + \beta^1_x(j)}\right) - \mu^2\right) \\ &= a_1(\mu^1_x - \mu^2). \end{aligned} \tag{11}$$

Fix $x^* = x+1$. We first prove that $\mu^1_{x^*} - \mu^1_x \geq 0$. For this purpose, we study separately the t-terms in the sum of μ^1_x and $\mu^1_{x^*}$ when $t \leq x$, $t = x^*$ and $t > x^*$. When x is an exact observation,

- The t-terms with $t \leq x$ are the same in μ^1_x and $\mu^1_{x^*}$.
- For $t = x^*$, in μ^1_x we have

  $$\prod_{j=0}^{t-1}\left(\frac{\beta^1(j)+1}{\alpha^1\{j\} + \beta^1(j) + 1}\right)\frac{\beta^1(x)}{\alpha^1\{x\} + \beta^1(x) + 1},$$

  whilst in $\mu^1_{x^*}$,

  $$\prod_{j=0}^{t-1}\left(\frac{\beta^1(j)+1}{\alpha^1\{j\} + \beta^1(j) + 1}\right)\frac{\beta^1(x)+1}{\alpha^1\{x\} + \beta^1(x) + 1},$$

  and the x^*-term of $\mu^1_{x^*}$ is higher than or equal to the corresponding term in μ^1_x.
- For $t > x^*$, the t-term of μ^1_x is

  $$\prod_{j=0}^{x-1}\left(\frac{\beta^1(j)+1}{\alpha^1\{j\} + \beta^1(j) + 1}\right)\frac{\beta^1(x)}{\alpha^1\{x\} + \beta^1(x) + 1}$$
  $$\cdot\frac{\beta^1(x^*)}{\alpha^1\{x^*\} + \beta^1(x^*)}\prod_{j=x^*+1}^{t-1}\frac{\beta^1(j)}{\alpha^1\{j\} + \beta^1(j)},$$

whilst the t-term of $\mu^1_{x^*}$ is

$$\prod_{j=0}^{x-1}\left(\frac{\beta^1(j)+1}{\alpha^1\{j\}+\beta^1(j)+1}\right)\frac{\beta^1(x)+1}{\alpha^1\{x\}+\beta^1(x)+1}$$
$$\cdot\frac{\beta^1(x^*)}{\alpha^1\{x^*\}+\beta^1(x^*)+1}\prod_{j=x^*+1}^{t-1}\frac{\beta^1(j)}{\alpha^1\{j\}+\beta^1(j)};$$

the t-term of $\mu^1_{x^*}$ is higher than or equal to the corresponding term in μ^1_x 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 μ^1_x can be proved when x is a right-censored observation: the t-terms with $t\leq x^*$ are the same in μ^1_x and $\mu^1_{x^*}$, whilst for $t>x^*$ the two terms in, respectively, μ^1_x and $\mu^1_{x^*}$ are

$$\prod_{j=0}^{x-1}\left(\frac{\beta^1(j)+1}{\alpha^1\{j\}+\beta^1(j)+1}\right)\frac{\beta^1(x)+1}{\alpha^1\{x\}+\beta^1(x)+1}$$
$$\cdot\frac{\beta^1(x^*)}{\alpha^1\{x^*\}+\beta^1(x^*)}\prod_{j=x^*+1}^{t-1}\frac{\beta^1(j)}{\alpha^1\{j\}+\beta^1(j)},$$
$$\prod_{j=0}^{x-1}\left(\frac{\beta^1(j)+1}{\alpha^1\{j\}+\beta^1(j)+1}\right)\frac{\beta^1(x)+1}{\alpha^1\{x\}+\beta^1(x)+1}$$
$$\cdot\frac{\beta^1(x^*)+1}{\alpha^1\{x^*\}+\beta^1(x^*)+1}\prod_{j=x^*+1}^{t-1}\frac{\beta^1(j)}{\alpha^1\{j\}+\beta^1(j)},$$

where the term in $\mu^1_{x^*}$ is higher than or equal to the corresponding term in μ^1_x. Then, for $n=1$ the statement is true since μ^1_x is nondecreasing in x and $a_1\geq 0$. From the induction hypothesis, we assume the monotonic property for $n=m-1$ for some natural number $m>1$. By (10),

$$\begin{aligned}\Delta(\{\alpha^1_x,\beta^1_x\},\{\alpha^2,\beta^2\};\mathbf{A}_m)&=(a_1-a_2)(\mu^1_x-\mu^2)\\&+\mathbb{E}\left[\Delta^+(\{\alpha^1_{[x,X_2]},\beta^1_{[x,X_2]}\},\{\alpha^2,\beta^2\};\mathbf{A}^1_m)\right]\\&+\mathbb{E}\left[\Delta^-(\{\alpha^1_x,\beta^1_x\},\{\alpha^2_{Y_2},\beta^2_{Y_2}\};\mathbf{A}^1_m)\right].\end{aligned}\tag{12}$$

The first term in the right hand side of (12) is nondecreasing in x since μ^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 4.2. The constraints $\beta^1(t)\leq\beta^1(t+1)+\alpha^1\{t+1\}$, for all $t\in\mathbb{N}$, are enough for the monotonicity of μ^1_x. The constraint is naturally verified in the

Dirichlet two-armed problem, obtained from the beta-Stacy in the special case of $\beta^1(t) = \beta^1(t+1) + \alpha^1\{t+1\}$, $t \in \mathbb{N}$. Also, a bandit problem with simple homogeneous processes (Susarla and Van Ryzin 1976; Ferguson and Phadia 1979) for each arm, corresponding to the case $\beta^1(t+1) = \beta^1(t)$ for all $t \in \mathbb{N}$, satisfies the constraints.

4.3. Existence of break-even points

We now prove the existence of stay-with-a-winner and stay-with-a-winner/switch-on-a-loser break-even points in a bandit problem with discrete beta-Stacy processes. The following proposition is preliminary to Theorems 4.5 and 4.6.

Proposition 4.3. *For all α^1, β^1 and α^2, β^2 as in Proposition 4.1 and for all nonincreasing discount sequences $\mathbf{A}_n$,*

$$\Delta\left(\{\alpha^1_{x=0}, \beta^1_{x=0}\}, \{\alpha^2, \beta^2\}; \mathbf{A}_n\right) = \inf_x \Delta\left(\{\alpha^1_x, \beta^1_x\}, \{\alpha^2, \beta^2\}; \mathbf{A}_n\right).$$

Furthermore, if the condition

$$\prod_{t\in\mathbb{N}} \left(1 - \frac{\alpha^1\{t\}}{\alpha^1\{t\} + \beta^1(t) + 1}\right) > 0 \tag{13}$$

is verified, then

$$\lim_{x\to+\infty} \Delta\left(\{\alpha^1_x, \beta^1_x\}, \{\alpha^2, \beta^2\}; \mathbf{A}_n\right) = \infty.$$

Proof. The result for the $x \to 0$ is a direct consequence of the monotonicity property shown in Proposition 4.1. We are left to prove the limit to $+\infty$. Consider x increasing to ∞. By induction, for $n = 1$, μ^1_x diverges to $+\infty$ as $x \to +\infty$ since

$$\begin{aligned} \lim_{x\to+\infty} \mu^1_x \quad &\geq \quad \sum_{t=1}^{+\infty} \prod_{j\in\mathbb{N}} \frac{\beta^1(j)+1}{\alpha^1\{j\} + \beta^1(j) + 1} \\ &= \quad \prod_{j\in\mathbb{N}} \left(1 - \frac{\alpha^1\{j\}}{\alpha^1\{j\} + \beta^1(j) + 1}\right) \cdot \sum_{t=1}^{+\infty} 1 = +\infty. \end{aligned} \tag{14}$$

Then, $\Delta\left(\{\alpha^1_x, \beta^1_x\}, \{\alpha^2, \beta^2\}; \mathbf{A}_1\right) = a_1(\mu^1_x - \mu^2)$ goes to $+\infty$ since μ^1_x is divergent and $a_1 > 0$. Assume now that the statement is true for $n = m-1$, for some natural number $m > 1$. By (12),

$$\begin{aligned} \Delta(\{\alpha^1_x, \beta^1_x\}, \{\alpha^2, \beta^2\}; \mathbf{A}_m) &= (a_1 - a_2)(\mu^1_x - \mu^2) \\ &+ \mathbb{E}\left[\Delta^+(\{\alpha^1_{[x,X_2]}, \beta^1_{[x,X_2]}\}, \{\alpha^2, \beta^2\}; \mathbf{A}^1_m)\right] \\ &+ \mathbb{E}\left[\Delta^-(\{\alpha^1_x, \beta^1_x\}, \{\alpha^2_{Y_2}, \beta^2_{Y_2}\}; \mathbf{A}^1_m)\right]. \end{aligned}$$

