Bernoulli **23**(4B), 2017, 3685–3710
DOI: 10.3150/16-BEJ862

Some monotonicity properties of parametric and nonparametric Bayesian bandits

YAMING YU

Department of Statistics, University of California, Irvine, CA 92697, USA. E-mail: yamingy@uci.edu

One of two independent stochastic processes (arms) is to be selected at each of n stages. The selection is sequential and depends on past observations as well as the prior information. The objective is to maximize the expected future-discounted sum of the n observations. We study structural properties of this classical bandit problem, in particular how the maximum expected payoff and the optimal strategy vary with the priors, in two settings: (a) observations from each arm have an exponential family distribution and different arms are assigned independent conjugate priors; (b) observations from each arm have a nonparametric distribution and different arms are assigned independent Dirichlet process priors. In both settings, we derive results of the following type: (i) for a particular arm and a fixed prior weight, the maximum expected payoff increases as the prior mean yield increases; (ii) for a fixed prior mean yield, the maximum expected payoff increases as the prior weight decreases. Specializing to the one-armed bandit, the second result captures the intuition that, given the same immediate payoff, the less one knows about an arm, the more desirable it becomes because there remains more information to be gained when selecting that arm. In the parametric case, our results extend those of (*Ann. Statist.* **20** (1992) 1625–1636) concerning Bernoulli and normal bandits (see also (In *Time Series and Related Topics* (2006) pp. 284–294 IMS)). In the nonparametric case, we extend those of (*Ann. Statist.* **13** (1985) 1523–1534). A key tool in the derivation is stochastic orders.

Keywords: Bernoulli bandits; convex order; Dirichlet bandits; log-concavity; optimal stopping; sequential decision; two-armed bandits

1. Introduction

At each of n stages, an experimenter must take an observation from one of two stochastic processes (arms). Let us adopt the Bayesian framework and assume that the experimenter's belief about an unknown arm is updated according to Bayes theorem after each observation. A strategy specifies which process to select at each stage. The objective is to maximize the expected payoff, $\sum_{i=1}^n a_i Z_i$, where Z_i is the observation at stage i and $A_n \equiv (a_1, a_2, \dots, a_n)$ is a discount sequence satisfying $a_i \geq 0$ and $\sum_{i=1}^n a_i > 0$. A strategy is optimal if it achieves the maximum expected payoff. An arm is optimal initially if there exists an optimal strategy that selects that arm at the first stage. This is a finite-horizon two-armed bandit [4], a classical problem in sequential decision theory.

Bernoulli bandits, where each arm generates binary observations, are important as a model for clinical trials, and have received considerable attention [1,2,4]. Others such as normal bandits [8,9,25] have also been extensively studied. Clayton and Berry [10] have introduced a one-armed Bayesian nonparametric bandit using Dirichlet process priors [11]. Chattopadhyay [7] extends this and studies the two-armed Dirichlet bandit. We shall consider both parametric and nonpara-

1350-7265 © 2017 ISI/BS

metric bandits in this paper. Moreover, our results are not limited to two-armed cases, although we present them in such terms for notational convenience.

For such problems, one must balance the desire to maximize the immediate payoff and the need to explore a less known arm in the hope of higher payoff later on (the exploitation versus exploration dilemma). From a Bayesian perspective, the optimal strategy is easily specified through backward induction, although its computation can be nontrivial. If the discount sequence is geometric, then the problem reduces to several one-armed bandits [12,14,18,24] and the optimal strategy is to choose an arm with the highest dynamic allocation index, or Gittins index. For a recent exposition to Gittins index theory, see [13]. Optimal strategies for general discount sequences are less tractable.

The Gittins index possesses intriguing monotonicity properties with respect to prior specifications. For example, [15] show that the Gittins index decreases in $\tau > 0$ for some special bandit arms: a Bernoulli arm whose unknown parameter has a $\mathrm{Beta}(\tau s, \tau(1-s))$ prior $(0 < s < 1)$, or a normal arm whose unknown mean has a $\mathrm{N}(\mu, 1/\tau)$ prior $(\mu \in \mathbb{R})$. In both cases, τ is naturally interpreted as the amount of prior information. Such monotonicity results therefore capture an aspect of the exploration–exploitation dilemma in precise terms: given the same immediate payoff, the less one knows about an arm, the more desirable it becomes since there is more room for exploration. Though easy to state and intuitively appealing, such results are often difficult to prove. We mention a long-standing conjecture of [2], which states that for a finite-horizon Bernoulli two-armed bandit with uniform discounting and independent $\mathrm{Beta}(u_i, v_i)$ priors, $i = 1, 2$, for arms 1 and 2 respectively, if $u_1/v_1 = u_2/v_2$ and $u_1 + v_1 < u_2 + v_2$, then arm 1 is preferred to arm 2 at the initial pull. If, instead of finite-horizon uniform discounting, we assume infinite-horizon geometric discounting, then the corresponding conjecture is true, as shown by [15]. While our results are not strong enough to confirm Berry's original conjecture, they add evidence that it is likely to hold.