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 $\Delta^+(\{\alpha^1_{[x,X_2]}, \beta^1_{[x,X_2]}\}, \{\alpha^2, \beta^2\}; \mathbf{A}^1_m)$ is a nondecreasing sequence in x (by Proposition 4.1), bounded below by 0 (by definition) and divergent to $+\infty$ (by the induction hypothesis). We can then apply the monotone convergence theorem and obtain

$$\begin{aligned}&\lim_{x\to+\infty} \mathbb{E}\left[\Delta^+(\{\alpha^1_{[x,X_2]}, \beta^1_{[x,X_2]}\}, \{\alpha^2, \beta^2\}; \mathbf{A}^1_m)\right]\\ &\quad = \mathbb{E}\left[\lim_{x\to+\infty} \Delta^+(\{\alpha^1_{[x,X_2]}, \beta^1_{[x,X_2]}\}, \{\alpha^2, \beta^2\}; \mathbf{A}^1_m)\right] = +\infty.\end{aligned}$$

For the third term, notice that, for all $y \in \mathbb{N}$,

$$\Delta(\{\alpha^1_x, \beta^1_x\}, \{\alpha^2_y, \beta^2_y\}; \mathbf{A}^1_m) = -\Delta(\{\alpha^2_y, \beta^2_y\}, \{\alpha^1_x, \beta^1_x\}; \mathbf{A}^1_m).$$

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

$$\begin{aligned}&\lim_{x\to+\infty} \mathbb{E}\left[\Delta^-(\{\alpha^1_x, \beta^1_x\}, \{\alpha^2_{Y_2}, \beta^2_{Y_2}\}; \mathbf{A}^1_m)\right]\\ &\quad = -\lim_{x\to+\infty} \mathbb{E}\left[\Delta^+(\{\alpha^2_{Y_2}, \beta^2_{Y_2}\}, \{\alpha^1_x, \beta^1_x\}; \mathbf{A}^1_m)\right]\\ &\quad = -\mathbb{E}\left[\lim_{x\to+\infty} \Delta^+(\{\alpha^2_{Y_2}, \beta^2_{Y_2}\}, \{\alpha^1_x, \beta^1_x\}; \mathbf{A}^1_m)\right] = 0.\end{aligned}$$

□

Remark 4.4. In Proposition 4.3 condition (13) is a sufficient condition, and the discrete beta-Stacy process is defined such that condition (2) is verified. Both conditions are satisfied when their ratio diverges, that is when

$$\lim_{t\to\infty} \prod_{j=0}^{t-1} \left(1 + \frac{1}{\beta^1(j)}\right) \frac{\alpha^1\{j\} + \beta^1(j)}{\alpha^1\{j\} + \beta^1(j) + 1} = +\infty.$$

This constraint does not pose restrictions, and it is satisfied, as expected, in the special cases of the simple 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 stay-with-a-winner/switch-on-a-loser 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 4.5. *For all α^1, β^1 and α^2, β^2 as in Proposition 4.1, for all non-increasing discount sequences $\mathbf{A}_n$ and $n > 1$, there exists a break-even point $b\left(\{\alpha^1, \beta^1\}, \{\alpha^2, \beta^2\}; \mathbf{A}_n\right) \in \mathbb{N}$ such that*

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

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

$$\Delta\left(\{\alpha^1_x,\beta^1_x\},\{\alpha^2,\beta^2\};\mathbf{A}^1_n\right) < \Delta\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};\mathbf{A}_n\right)$$

if $x < b\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};\mathbf{A}_n\right)$.

Proof. From Proposition 4.1, $\Delta\left(\{\alpha^1_x,\beta^1_x\},\{\alpha^2,\beta^2\};\mathbf{A}^1_n\right)$ is non decreasing in x, starting from a value lower than $\Delta\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};\mathbf{A}_n\right)$ and going to infinity (Proposition 4.3). This is enough to claim that there exists a break-even point b which satisfies the properties in the theorem. □

Theorem 4.6. *For all* α^1,β^1 *and* α^2,β^2 *as in Proposition 4.1, for all nonincreasing discount sequences* $\mathbf{A}_n$ *and* $n>1$*, if the condition*

$$\Delta\left(\{\alpha^1_{x=0},\beta^1_{x=0}\},\{\alpha^2,\beta^2\};\mathbf{A}^1_n\right) \leq 0 \tag{15}$$

holds, there exists a break-even point $d\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};\mathbf{A}_n\right) \in \mathbb{N}$ *such that*

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

and

$$\Delta\left(\{\alpha^1_x,\beta^1_x\},\{\alpha^2,\beta^2\};\mathbf{A}^1_n\right) < 0 \quad \text{if } \ x < d\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};\mathbf{A}_n\right).$$

Proof. As in the proof of Theorem 4.5, there exists a point d satisfying the properties. □

Remark 4.7. Sufficient condition (15) arises because the support of the base measure of the beta-Stacy process is bounded below by zero. Both Clayton and Berry (1985) and Chattopadhyay (1994) notice that when the support is bounded, additional conditions at the boundaries are sufficient for the existence of break-even observations. In particular, the condition intuitively means that if a very bad observation from arm 1 is extracted at stage 1 (x close to 0), the alternative arm 2 is preferred under the current strategy. Note that in Theorem 4.5 it is superfluous a condition of the kind

$$\Delta\left(\{\alpha^1_{x=0},\beta^1_{x=0}\},\{\alpha^2,\beta^2\};\mathbf{A}^1_n\right) \leq \Delta\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};\mathbf{A}_n\right),$$

that is a condition that imposes a reduction in the expected advantage of arm 1 after the observation of $x=0$: the worst observation $x=0$ always causes a decrease in the expected advantage. On the other hand, without condition (15) in Theorem 4.6 we cannot exclude cases of prior values of α^i and β^i, $i=1,2$ such that $\mu_1 >> \mu_2$, with an expected advantage that does not change sign after the observation of $x=0$.

5. Bandit solution with continuous beta-Stacy processes

5.1. Framework setting

In the continuous beta-Stacy discrete-stage two-armed problem, X_k and Y_k, respectively from arm 1 and arm 2 at stage $k \in \{1,2,\ldots,n\}$, can assume values in $\mathbb{R}^+$ and α^1 and β^1 (and also α^2 and β^2) are, respectively, a continuous measure and a positive function, both defined on $\mathbb{R}^+$. α^1 and α^2 are assumed a priori to have no discontinuity points.

Recalling that we omit the dependence on α^1, α^2, β^1 and β^2, the results in Section 2.1 say that, for $t \in \mathbb{R}^+$,

$$\begin{aligned}\mathbb{P}(X_1 \leq t) &= 1 - \exp\left\{-\int_0^t \frac{d\alpha^1(s)}{\beta^1(s)}\right\} \\ &=: 1 - \prod_{[0,t]} \left(1 - \frac{d\alpha^1(s)}{\beta^1(s) + \alpha^1\{s\}}\right)\end{aligned}$$

where $\prod_{[0,t]}$ denotes the *product integral*, an operator commonly used in the survival analysis literature. For any partition $k_1 = z_0 < z_1 < \cdots < z_m = k_2$, if $l_m = \max_{i=1,\dots,m} |z_i - z_{i-1}|$, the product integral for a function $f : [k_1, k_2] \to \mathbb{R}^+$ is defined as

$$\prod_{[k_1,k_2]} \{1 + f(z)dz\} := \lim_{l_m \to 0} \prod_{i=1}^{m} \{1 + f(z_j) - f(z_{j-1})\},$$

where the limit is taken over all partitions of the interval $[k_1, k_2]$ with l_m approaching zero, for $k_1 < k_2$ both in $\mathbb{R}^+$. See Gill and Johansen (1990) for a survey of applications of product integrals to survival analysis. We can compute, in analogy with the discrete case,

$$\mathbb{E}[X_1] = \int_0^{+\infty} \mathbb{P}(X_1 > t)dt = \int_0^{+\infty} \prod_{[0,t]} \left(1 - \frac{d\alpha^1(s)}{\beta^1(s)}\right) dt =: \mu^1$$

and, similarly,

$$\mathbb{E}[Y_1] = \int_0^{+\infty} \prod_{[0,t]} \left(1 - \frac{d\alpha^2(s)}{\beta^2(s)}\right) dt =: \mu^2$$

assuming, without loss of generality, that $\mu^1 \leq \mu^2$.

For the stage $k \in \{1, 2, \dots, n\}$, from the results in Section 2.3,

$$\begin{aligned}\mathbb{P}(X_k \leq t|\mathbf{x}_{k-1}) &= 1 - \exp\left\{-\int_0^t \frac{d\alpha^1(s)}{\beta^1(s) + N_{\mathbf{x}_{k-1}}\{s\} + M_{\mathbf{x}_{k-1}}(s)}\right\} \\ &\quad \cdot \left(1 - \frac{N_{\mathbf{x}_{k-1}}\{t\}}{\beta^1(t) + N_{\mathbf{x}_{k-1}}\{s\} + M_{\mathbf{x}_{k-1}}(t)}\right) \\ &= 1 - \prod_{[0,t]} \left(1 - \frac{d\alpha^1_{\mathbf{x}_{k-1}}(s)}{\beta^1_{\mathbf{x}_{k-1}}(s) + \alpha^1_{\mathbf{x}_{k-1}}\{s\}}\right)\end{aligned}$$

and, partitioning $\mathbf{x}_{k-1} = [\mathbf{x}^{exact}_{k-1}, \mathbf{x}^{cens}_{k-1}]$ for respectively exact and censored observations, the posterior mean is

$$\mathbb{E}[X_k|\mathbf{x}_{k-1}] \quad = \quad \mathbb{P}(X_k \notin \mathbf{x}^{exact}_{k-1}|\mathbf{x}_{k-1}) \cdot \int_0^{+\infty} \mathbb{P}(X_k > t|X \notin \mathbf{x}^{exact}, \mathbf{x}_{k-1})dt +$$

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

5.2. Monotonicity of the expected advantage

The function $\Delta(\{\alpha^1, \beta^1\}, \{\alpha^2, \beta^2\}; \mathbf{A}_n)$ can be expressed as in (10). In the following propositions in the present and next subsections we will study its properties of monotonicity and continuity, with the aim of proving in Section 5.4 the existence of break-even observations of stay-with-a-winner and stay-with-a-winner/switch-on-a-loser strategies.

Proposition 5.1. *For all α^1, β^1 and α^2, β^2 such that $-\frac{\partial}{\partial t}\beta^1(t) \geq \frac{\partial}{\partial t}\alpha^1(t)$, $t \in \mathbb{R}^+$, and for all nonincreasing discount sequences $\mathbf{A}_n$,*

$$\Delta\left(\{\alpha^1_x, \beta^1_x\}, \{\alpha^2, \beta^2\}; \mathbf{A}_n\right)$$

is nondecreasing in x, for all $x \in \mathbb{R}^+$.

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

$$
\begin{aligned}
\Delta\left(\{\alpha^1_x, \beta^1_x\}, \{\alpha^2, \beta^2\}; \mathbf{A}_1\right) &= a_1(\mu^1_x - \mu^2) \\
&= a_1 \left(\int_0^{+\infty} \prod_{[0,t]} \left(1 - \frac{d\alpha^1_x(s)}{\beta^1_x(s) + N_x\{s\}} \right) dt - \mu^2 \right).
\end{aligned}
$$

We first show that μ^1_x is nondecreasing in x, for x being censored to the right. Notice that μ^1_x can be written as

$$
\begin{aligned}
\mu^1_x \; = \; & \int_0^{+\infty} \exp\left\{ - \int_0^t \frac{d\alpha^1(s)}{\beta^1_x(s) + N_x(s)} \right\} \\
& \left(1 - \frac{N_x\{t\}}{\beta^1_x(s) + N_x\{s\}} \right) dt \\
= \; & \int_0^x \exp\left\{ - \int_0^t \frac{d\alpha^1(s)}{\beta^1(s) + 1} \right\} dt \\
& + \int_x^{+\infty} \exp\left\{ - \left(\int_0^x \frac{d\alpha^1(s)}{\beta^1(s) + 1} + \int_x^t \frac{d\alpha^1(s)}{\beta^1(s)} \right) \right\} dt.
\end{aligned}
$$

The integrand in μ^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 μ^1_x. Take now any $x^* > x$, and separate the cases $t \leq x$, $t \in (x, x^*)$ and $t \geq x^*$:

- When $t \leq x$, the integrands in μ_x^1 and $\mu_{x^*}^1$ are the same.
- When $t \geq x^*$, the integrand in μ_x^1 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 $\mu_{x^*}^1$ 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 $\mu_{x^*}^1$ always greater than or equal to the one in μ_x^1.
- Finally, when $t \in (x, x^*)$, the integrands in μ_x^1 and $\mu_{x^*}^1$ 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 $\mu_{x^*}^1 \geq \mu_x^1$ and that the statement is true for $n = 1$. On the other hand, when x is not censored

$$\begin{aligned}
\mu_x^1 &= \mathbb{P}(X_2 \neq x|x)\mu^1 + \mathbb{P}(X_2 = x|x)x \\
&= \left(1 - \exp\left\{-\int_0^x \frac{d\alpha^1(s)}{\beta^1(s)+1}\right\}\frac{1}{\beta^1(x)+1}\right)\mu^1 \\
&\quad + \exp\left\{-\int_0^x \frac{d\alpha^1(s)}{\beta^1(s)+1}\right\}\frac{x}{\beta^1(x)+1},
\end{aligned}$$

and $\mu_{x^*}^1 \geq \mu_x^1$ for all $x^* > x$, if and only if $\mathbb{P}(X_2 = x|x)$, the probability of X_2 from arm 1 at stage 2 being equal to the previous exact observation, is nondecreasing in x. This condition is equivalent to $-\frac{\partial}{\partial t}\beta^1(t) \geq \frac{\partial}{\partial t}\alpha^1(t)$, for $t \in \mathbb{R}^+$, as required in the proposition.

By induction, assuming the monotonicity property for $n = m - 1$, with some natural number $m > 1$, the proof is completed along the lines of Proposition 4.1. □

Remark 5.2. 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 $\beta^1(t) \leq \beta^1(t+1) + \alpha^1\{t+1\}$, $t \in \mathbb{N}$, in Proposition 4.1 for the discrete beta-Stacy process, finds its continuous analogue $-\frac{\partial}{\partial t}\beta^1(t) \geq \frac{\partial}{\partial t}\alpha^1(t)$, $t \in \mathbb{R}^+$, in Proposition 5.1. As with the beta-Stacy bandit problem, the special cases of Dirichlet and simple homogeneous processes are included, and they correspond, respectively, to $-\frac{\partial}{\partial t}\beta^1(t) = \frac{\partial}{\partial t}\alpha^1(t)$ and to $\frac{\partial}{\partial t}\beta^1(t) = 0$.

5.3. Continuity of the expected advantage

Proposition 5.3. *For all α^1, β^1 and α^2, β^2 and all nonincreasing discount sequences $\mathbf{A}_n$, the expected advantage $\Delta\left(\{\alpha^1_x, \beta^1_x\}, \{\alpha^2, \beta^2\}; \mathbf{A}_n\right)$ is a continuous function of x, for $x \in \mathbb{R}^+$ censored to the right.*

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

$$\Delta\left(\{\alpha^1_x, \beta^1_x\}, \{\alpha^2, \beta^2\}; \mathbf{A}_n\right) \to \Delta\left(\{\alpha^1_{x_0}, \beta^1_{x_0}\}, \{\alpha^2, \beta^2\}; \mathbf{A}_n\right).$$

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

$$\Delta\left(\{\alpha^1_x, \beta^1_x\}, \{\alpha^2, \beta^2\}; \mathbf{A}_1\right) = a_1(\mu^1_x - \mu^2).$$

The continuity in x is shown through the continuity of μ^1_x. Taking any increasing sequence converging to x_0, then

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

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 μ^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