The Bernoulli and normal bandits can be regarded as special cases of a general bandit where observations from each arm have an exponential family distribution. Assume each arm is assigned an independent conjugate prior, which is characterized by a prior mean yield and a prior weight. The prior mean yield specifies the immediate payoff of an arm, whereas the prior weight reflects the associated uncertainty. In this more general setting, we show that: (i) for a fixed prior weight, the maximum expected payoff increases as the prior mean yield for any arm increases; (ii) for a fixed prior mean yield, the maximum expected payoff increases as the prior weight for any arm decreases. These generalize and unify several results in the literature concerning specific distributions. We do not present numerical calculations but it would be interesting to see to what extent such monotonicity results still hold when the assumptions such as conjugate priors are relaxed.

We also study Dirichlet bandits, which do not fit in the one-parameter exponential family framework. For Dirichlet bandits with known arm 2, [10] obtain several structural results. In particular, the maximum expected payoff increases as F_1, the mean of the Dirichlet process prior for arm 1, increases in the usual stochastic order. Also, a version of the stay-on-a-winner rule [2,5] holds: if arm 1 is optimal initially then it is optimal at the next stage provided that the initial observation from arm 1 is sufficiently large. Such results have been extended to the general two-armed Dirichlet bandits [7]. In this paper, we obtain further structural properties of Dirichlet bandits concerning how the value of the bandit (i.e., the maximum expected payoff) varies with the Dirichlet process priors. In particular, we show that (i) the value increases as the mean of the Dirichlet process for any arm becomes larger in the increasing convex order; (ii) the value decreases as the prior weight of the Dirichlet process of an arm increases. We confirm some conjectures of [10] along the way.

A key tool in our derivation is the notion of stochastic ordering [20,22]. We shall use the usual stochastic order $\leq_{\mathrm{st}}$, the convex order $\leq_{\mathrm{cx}}$, the increasing convex order $\leq_{\mathrm{icx}}$, the likelihood ratio order $\leq_{\mathrm{lr}}$, and the relative log-concavity order $\leq_{\mathrm{lc}}$. For random variables Z_1 and Z_2 taking values on $\mathbb{R}$, we write $Z_1 \leq_{\mathrm{st}} Z_2$ (respectively, $Z_1 \leq_{\mathrm{cx}} Z_2$), if

$$E\phi(Z_1) \leq E\phi(Z_2) \tag{1}$$

for every increasing (respectively, convex) function ϕ such that the expectations exist. If $Z_1 \leq_{\mathrm{st}} Z_2$ then we also say Z_2 is to the right of Z_1. We say Z_1 is smaller than Z_2 in the increasing convex order, written as $Z_1 \leq_{\mathrm{icx}} Z_2$, if (1) holds for every increasing and convex function ϕ such that the expectations exist. Hence, $\leq_{\mathrm{icx}}$ is implied by either $\leq_{\mathrm{st}}$ or $\leq_{\mathrm{cx}}$. The convex order is especially important in deriving our results concerning Dirichlet bandits.