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

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 $h \in \mathbb{R}^+$. 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 $n = m-1$, and some natural number $m > 1$. By (12),

$$\begin{aligned}\Delta(\{\alpha^1_x,\beta^1_x\},\{\alpha^2,\beta^2\};\mathbf{A}_m) &= (a_1-a_2)(\mu^1_x-\mu^2) \\ &+ \mathbb{E}\left[\Delta^+(\{\alpha^1_{[x,X_2]},\beta^1_{[x,X_2]}\},\{\alpha^2,\beta^2\};\mathbf{A}^1_m)\right] \\ &+ \mathbb{E}\left[\Delta^-(\{\alpha^1_x,\beta^1_x\},\{\alpha^2_{Y_2},\beta^2_{Y_2}\};\mathbf{A}^1_m)\right].\end{aligned}$$

The first term $(a_1-a_2)(\mu^1_x-\mu^2)$ on the right hand side is continuous in x (from the continuity of μ^1_x). For the second term, note that

$$\Delta^+(\{\alpha^1_{[x,X_2]},\beta^1_{[x,X_2]}\},\{\alpha^2,\beta^2\};\mathbf{A}^1_m)$$

is a nondecreasing sequence in x (by Proposition 5.1), bounded below by 0 (by its definition) and convergent to $\Delta^+(\{\alpha^1_{[x_0,X_2]},\beta^1_{[x_0,X_2]}\},\{\alpha^2,\beta^2\};\mathbf{A}^1_m)$ (by the induction hypothesis). We can then apply the monotone convergence theorem:

$$\begin{aligned}&\lim_{x\to x_0}\mathbb{E}\left[\Delta^+(\{\alpha^1_{[x,X_2]},\beta^1_{[x,X_2]}\},\{\alpha^2,\beta^2\};\mathbf{A}^1_m)\right] \\ &= \mathbb{E}\left[\lim_{x\to x_0}\Delta^+(\{\alpha^1_{[x,X_2]},\beta^1_{[x,X_2]}\},\{\alpha^2,\beta^2\};\mathbf{A}^1_m)\right] \\ &= \mathbb{E}\left[\Delta^+(\{\alpha^1_{[x_0,X_2]},\beta^1_{[x_0,X_2]}\},\{\alpha^2,\beta^2\};\mathbf{A}^1_m)\right]\end{aligned}$$

For the third term, notice that, for $y \in \mathbb{R}^+$,

$$\Delta(\{\alpha^1_x,\beta^1_x\},\{\alpha^2_y,\beta^2_y\};\mathbf{A}^1_m) = -\Delta(\{\alpha^2_y,\beta^2_y\},\{\alpha^1_x,\beta^1_x\};\mathbf{A}^1_m).$$

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

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

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

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

By the dominated convergence theorem,

$$\begin{aligned}
&\lim_{x\to x_0} \mathbb{E}\left[\Delta^-(\{\alpha_x^1, \beta_x^1\}, \{\alpha_{Y_2}^2, \beta_{Y_2}^2\}; \mathbf{A}_m^1)\right] \\
&\quad = -\lim_{x\to x_0} \mathbb{E}\left[\Delta^+(\{\alpha_{Y_2}^2, \beta_{Y_2}^2\}, \{\alpha_x^1, \beta_x^1\}; \mathbf{A}_m^1)\right] \\
&\quad = -\mathbb{E}\left[\lim_{x\to x_0} \Delta^+(\{\alpha_{Y_2}^2, \beta_{Y_2}^2\}, \{\alpha_x^1, \beta_x^1\}; \mathbf{A}_m^1)\right] \\
&\quad = \mathbb{E}\left[\Delta^-(\{\alpha_{x_0}^1, \beta_{x_0}^1\}, \{\alpha_{Y_2}^2, \beta_{Y_2}^2\}; \mathbf{A}_m^1)\right],
\end{aligned}$$

proving continuity for the generic bandit horizon n. □

5.4. Existence of break-even points

Proposition 5.4. *For all α^1, β^1 and α^2, β^2 as in Proposition 5.1, for all $x \in \mathbb{R}^+$ and all nonincreasing discount sequences $\mathbf{A}_n$,*

$$\Delta\left(\{\alpha_{x=0}^1, \beta_{x=0}^1\}, \{\alpha^2, \beta^2\}; \mathbf{A}_n\right) = \inf_x \Delta\left(\{\alpha_x^1, \beta_x^1\}, \{\alpha^2, \beta^2\}; \mathbf{A}_n\right).$$

Furthermore, if $\int_0^\infty d\alpha^1(t)/(\beta^1(t)+1) < \infty$, then

$$\lim_{x\to+\infty} \Delta\left(\{\alpha_x^1, \beta_x^1\}, \{\alpha^2, \beta^2\}; \mathbf{A}_n\right) = \infty.$$

Proof. The case for $x = 0$ is an immediate consequence of Proposition 5.1. To study the case where x diverges, we proceed by induction. Note that for $n = 1$, $\Delta\left(\{\alpha_x^1, \beta_x^1\}, \{\alpha^2, \beta^2\}; \mathbf{A}_1\right) = a_1(\mu_x^1 - \mu^2)$ and

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

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

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

This proves that $\lim_{x\to+\infty} \Delta\left(\{\alpha_x^1, \beta_x^1\}, \{\alpha^2, \beta^2\}; \mathbf{A}_1\right) = \infty$. The rest of the proof follows the same lines as the proof of Proposition 4.3. □

Remark 5.5. Coherently with the conditions required in Proposition 4.3 for the discrete beta-Stacy bandit problem, in the above proposition the additional condition $\int_0^\infty d\alpha^1(t)/(\beta^1(t)+1) < \infty$ is a sufficient condition. Note that the beta-Stacy process is defined such that $\int_0^\infty d\alpha^1(t)/\beta^1(t) = \infty$. 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

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

For finite β^1, this is satisfied, and, as expected, includes the special cases of the simple homogeneous process and the Dirichlet process. In short, the additional constraint rules out cases of exploding β^1. Usually, β^1 is fixed such that $\beta^1(t) = M\cdot F_0[t,\infty)$, converging to 0 as t diverges (see Walker and Muliere 1997).

Theorem 5.6. *For all α^1,β^1, α^2,β^2 as in Proposition 5.4, for all nonincreasing discount sequences $\mathbf{A}_n$ and $n>1$, there exists a break-even point $b\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};\mathbf{A}_n\right)\in\mathbb{R}^+$ such that*

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

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

$$\Delta\left(\{\alpha^1_x,\beta^1_x\},\{\alpha^2,\beta^2\};\mathbf{A}^1_n\right)<\Delta\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};\mathbf{A}_n\right)$$

if $x<b\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};\mathbf{A}_n\right)$.

Proof. From Propositions 5.1 and 5.3, $\Delta\left(\{\alpha^1_x,\beta^1_x\},\{\alpha^2,\beta^2\};\mathbf{A}^1_n\right)$ is nondecreasing in x and continuous (the latter only for x censored), starting from a value lower than

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

and growing to infinity (Proposition 5.4). Then the point $b(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};\mathbf{A}_n)$ exists and satisfies the properties of the theorem. □

Theorem 5.7. *For all $\alpha^1,\beta^1,\alpha^2,\beta^2$ as in Proposition 5.4, for all nonincreasing discount sequences $\mathbf{A}_n$ and $n>1$, if condition (15) holds, there exists a break-even point $d\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};\mathbf{A}_n\right)\in\mathbb{R}^+$ such that*

$$\Delta\left(\{\alpha^1_x,\beta^1_x\},\{\alpha^2,\beta^2\};\mathbf{A}^1_n\right)\geq 0\qquad \textit{if}\quad x\geq d\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};\mathbf{A}_n\right)$$

and

$$\Delta\left(\{\alpha^1_x,\beta^1_x\},\{\alpha^2,\beta^2\};\mathbf{A}^1_n\right)<0\qquad \textit{if}\quad x<d\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};\mathbf{A}_n\right).$$

Proof. As in Theorem 5.6, there exists a point $d\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};\mathbf{A}_n\right)$ satisfying the properties. □

Remark 5.8. For Theorems 5.6 and 5.7, Remark 4.7 is still valid, on the sufficiency of an additional boundary condition (in Theorem 5.7, but not in 5.6) for finding break-even points in bandit problems with base measures having bounded supports.

6. Applications

6.1. Discrete beta-Stacy bandit examples

Consider the discrete beta-Stacy two-armed bandit problem, with $M^i = \alpha^i(\mathbb{N})$, $i = 1, 2$ be the total masses of the measures α^1 and α^2. A higher value of M^i is interpreted as a stronger prior knowledge of the beta-Stacy process related to arm i. We observe censored to the right observations, fix the bandit horizon to $n = 3$ and the discount sequence is $\mathbf{A}_3 = (1, 0.9, 0.8)$. Choosing higher values for n is feasible and to higher values correspond higher processing times. Denoting by $X_k^{(l)}$ (respectively $Y_k^{(l)}$) the l-th sampled extraction from arm 1 (arm 2) at stage k, we first sample $X_1^{(l)}$ and $Y_1^{(l)}$ from the two arms, for $l = 1, \dots, T$ with $T = 100$. Then for each $X_1^{(l)}$ we sample T times X_2 from the first arm, and for each $Y_1^{(l)}$ we sample T times 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.

6.1.1. Discrete numerical example 1

Fix, for all $t \in \mathbb{N} \setminus \{0\}$, $\alpha^1\{t\} = 0.1M^1 \cdot 0.9^{t-1}$ and $\beta^1(t) = M^1 \cdot 0.9^t$ for the first arm, and $\alpha^2\{t\} = 0.08M^2 \cdot 0.92^{t-1}$ and $\beta^2(t) = M^2 \cdot 0.92^t$ and for the second arm. For $t = 0$, $i = 1, 2$, $\alpha^i\{t\} = \beta^i(t) = 0$, and for all $t \notin \mathbb{N}$, $i = 1, 2$, $\alpha^i\{t\} = 0$ and $\beta^i(t) = \beta^i(\lfloor t \rfloor)$, where $\lfloor t \rfloor$ is the largest integer lower or equal to t. Note that $\mu^1 < \mu^2$ a priori and that different values of M^1 and M^2 do not affect the prior means, μ^1 and μ^2, but only the posterior means, since

$$\begin{aligned} \mu^1 &= \sum_{t=1}^{\infty} \prod_{j=0}^{t-1} \frac{\beta^1\{j\}}{\alpha^1\{j\} + \beta^1(j)} \\ &= \sum_{t=1}^{\infty} \prod_{j=0}^{t-1} \frac{M^1 \cdot 0.9^j}{0.1M^1 \cdot 0.9^{j-1} + M^1 \cdot 0.9^j} = \sum_{t=1}^{\infty} 0.9^t = 9, \end{aligned}$$

and similar calculations show that $\mu^2 = 11.5$. The assumption of Proposition 4.1 is satisfied, since for all $t \in \mathbb{N}\setminus\{0\}$ we have $\beta^1(t) = \beta^1(t+1) + \alpha^1\{t+1\} = M^1 \cdot 0.9^t$ and $\beta^2(t) = \beta^2(t+1) + \alpha^2\{t+1\} = M^2 \cdot 0.92^t$, for $t = 0$ and $i = 1, 2$ we have $\beta^i(t) = 0 < M^i = \beta^i(t+1) + \alpha^i\{t+1\}$, and for $t \notin \mathbb{N}$ and $i = 1, 2$ we have $\beta^i(t) = \beta^i(t+1) + \alpha^i\{t+1\} = 0$.

For each scenario, we evaluate $\mu^1_{\mathbf{x}_1}$, $\mu^1_{\mathbf{x}_2}$, $\mu^2_{\mathbf{y}_1}$ and $\mu^2_{\mathbf{y}_2}$; we then evaluate $\Delta\left(\{\alpha^1, \beta^1\}, \{\alpha^2, \beta^2\}; \mathbf{A}_3\right)$, reported in Table 1 for different values of M^1 and M^2. There is a tendency for $\Delta\left(\{\alpha^1, \beta^1\}, \{\alpha^2, \beta^2\}; \mathbf{A}_3\right)$ to increase in M^2 and decrease in M^1, holding everything else constant. 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

Table 1

Estimated $\Delta\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};\mathbf{A}_3\right)$*, with parameters as specified in Section 6.1.1, 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

arm, since more information can be gained from its exploration: when M^1 increases, more weight is given to the prior belief of arm 1, that is then considered better known, with consequent higher tendency of exploring arm 2, as reflected in the lower expected advantage of arm 1 over arm 2. A specular reasoning on M^2 leads to a higher expected advantage of arm 1 over arm 2 as M^2 increases. Furthermore, as both M^1 and M^2 increase, prior information on both arms assumes more relevance, relative to the observation coming from the observation of the arms, up to the case where $\Delta\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};\mathbf{A}_3\right)$ approaches $\mu^1-\mu^2=-2.5$ (the prior mean difference), with no impact of the observations on the choice of the arms to observe. When $\Delta\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};\mathbf{A}_3\right)$ 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 $M^1<<M^2$: these are situations where the higher prior uncertainty (lower M^1) of the first arm, relative to the higher prior confidence in the second arm (larger M^2), makes the first arm preferable to be explored, even if a priori arm 2 is believed to be better.

6.1.2. Discrete numerical example 2

The beta-Stacy parameters in the two-armed bandit problem are fixed, for $i=1,2$ and $t\in\mathbb{N}$, as

$$\alpha^i\{t\} \;=\; M^i\frac{1}{2h^i+1}\mathbb{1}_{\{c^i-h^i,\dots,c^i+h^i\}}(t), \tag{16}$$

$$\beta^i(t) \;=\; M^i\left(\frac{h^i+c^i-t}{2h^i+1}\mathbb{1}_{\{c^i-h^i,\dots,c^i+h^i\}}(t)+\mathbb{1}_{[0,c^i-h^i)}(t)\right), \tag{17}$$

where $c^i,h^i\in\mathbb{N}$ and $h^i<c^i$. For all $t\notin\mathbb{N}$, $i=1,2$, $\alpha^i\{t\}=0$ and $\beta^i(t)=\beta^i(\lfloor t\rfloor)$. Note that, for $i=1,2$,

$$\mu^i \;=\; \sum_{t=1}^{c^i+h^i}\prod_{j=c^i-h^i}^{t-1}\frac{\beta^i\{j\}}{\alpha^i\{j\}+\beta^i(j)}=c^i-h^i-1+\sum_{t=c^i-h^i}^{c^i+h^i}\frac{h^i+c^i-t+1}{2h^i+1}=c^i,$$

and therefore we can fix the prior means through c^1 and c^2. The assumption of Proposition 4.1 is satisfied, since for $i=1,2$ and $t\le c^i-h^i-1$ we have $\beta^i(t)=$

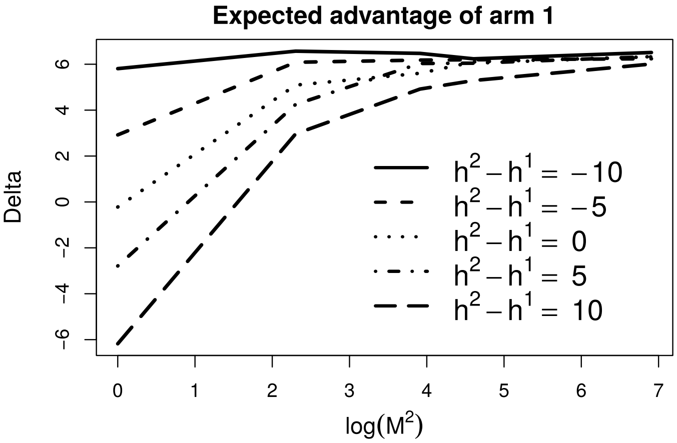

FIG 1. *Estimated* $\Delta\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};\mathbf{A}_3\right)$*, with parameters as specified in equations* (16) *and* (17)*, for different variability around the base measure (*M^2*) and different base measure variability (*h^2*) of the beta-Stacy process from the second arm. The prior means are both equal to* $\mu^1 = \mu^2 = 20$*, and* $M^1 = h^1 = 1$*.*

$\beta^i(t+1)+\alpha^i\{t+1\} = M^i$, for $t \geq c^i + h^i$ we have $\beta^i(t) = \beta^i(t+1)+\alpha^i\{t+1\} = 0$, and for $t \in \{c^i - h^i, \dots, c^i + h^i - 1\}$, we have $\beta^i(t) = \beta^i(t+1) + \alpha^i\{t+1\} = M^i(h^i + c^i - t)/(2h^i + 1)$.

The parameter h^i is positively related to the variability of the base measure of the beta-Stacy process related to arm i, whilst M^i 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 arm 1 over the other arm is affected by a change in h^2 and in M^2. In this way we set up an experiment in which we can isolate the effect on the expected advantage of arm 1 of a change in the prior variability of the beta-Stacy base measure related to arm 2 (a change in h^2) from the effect of a change in the prior belief in the base measure (a change in M^2). The prior means are fixed equal to avoid the results being affected by a dominant prior mean. For instance, holding fixed M^2, an increase in h^2 leaves unaltered the prior mean of arm 2, but the support of α^2 is more spread, with a consequent increase in the variability of responses from arm 2 and a more convenient exploration of arm 2.