If Z_1 and Z_2 have densities $f_1(z)$ and $f_2(z)$ respectively, supported on the same interval, then we write $Z_1 \leq_{\mathrm{lr}} Z_2$ (respectively, $Z_1 \leq_{\mathrm{lc}} Z_2$) if $\log(f_1(z)/f_2(z))$ is decreasing (respectively, concave) in z. For example, the $\mathrm{Beta}(\tau s, \tau(1-s))$ density increases in the likelihood ratio order as $s \in (0,1)$ increases, and decreases in the relative log-concavity order as τ increases. (We use $\leq_{\mathrm{st}}$, $\leq_{\mathrm{cx}}$, $\leq_{\mathrm{icx}}$, $\leq_{\mathrm{lr}}$ and $\leq_{\mathrm{lc}}$ with densities as well as random variables.) A basic property is $\leq_{\mathrm{lr}} \Longrightarrow \leq_{\mathrm{st}}$. Assuming equal means, it also holds that $\leq_{\mathrm{lc}}$ implies $\leq_{\mathrm{cx}}$. Intuitively, the relative log-concavity order compares the amount of information as it is defined through curvatures of the log density functions. Both $\leq_{\mathrm{lr}}$ and $\leq_{\mathrm{lc}}$ are preserved under the prior-to-posterior updating, which makes them ideal for studying structural properties in parametric bandit problems. The log-concavity order is also useful in other seemingly unrelated contexts [23,26–28].

The rest of the paper is organized as follows. In Section 2, we consider the parametric case with exponential family distributions and conjugate priors. After setting up the exponential family framework in Section 2.1, we present basic structural results such as a stay-on-a-winner rule in Section 2.2. Section 2.3 contains the main results, including monotonicity of the value function with respect to prior weights. Section 2.4 applies the results in Section 2.3 to one-armed bandits. In particular, we show that the break-even value, or Gittins index, decreases as the prior weight of the unknown arm increases. In Sections 2.5 and 2.6, we extend the monotonicity results to nonconjugate priors for Bernoulli and normal bandits, respectively. Section 2.7 concludes the parametric case with a brief discussion on an open problem. In Section 3, we consider the nonparametric case with Dirichlet process priors where similar results are derived. We mention that although the results are similar to the parametric case in form, the proofs are somewhat different and are not straightforward extensions.

2. The parametric case with exponential family distributions and conjugate priors

2.1. Preliminaries

Let ν be a σ-finite measure on $\mathbb{R}$ that is not a point mass. Denote

$$\psi(\theta) = \log \int e^{\theta x}\, \mathrm{d}\nu(x), \qquad \theta \in \Theta,$$

where Θ is the natural parameter space defined as the set of $\theta \in \mathbb{R}$ such that $\psi(\theta)$ is finite. We assume that Θ has a non-empty interior. Suppose that given θ_i, observations from arm i are independent and identically distributed (i.i.d.) according to the density (relative to ν)

$$f(x|\theta_i) = e^{\theta_i x - \psi(\theta_i)}. \tag{2}$$

Let us assume independent conjugate priors on θ_i, $i = 1, 2$, with Lebesgue density

$$f(\theta_i|\gamma_i, \tau_i) \propto e^{\theta_i \gamma_i - \tau_i \psi(\theta_i)}, \qquad \theta_i \in \Theta. \tag{3}$$

Let $\mathcal{K}$ denote the smallest open interval such that ν assigns no mass outside of the closure $\bar{\mathcal{K}}$. To ensure that the priors are proper, we require $\tau_i > 0$ and $\gamma_i/\tau_i \in \mathcal{K}$ ([6], Chapter 4). As usual τ_i is regarded as the "prior sample size" and γ_i the "prior sum of observations". We refer to (3) as the (γ_i, τ_i) prior and call this two-armed bandit with discount sequence A_n the $(\gamma_1, \tau_1; \gamma_2, \tau_2; A_n)$ bandit. Its value (i.e., maximum expected payoff) is denoted by $V(\gamma_1, \tau_1; \gamma_2, \tau_2; A_n)$.

This framework unifies several well-studied bandit reward structures: (i) Bernoulli rewards whose unknown parameter has a Beta$(\gamma, \tau - \gamma)$ prior; (ii) normal rewards whose unknown mean has a N$(\gamma/\tau, 1/\tau)$ prior; (iii) exponential rewards whose unknown rate parameter has a Gamma$(\tau + 1, \gamma)$ prior; (iv) Poisson rewards whose unknown rate parameter has a Gamma(γ, τ) prior. Extensions to general priors for (i) and (ii) are considered in Sections 2.5 and 2.6, respectively.