In Figure 1 we report the value of $\Delta\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};\mathbf{A}_3\right)$ for different h^2 and M^2. The dotted line, corresponding to the case $h^1 - h^2 = 0$, 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 $\Delta\left(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};\mathbf{A}_3\right) \approx 6$ for $h^2 - h^1 = -10$. Viceversa, higher positive values of $h^2 - h^1$ correspond to higher preference for arm 2. Furthermore, in the considered setting the effect of a change in M^2 seems to dominate when it becomes very large: 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.

6.2. Continuous beta-Stacy bandit numerical example

We adopt in the two-armed bandit problem the beta-Stacy process suggested by the numerical example in Ferguson and Phadia (1979) and Walker and Muliere (1997): for the first arm we choose $d\alpha^1(t) = \exp(-t/10)/10dt$ and $\beta^1(t) = \exp(-t/10)$, whist for the second arm $d\alpha^2(t) = \exp(-t/12)/12dt$ and $\beta^2(t) = \exp(-t/12)$, for $t \in \mathbb{R}^+$. Note that $\mu^1 < \mu^2$ a priori since

$$\mu^1 \quad = \quad \int_0^\infty \exp\left\{-\int_0^t \frac{d\alpha^1(s)}{\beta^1(s)}\right\} dt = \int_0^\infty \exp\left\{-\int_0^t \frac{e^{-s/10}}{10e^{-s/10}}\right\} dt = 10,$$

and similar calculations show that $\mu^2 = 12$. The assumption in Proposition 5.1 is satisfied, since for all $t \in \mathbb{R}^+$, $-\frac{\partial}{\partial t}\beta^1(t) = \frac{\partial}{\partial t}\alpha^1(t) = \exp\{-t/10\}/10$ and $-\frac{\partial}{\partial t}\beta^2(t) = \frac{\partial}{\partial t}\alpha^2(t) = \exp\{-t/12\}/12$. All the rest is fixed as in the previous examples. Denoting by $X_k^{(l)}$ (respectively $Y_k^{(l)}$) the l-th sampled extraction from arm 1 (arm 2) at stage k, we sample $X_1^{(l)}$ and $Y_1^{(l)}$ from the two arms, for $l = 1, \ldots, T$ and $T = 150$. Then for each $X_1^{(l)}$ we sample T times X_2 from the first arm, and for each $Y_1^{(l)}$ we sample T times Y_2 from the second arm, using Algorithms 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 $\mu^1_{\mathbf{x}_1}$, $\mu^1_{\mathbf{x}_2}$, $\mu^2_{\mathbf{y}_1}$, $\mu^2_{\mathbf{y}_2}$; $\Delta\left(\{\alpha^1, \beta^1\}, \{\alpha^2, \beta^2\}; \mathbf{A}_3\right)$ is then evaluated and $\Delta\left(\{\alpha^1_x, \beta^1_x\}, \{\alpha^2, \beta^2\}; \mathbf{A}^1_3\right)$, for $x \in \mathbb{R}^+$. In the top-right plot of Figure 2 we show the expected advantage of arm 1 over arm 2 as a function of the observed x at stage 1, x from arm 1 and exact. Monotonicity in x of $\Delta\left(\{\alpha^1_x, \beta^1_x\}, \{\alpha^2, \beta^2\}; \mathbf{A}^1_3\right)$ is numerically verified. Note also that condition (15) is satisfied, so that the two break-even points exist. In particular, the break-even observation for the stay-with-a-winner strategy is $b\left(\{\alpha^1, \beta^1\}, \{\alpha^2, \beta^2\}; \mathbf{A}_3\right) = 15.27$, whilst for the stay-with-a-winner/switch-on-a-loser strategy is $d\left(\{\alpha^1, \beta^1\}, \{\alpha^2, \beta^2\}; \mathbf{A}_3\right) = 18.89$. The strategies can be completely determined. For instance, arm 2 is optimally selected at stage 1 by both stay-with-a-winner and stay-with-a-winner/switch-on-a-loser strategies, since $\Delta\left(\{\alpha^1, \beta^1\}, \{\alpha^2, \beta^2\}; \mathbf{A}_3\right) < 0$. If an exact observation from arm 2 is extracted equal, say, to $y_1 = 4$, the stay-with-a-winner strategy chooses arm 1 at stage 2 since $\Delta\left(\{\alpha^1, \beta^1\}, \{\alpha^2_{y_1}, \beta^2_{y_1}\}; \mathbf{A}^1_3\right) = -1.60$, greater than

$$\Delta\left(\{\alpha^1, \beta^1\}, \{\alpha^2, \beta^2\}; \mathbf{A}_3\right) = -2.47.$$

Since $\Delta\left(\{\alpha^1, \beta^1\}, \{\alpha^2_{y_1}, \beta^2_{y_1}\}; \mathbf{A}^1_3\right) < 0$, at stage 2 arm 1 would not be chosen by the stay-with-a-winner/switch-on-a-loser strategy, but it will choose again the optimal arm 2. Following the stay-with-a-winner strategy, suppose now a censored observation from arm 1 equal, say, to $x_2 = 15.5$ is observed, for which $\Delta\left(\{\alpha^1_{x_2}, \beta^1_{x_2}\}, \{\alpha^2_{y_1}, \beta^2_{y_1}\}; \mathbf{A}^2_3\right) = -1$, greater than $\Delta\left(\{\alpha^1, \beta^1\}, \{\alpha^2_{y_1}, \beta^2_{y_1}\}; \mathbf{A}^1_3\right)$. Therefore in the third stage the observation of arm 1 is again dictated by the stay-with-a-winner strategy.

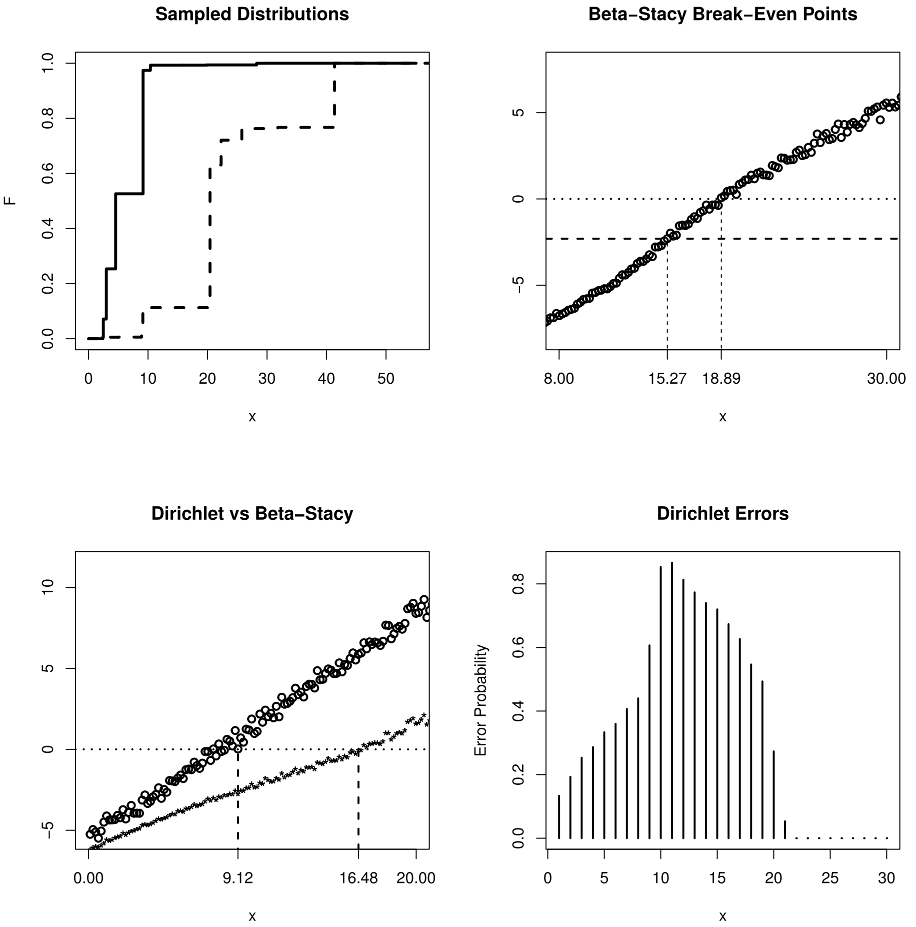

FIG 2. *Top-left: Two distributions sampled from the beta-Stacy process of Walker and Muliere (1997), with algorithm A in Al Labadi and Zarepour (2013). The solid line is from arm 1, the dashed one from arm 2, with parameters as specified in Section 6.2. Top-right: in circles we report* $\Delta\left(\{\alpha^1_x, \beta^1_x\}, \{\alpha^2, \beta^2\}; \mathbf{A}^1_3\right)$*, the dashed horizontal line is* $\Delta\left(\{\alpha^1, \beta^1\}, \{\alpha^2, \beta^2\}; \mathbf{A}_3\right)$*. The horizontal line at 0 is also highlighted with a dotted line. The intersections determine the break-even observations of the two strategies outlined in the text. Bottom-left: in circles* $\Delta\left(\{\alpha^1_x, \beta^1_x\}, \{\alpha^2, \beta^2\}; \mathbf{A}^1_3\right)$ *for the beta-Stacy bandit problem. With the asterisks we represent the corresponding quantity for the Dirichlet bandit that ignores the censorship. The horizontal line at 0 is also highlighted with a dotted line. Bottom-right: Error probability of the stay-with-a winner strategy implied by the Dirichlet bandit problem, when censorship is ignored, as function of the first right-censored observation from arm 1.*

In the bottom-left plot of Figure 2 we report $\Delta\left(\{\alpha^1_x, \beta^1_x\}, \{\alpha^2, \beta^2\}; \mathbf{A}^1_3\right)$ when $x \in \mathbb{R}^+$ 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 and highlight the relevance of properly accounting for right-censored data. There is a range of values from 9.12 to 16.48 at which the Dirichlet bandit problem would take the wrong strategy, since $\Delta\left(\{\alpha_x^1,\beta_x^1\},\{\alpha^2,\beta^2\};\mathbf{A}_3^1\right)$ would be of opposite sign, relative to the corresponding beta-Stacy quantity. The break-even point 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 future stages. If we repeat the experiment 150 times for each value of x from 1 to 30, we can compute the probability the Dirichlet bandit being in error in the choice of the optimal arm, after the observation of x from arm 1 at stage 1. This probability is reported in the bottom-right plot of Figure 2.

7. 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 (Walker and Muliere 1997). The proposed framework extends the one-armed and two-armed Dirichlet bandit problem of Clayton and Berry (1985) and Chattopadhyay (1994) since the beta-Stacy process reduces 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, under non-restrictive constraints on the process parameters, the existence of stay-with-a-winner and stay-with-a-winner/switch-on-a-loser break-even points that partially characterize optimal strategies.

7.1. Relation to delayed responses

A stream of literature close to the proposed setting is that on bandit problems with delayed responses: new subjects arrive in the bandit problem and optimal assigment to arms has to be performed before having observed the responses of past subjects. Delayed bandits have been first proposed by Eick (1985, 1988a) for the one-armed and two-armed bandit and by Eick (1988b) for the multi-armed bandit problem. Among later significant developements, Hardwick, Oehmke and Stout (2001, 2006) proposed the two-armed Bernoulli bandit problem with subjects arriving according to a Poisson process and exponential responses; Wang (2000) studied boundary and monotonicity properties of the break-even point and finite-stage optimal stopping of the Eick (1988a) bandit problem; the one-armed Eick (1988a) with continuous stages was introduced by Wang and Bickis (2003); Caro and Yoo (2010) proved that discrete-stage bandit problems with stationary random delays satisfy the indexability criterion (Whittle 1988) as long as the delayed responses do not cross over; a computationally efficient approximation to the solution of a multi-armed bandit problem with delayed responsed was implemented in Guha, Munagala and Pál (2013).

The relation between censored and delayed responses lays in the fact that past responses not yet observed in delayed bandits are typically treated as censored observations. But there is a significant difference that prevents the approach of the present paper fitting into the framework of delayed responses: in delayed bandits the censored observations are potentially exact observations with exact value not yet realized, whilst in our setting a censored observation will never become exact. In the patients' treatment example, the censored observation in the delayed bandit is the survival time returned at the current stage by a treated patient not died yet, and this piece of data will become exact at the stage at which the patient will die. On the contrary, we are not able to treat delayed responses since the time index of the beta-Stacy process driving the generation of the response is detached from the stage index of the bandit problem: the new bandit stage begins (the new subject arrives) only when past responses (exact or censored) are given. The extension of our setting to the case with delayed responses is an interesting research question beyond the scope of the present paper, but similarities between the two approaches give first suggestions on how this extension can be performed: from Walker and Muliere (1997), the beta-Stacy process can be expressed in a product form that resembles and extends the geometric distribution of the delayed responses in Eick (1985, 1988a).

7.2. Multi and one-armed bandits

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, Bulla and Walker (2007). In this direction, Battiston, Favaro and Teh (2016) adopt hierarchical Poisson-Dirichlet processes in multi-armed bandit problems. The introduction of dependence between the arms also suggests the extension to restless bandit problems, in which the parameters of the beta-Stacy process of one arm are updated even if no response from that arm is observed, but as a consequence of the observation of a response from a dependent arm.

We highlight that in the setting of the current paper the parameters associated to the beta-Stacy process of arm i, $i = 1, 2$, are updated according to the rules given in Section 2.3 only when a new response is observed from arm i, ruling out from the present setting restless bandits (Whittle 1988). This feature, together with the independence of the two arms, makes the current setting (and the one with a generic number of arms) a classic bandit problem that, in the special case of a discount sequence that is geometric at least up to a constant of proportionality and common to all arms, is solvable in priciple by a Gittins index policy (Gittins 1979). The feasibility of the Gittins index calculation associated to each beta-Stacy arm is not trivial and is object of current investigation from the authors in the aim of generalizing the proposed setting to a multi-armed framework. We conjecture that one of the methods surveyed in Chakravorty and Mahajan (2014) could be adopted, but an effort is needed to derive the transition probability matrix corresponding to the Markovian update of the bandit process.

On the other hand, the present framework is reduced to a one-arm beta-Stacy bandit problem if we let the total measure $M^2 := \alpha^2(\mathbb{R}^+)$ diverge to $+\infty$. The value of M^2 gives indication on the strength of the prior belief in the base measure of the beta-Stacy process associated to arm 2. The extreme case of an infinite M^2 means *de facto* a sure knowledge of the distribution driving the generation of the responses from arm 2. The result would be an arm 1 whose responses are driven by a beta-Stacy process whose parameters are updated according to the specified rules as new observations from arm 1 are collected; and an arm 2 with known distribution equal to the mean distribution of the beta-Stacy process of arm 2, whose parameters are never updated, regardless of the responses observed from arm 2. All subsequent results would remain the same, with the exception that the mean response of arm 2, affecting the expected bandit payoff, would not change from stage to stage, but it would remain fixed to its prior value.

7.3. Other directions of investigation

Our framework can be further extended to different bandit 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, Paganoni and Secchi 2006); epsilon-decreasing and VBDE strategies (Cesa-Bianchi and Fisher 1998; Tokic 2010) would require a beta-Stacy parameter update mechanism dependent on the number of steps or on the values extracted from the arms. Third, 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 measure of the beta-Stacy processes can change after some stage.