Let $V^i(\gamma_1, \tau_1; \gamma_2, \tau_2; A_n)$ be the expected payoff when selecting arm i initially and using an optimal strategy thereafter. Then

$$V(\gamma_1, \tau_1; \gamma_2, \tau_2; A_n) = \max\bigl\{V^1(\gamma_1, \tau_1; \gamma_2, \tau_2; A_n), V^2(\gamma_1, \tau_1; \gamma_2, \tau_2; A_n)\bigr\}, \tag{4}$$

and it is optimal to start with the arm whose V^i is larger. Suppose arm 1 is selected, resulting in an observation X. By conjugacy, the posterior for θ_1 is again of the form of (3) with $(\gamma_1 + X, \tau_1 + 1)$ in place of (γ_1, τ_1). Thus, we have

$$V^1(\gamma_1, \tau_1; \gamma_2, \tau_2; A_n) = a_1\mu_1 + E\bigl[V\bigl(\gamma_1 + X, \tau_1 + 1; \gamma_2, \tau_2; A_n^1\bigr)|\gamma_1, \tau_1\bigr], \tag{5}$$

$$V^2(\gamma_1, \tau_1; \gamma_2, \tau_2; A_n) = a_1\mu_2 + E\bigl[V\bigl(\gamma_1, \tau_1; \gamma_2 + Y, \tau_2 + 1; A_n^1\bigr)|\gamma_2, \tau_2\bigr], \tag{6}$$

where $A_n^1 = (a_2, a_3, \dots, a_n)$ and μ_i denotes the expected value of an observation from arm i under the (γ_i, τ_i) prior. This μ_i is simply $\mu_i = \gamma_i/\tau_i$, which we refer to as the prior mean yield (this is not to be confused with $E\theta$). In $E[g(X)|\gamma_1, \tau_1]$, we use X to denote a generic observation from arm 1 under the (γ_1, τ_1) prior; similarly for Y. That is, the density of X relative to ν is

$$f(x) \propto \int_\Theta e^{\theta(\gamma_1 + x) - (\tau_1 + 1)\psi(\theta)}\, \mathrm{d}\theta. \tag{7}$$

The dynamic programming equations (4)–(6) are crucial for both theoretical analysis and numerical computation of the optimal strategy.

2.2. Stay-on-a-winner

This subsection derives a basic monotonicity property of the optimal strategy: as the observation from an arm becomes larger, the inclination to pull that arm again also increases. Under suitable conditions, we prove a generalized stay-on-a-winner rule, which is a natural extension of the results for Bernoulli bandits [2,4,5].

Let us define the advantage of arm 1 over arm 2 as

$$\Delta(\gamma_1, \tau_1; \gamma_2, \tau_2; A_n) = V^1(\gamma_1, \tau_1; \gamma_2, \tau_2; A_n) - V^2(\gamma_1, \tau_1; \gamma_2, \tau_2; A_n).$$

Define $\Delta^+ = \max\{\Delta, 0\}$ and $\Delta^- = \min\{\Delta, 0\}$. By considering the initial two pulls, one can show [2]

$$\Delta(\gamma_1, \tau_1; \gamma_2, \tau_2; A_n) = (a_1 - a_2)\left(\frac{\gamma_1}{\tau_1} - \frac{\gamma_2}{\tau_2}\right) \tag{8}$$

$$+ E\left[\Delta^+\left(\gamma_1 + X, \tau_1 + 1; \gamma_2, \tau_2; A_n^1\right)|\gamma_1, \tau_1\right] \tag{9}$$

$$+ E\left[\Delta^-\left(\gamma_1, \tau_1; \gamma_2 + Y, \tau_2 + 1; A_n^1\right)|\gamma_2, \tau_2\right]. \tag{10}$$

Proposition 1 states that as the prior mean yield of arm 1 increases, so does the advantage of arm 1 over arm 2, assuming A_n is decreasing. This can be extended to non-conjugate priors. Specifically, Δ increases as the prior for arm 1 becomes larger in the likelihood ratio order. Extensions to general Markov decision problems are also possible [21]. We provide a complete proof which serves as an introduction to the derivation of the main results in Section 2.3.