Appendix

Following the notation introduced in Section 3.2,

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

$$
\begin{aligned}
&W^2(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};\mathbf{A}_n)\\
&\quad= W(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};\mathbf{A}_n) - \Delta^+(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};\mathbf{A}_n).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&W^1(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};\mathbf{A}_n)\\
&\quad= a_1\mu^1 + \mathbb{E}\left[W(\{\alpha^1_{X_1},\beta^1_{X_1}\},\{\alpha^2,\beta^2\};\mathbf{A}^1_n)\right]\\
&\quad= a_1\mu^1 + \mathbb{E}\left[W^2(\{\alpha^1_{X_1},\beta^1_{X_1}\},\{\alpha^2,\beta^2\};\mathbf{A}^1_n)\right]\\
&\qquad+ \mathbb{E}\left[\Delta^+(\{\alpha^1_{X_1},\beta^1_{X_1}\},\{\alpha^2,\beta^2\};\mathbf{A}^1_n)\right],\\
&W^2(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};\mathbf{A}_n)\\
&\quad= a_1\mu^2 + \mathbb{E}\left[W(\{\alpha^1,\beta^1\},\{\alpha^2_{Y_1},\beta^2_{Y_1}\};\mathbf{A}^1_n)\right]\\
&\quad= a_1\mu^2 + \mathbb{E}\left[W^1(\{\alpha^1,\beta^1\},\{\alpha^2_{Y_1},\beta^2_{Y_1}\};\mathbf{A}^1_n)\right]\\
&\qquad- \mathbb{E}\left[\Delta^-(\{\alpha^1,\beta^1\},\{\alpha^2_{Y_1},\beta^2_{Y_1}\};\mathbf{A}^1_n)\right],
\end{aligned}
$$

and

$$
\begin{aligned}
&\Delta(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};\mathbf{A}_n)\\
&\quad= W^1(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};\mathbf{A}_n) - W^2(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};\mathbf{A}_n)\\
&\quad= a_1\mu^1 + \mathbb{E}\left[W^2(\{\alpha^1_{X_1},\beta^1_{X_1}\},\{\alpha^2,\beta^2\};\mathbf{A}^1_n)\right]\\
&\qquad- a_1\mu^2 - \mathbb{E}\left[W^1(\{\alpha^1,\beta^1\},\{\alpha^2_{Y_2},\beta^2_{Y_2}\};\mathbf{A}^1_n)\right]\\
&\qquad+ \mathbb{E}\left[\Delta^+(\{\alpha^1_{X_1},\beta^1_{X_1}\},\{\alpha^2,\beta^2\};\mathbf{A}^1_n)\right]\\
&\qquad+ \mathbb{E}\left[\Delta^-(\{\alpha^1,\beta^1\},\{\alpha^2_{Y_1},\beta^2_{Y_1}\};\mathbf{A}^1_n)\right].
\end{aligned}
$$

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

$$a_1\mu^1 + \mathbb{E}\left[W^2(\{\alpha^1_{X_1},\beta^1_{X_1}\},\{\alpha^2,\beta^2\};\mathbf{A}^1_n)\right]$$

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

$$a_1\mu^2 + \mathbb{E}\left[W^1(\{\alpha^1,\beta^1\},\{\alpha^2_{Y_1},\beta^2_{Y_1}\};\mathbf{A}^1_n)\right]$$

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,

$$
\begin{aligned}
\Delta(\{\alpha^1,\beta^1\},\{\alpha^2,\beta^2\};\mathbf{A}_n) \quad = \quad & (a_1-a_2)(\mu^1-\mu^2)\\
&+ \mathbb{E}\left[\Delta^+(\{\alpha^1_{X_1},\beta^1_{X_1}\},\{\alpha^2,\beta^2\};\mathbf{A}^1_n)\right]\\
&+ \mathbb{E}\left[\Delta^-(\{\alpha^1,\beta^1\},\{\alpha^2_{Y_1},\beta^2_{Y_1}\};\mathbf{A}^1_n)\right]
\end{aligned}
$$

Acknowledgments

We thank three anonymous referees, the Editor and the Associate Editor, for their detailed and insightful comments that contributed significanty to improve the paper. We also thank Andrea Arfè, Bruno Buonaguidi and Antonio Lijoi for their precious suggestions.

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. MR3101595

Antoniak, C. (1974). Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *Annals of Statistics* **2** 1152-1174. MR0365969

Battiston, M., Favaro, S. and Teh, Y. W. (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. MR0079386

Berry, D. A. (1972). A Bernoulli Two-Armed Bandit. *The Annals of Mathematical Statistics* **43** 871-897. MR0305531

Berry, D. A. and Fristedt, B. (1979). Bernoulli One-Armed Bandits - Arbitrary Discount Sequences. *The Annals of Statistics* **7** 1086-1105. MR0536512

Berry, D. A. and Fristedt, B. (1985). *Bandit Problems: Sequential Allocation of Experiments*. Chapman and Hall, New York. MR0813698

Bradt, R. N., Johnson, S. M. and Karlin, S. (1956). On Sequential Designs for Maximizing the Sum of n Observations. *The Annals of Mathematical Statistics* **33** 847-856. MR0087288

Caro, F. and Yoo, O. S. (2010). Indexability of Bandit Problems with Response Delays. *Probability in the Engineering and Informational Sciences* **24** 349-374. MR2653796

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* 100-108.

Chakravorty, J. and Mahajan, A. (2014). Multi-armed bandits, Gittins index, and its calculation. In *Methods and Applications of Statistics in Clinical Trials: Planning, Analysis, and Inferential Methods, Volume 2* 416-435. MR3287464

Chattopadhyay, M. K. (1994). Two-Armed Dirichlet Bandits With Discounting. *The Annals of Statistics* **22** 1212-1221. MR1311973

Chernoff, H. (1968). Optimal Stochastic Control. *Sankhya* **30** 221-252. MR0241149

Clayton, M. K. and Berry, D. A. (1985). Bayesian Nonparametric Bandits. *The Annals of Statistics* **13** 1523-1534. MR0811507

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* 1814-1819.

de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré* **7** 1-68. MR1508036

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. MR2894871

Doksum, K. A. (1974). Tailfree and neutral random probabilities and their posterior distributions. *Annals of Probability* **2** 183-201. MR0373081

Eick, S. G. (1985). Two-armed bandits with delayed responses. University of Minnesota Statistics Technical Report 456.

Eick, S. G. (1988a). The two-armed bandit with delayed responses. *The Annals of Statistics* **16** 254-264. MR0924869

Eick, S. G. (1988b). Gittins procedures for bandits with delayed responses. *Journal of the Royal Statistical Society, Series B* **50** 125-132. MR0954739

Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics* **1** 209-230. MR0350949

Ferguson, T. S. and Phadia, E. G. (1979). Bayesian Nonparametric Estimation Based on Censored Data. *The Annals of Statistics* **7** 163-186. MR0515691

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. D. and Johansen, S. (1990). A Survey of Product Integration with a View Toward Application in Survival Analysis. *The Annals of Statistics* **18** 1501-1555. MR1074422

Gittins, J. C. (1979). Bandit Processes and Dynamic Allocation Indices (with discussion). *Journal of the Royal Statistical Society, Series B* **41** 148-177. MR0547241

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. MR0996417

Guha, S., Munagala, K. and Pál, M. (2013). Multi-armed bandit problems with delayed feedback. Available at https://arxiv.org/abs/1306.3525.

Hardwick, J., Oehmke, R. and Stout, Q. F. (1998). Adaptive allocation in the presence of missing outcomes. *Computing Science and Statistics* **30** 219-223.

Hardwick, J., Oehmke, R. and Stout, Q. F. (2001). Optimal adaptive designs for delayed response models: exponential case. In *MODA6: Model Oriented Data Analysis* 127-134. MR1865152

Hardwick, J., Oehmke, R. and Stout, Q. F. (2006). New adaptive designs for delayed response models. *Journal of Statistical Planning and Inference* **136** 1940-1955. MR2255605

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

Muliere, P., Bulla, P. and Walker, S. (2007). Bayesian Nonparametric Estimation of Bivariate Survival Function. *Statistica Sinica* **17** 427-444. MR2398429

Muliere, P., Paganoni, A. M. and Secchi, P. (2006). A randomly reinforced urn. *Journal of Statistical Planning and Inference* **136** 1853-1874. MR2255601

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

Phadia, E. G. (2013). *Prior Processes and Their Applications.* Springer-Verlag, Berlin. MR3087744

Robbins, H. (1952). Some Aspects of the Sequential Design of Experiments. *Bullettin of American Mathematical Society* **58** 527-535. MR0050246

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. MR0436445

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

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* 203-210.

Walker, S. and Muliere, P. (1997). Beta-Stacy Processes and a Generalization of the Pólya-Urn Scheme. *The Annals of Statistics* **25** 1762-1780. MR1463574

Walker, S. and Muliere, P. (2003). A Bivariate Dirichlet Process. *Statistics and Probability Letters* **64** 1-7. MR1995803

Wang, X. (2000). A bandit process with delayed responses. *Statistics & Probability Letters* **48** 303-307. MR1765756

Wang, X. and Bickis, M. G. (2003). One-armed bandit models with continuous and delayed responses. *Mathematical Methods of Operations Research* **58** 209-219. MR2015007

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

Whittle, F. (1988). Restless bandits: activity allocation in a changing world. *Journal of Applied Probability* **25** 287-298. MR0974588

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