Proposition 1. *Suppose A_n is decreasing. Then $\Delta(\gamma_1, \tau_1; \gamma_2, \tau_2; A_n)$ increases in γ_1.*

Proof. The $n = 1$ case is easy. Let us use induction for $n \geq 2$. In view of (8)–(10), we only need to show that

$$E\left[\Delta^+\left(\gamma_1 + X, \tau_1 + 1; \gamma_2, \tau_2; A_n^1\right)|\gamma_1, \tau_1\right] \quad \text{and} \tag{11}$$

$$E\left[\Delta^-\left(\gamma_1, \tau_1; \gamma_2 + Y, \tau_2 + 1; A_n^1\right)|\gamma_2, \tau_2\right] \tag{12}$$

both increase in γ_1. Monotonicity of (12) follows from the induction hypothesis. To handle (11), let us consider $\gamma_1 < \tilde{\gamma}_1$. Let θ_1 and $\tilde{\theta}_1$ have the (γ_1, τ_1) and $(\tilde{\gamma}_1, \tau_1)$ priors, respectively. Let

$g(x)$ (respectively, $\tilde{g}(x)$) be the marginal density of X if it is drawn according to (2) given θ_1 (respectively, $\tilde{\theta}_1$). Note that $\theta_1 \leq_{\mathrm{lr}} \tilde{\theta}_1$. In view of (7), we know that $g \leq_{\mathrm{lr}} \tilde{g}$ by total positivity considerations ([17], Chapter 3). It follows that $g \leq_{\mathrm{st}} \tilde{g}$. By the induction hypothesis,

$$\phi(x) \equiv \Delta^{+}\bigl(x, \tau_1 + 1; \gamma_2, \tau_2; A_n^1\bigr)$$

increases in x. Thus

$$\begin{aligned} E\bigl[\phi(\gamma_1 + X)|\gamma_1, \tau_1\bigr] &\leq E\bigl[\phi(\tilde{\gamma}_1 + X)|\gamma_1, \tau_1\bigr] \\ &\leq E\bigl[\phi(\tilde{\gamma}_1 + X)|\tilde{\gamma}_1, \tau_1\bigr], \end{aligned} \tag{13}$$

where (13) holds because $g \leq_{\mathrm{st}} \tilde{g}$. Hence, (11) increases in γ_1. □

Corollary 1. *Suppose A_n is a decreasing sequence, and an observation x is taken from arm* 1 *initially. Then, at the second stage, either arm* 1 *is optimal for all x, or arm* 2 *is optimal for all x, or there exists some $x_* \in \mathcal{K}$ such that arm* 1 *is optimal if $x \geq x_*$ and arm* 2 *is optimal if $x \leq x_*$.*

Proof. We can show that $\Delta(\gamma_1 + x, \tau_1 + 1; \gamma_2, \tau_2; A_n^1)$ is continuous in x, which is treated as a real number in $\bar{\mathcal{K}}$ even though an actual observation from arm 1 may be discrete. (One method is to use the convexity arguments of Proposition 2 in Section 2.3.) The claim then follows from Proposition 1. □

The next result, Theorem 1, is a generalized stay-on-a-winner rule: under suitable conditions if an arm is optimal initially then it continues to be optimal at the next stage provided that the initial observation from that arm is large enough.

Theorem 1. *Assume A_n is decreasing, $n \geq 2$, and either* (i) $a_1 = a_2$ *or* (ii) $\gamma_1/\tau_1 \leq \gamma_2/\tau_2$ *holds. Assume $\Delta(\gamma_1, \tau_1; \gamma_2, \tau_2; A_n) \geq 0$, that is, arm* 1 *is optimal initially. Then $\Delta(\gamma_1 + x, \tau_1 + 1; \gamma_2, \tau_2; A_n^1) \geq 0$ for sufficiently large $x \in \bar{\mathcal{K}}$.*

Proof. We may assume $a_i > 0$ for all $i \leq n$. Let U be the upper end point of $\mathcal{K}$. If $U = \infty$, then using (8)–(10), it is easy to show by induction that $\Delta(\gamma_1 + x, \tau_1 + 1; \gamma_2, \tau_2; A_n^1) > 0$ for sufficiently large x. That is, the claim holds even without assuming that arm 1 is optimal initially. Assume $U < \infty$ and $\Delta(\gamma_1, \tau_1; \gamma_2, \tau_2; A_n) \geq 0$. By (8)–(10), we have

$$0 \leq E\bigl[\Delta^{+}\bigl(\gamma_1 + X, \tau_1 + 1; \gamma_2, \tau_2; A_n^1\bigr)|\gamma_1, \tau_1\bigr] + E\bigl[\Delta^{-}\bigl(\gamma_1, \tau_1; \gamma_2 + Y, \tau_2 + 1; A_n^1\bigr)|\gamma_2, \tau_2\bigr]. \tag{14}$$

Suppose the claim does not hold, that is, $\Delta(\gamma_1 + x, \tau_1 + 1; \gamma_2, \tau_2; A_n^1) < 0$ for all $x \in \bar{\mathcal{K}}$. In particular,

$$\Delta\bigl(\gamma_1 + U, \tau_1 + 1; \gamma_2, \tau_2; A_n^1\bigr) < 0. \tag{15}$$

Then it is necessary that both expectations in (14) are zero. That is,

$$\Delta\bigl(\gamma_1, \tau_1; \gamma_2 + y, \tau_2 + 1; A_n^1\bigr) \geq 0 \qquad \text{for all } y \in \mathcal{K}.$$

By continuity, $\Delta(\gamma_1, \tau_1; \gamma_2 + U, \tau_2 + 1; A_n^1) \geq 0$. However, the $(\gamma_1 + U, \tau_1 + 1)$ prior is larger than the (γ_1, τ_1) prior in the likelihood ratio order. The argument of Proposition 1 yields

$$\begin{aligned}\Delta\bigl(\gamma_1 + U, \tau_1 + 1; \gamma_2, \tau_2; A_n^1\bigr) &\geq \Delta\bigl(\gamma_1, \tau_1; \gamma_2, \tau_2; A_n^1\bigr)\\ &\geq \Delta\bigl(\gamma_1, \tau_1; \gamma_2 + U, \tau_2 + 1; A_n^1\bigr) \geq 0,\end{aligned}$$

which contradicts (15). □

2.3. Monotonicity

Proposition 2 shows that the maximum expected payoff is an increasing and convex function of the prior mean yield of any arm. The convexity will be useful in proving Theorem 2 concerning monotonicity with respect to the prior weight.

Proposition 2. *$V(\gamma_1, \tau_1; \gamma_2, \tau_2; A_n)$ is increasing and convex in each of γ_i, $i = 1, 2$.*

Proof. Monotonicity holds by the same argument that proves Proposition 1. Let us focus on the convexity with respect to γ_1. The $n = 1$ case is easy. For $n \geq 2$ we use induction. Note that by (4)–(6) it suffices to show that both

$$E\bigl[V\bigl(\gamma_1 + X, \tau_1 + 1; \gamma_2, \tau_2; A_n^1\bigr)|\gamma_1, \tau_1\bigr] \quad \text{and} \tag{16}$$

$$E\bigl[V\bigl(\gamma_1, \tau_1; \gamma_2 + Y, \tau_2 + 1; A_n^1\bigr)|\gamma_2, \tau_2\bigr] \tag{17}$$

are convex in γ_1. The claim for (17) follows from the induction hypothesis. To deal with (16), suppose $\gamma_1 < \tilde{\gamma}_1$. Denote the marginal of X when the prior on θ is (γ_1, τ_1) (respectively, $(\tilde{\gamma}_1, \tau_1)$) by g (respectively, $\tilde{g}$). Then $g \leq_{\mathrm{st}} \tilde{g}$ as in the proof of Proposition 1. By the induction hypothesis,

$$\phi(x) \equiv V\bigl(x, \tau_1 + 1; \gamma_2, \tau_2; A_n^1\bigr)$$

is convex in x. Moreover, for fixed $\rho \in (0, 1)$, we have

$$\begin{aligned}&E\bigl[\phi(\gamma_1 + X)|\gamma_1, \tau_1\bigr] - E\bigl[\phi\bigl(\rho\gamma_1 + (1 - \rho)\tilde{\gamma}_1 + X\bigr)|\gamma_1, \tau_1\bigr]\\ &\quad\geq \rho^{-1}(1 - \rho)E\bigl[\eta(X)|\gamma_1, \tau_1\bigr] &&(18)\\ &\quad\geq \rho^{-1}(1 - \rho)E\bigl[\eta(X)|\tilde{\gamma}_1, \tau_1\bigr], &&(19)\end{aligned}$$

where

$$\eta(x) \equiv \phi\bigl(\rho\gamma_1 + (1 - \rho)\tilde{\gamma}_1 + x\bigr) - \phi(\tilde{\gamma}_1 + x).$$

The inequality (18) holds because ϕ is convex; (19) holds because η is decreasing and $g \leq_{\mathrm{st}} \tilde{g}$. Rearranging, we get

$$\rho E\bigl[\phi(\gamma_1 + X)|\gamma_1, \tau_1\bigr] + (1 - \rho)E\bigl[\phi(\tilde{\gamma}_1 + X)|\tilde{\gamma}_1, \tau_1\bigr] \geq E\phi\bigl(\rho\gamma_1 + (1 - \rho)\tilde{\gamma}_1 + X^*\bigr),$$

where X^* has the following distribution. Given θ, X^* is distributed according to (2); the prior on θ is a mixture of (γ_1, τ_1) and $(\tilde{\gamma}_1, \tau_1)$ with weights ρ and $1-\rho$ respectively. Denote this mixture density by $h^*(\theta)$, and the $(\rho\gamma_1 + (1-\rho)\tilde{\gamma}_1, \tau_1)$ prior density by $h(\theta)$. Then $h(\theta) \le_{\mathrm{lc}} h^*(\theta)$, because log-convexity is closed under mixtures [19]. Consider the difference between the marginal densities

$$D(x) \equiv \int_\Theta e^{x\theta - \psi(\theta)} \bigl[h(\theta) - h^*(\theta)\bigr] \, \mathrm{d}\theta.$$

Relative log-concavity implies that, as θ traverses Θ, $h(\theta) - h^*(\theta)$ changes signs at most twice and, in the case of two changes, the sign sequence is $-, +, -$. By the variation-diminishing properties of the Laplace transform ([17], Chapter 5), $D(x)$ has at most two changes of sign, and in the case of two changes, the sign sequence is $-, +, -$. Note that, when the prior is either h or h^*, the marginal mean of X is the same, namely $(\rho\gamma_1 + (1-\rho)\tilde{\gamma}_1)/\tau_1$. Hence it is not possible for $D(x)$ to change signs exactly once. Unless $D(x) \equiv 0$, its sign sequence must be $-, +, -$. It follows that the marginal distribution of X becomes larger in the convex order when $h^*(\theta)$ replaces $h(\theta)$ as the prior for θ (see, e.g., [28], Lemma 1). Using the convexity of ϕ again, we obtain

$$E\phi\bigl(\rho\gamma_1 + (1-\rho)\tilde{\gamma}_1 + X^*\bigr) \ge E\bigl[\phi\bigl(\rho\gamma_1 + (1-\rho)\tilde{\gamma}_1 + X\bigr) | \rho\gamma_1 + (1-\rho)\tilde{\gamma}_1, \tau_1\bigr].$$

It follows that $E[\phi(\gamma_1 + X)|\gamma_1, \tau_1]$, i.e., (16), is convex in γ_1, as required. □

Our main result, Theorem 2, shows that the value of the bandit decreases as the prior weight of an arm increases. That is, given the same immediate payoff, an arm becomes less desirable as the amount of information about it increases.

Theorem 2. *$V(c\gamma_1, c\tau_1; \gamma_2, \tau_2; A_n)$ decreases in $c \in (0, \infty)$.*

Proof. Let us use induction on n. The $n = 1$ case is easy. Suppose $n \ge 2$. In view of (4)–(6), we only need to show that

$$E\bigl[V\bigl(c\gamma_1 + X, c\tau_1 + 1; \gamma_2, \tau_2; A_n^1\bigr)|c\gamma_1, c\tau_1\bigr] \quad \text{and} \tag{20}$$

$$E\bigl[V\bigl(c\gamma_1, c\tau_1; \gamma_2 + Y, \tau_2 + 1; A_n^1\bigr)|\gamma_2, \tau_2\bigr] \tag{21}$$

both decrease in c. By the induction hypothesis, (21) decreases in c. To deal with (20), suppose $0 < c < \tilde{c}$ and denote $\xi = (c\tau_1 + 1)/(\tilde{c}\tau_1 + 1)$. We get

$$\begin{aligned}
&E\bigl[V\bigl(c\gamma_1 + X, c\tau_1 + 1; \gamma_2, \tau_2; A_n^1\bigr)|c\gamma_1, c\tau_1\bigr] \\
&\quad \ge E\bigl[V\bigl(\xi(\tilde{c}\gamma_1 + X), c\tau_1 + 1; \gamma_2, \tau_2; A_n^1\bigr)|c\gamma_1, c\tau_1\bigr] && (22)\\
&\quad \ge E\bigl[V\bigl(\tilde{c}\gamma_1 + X, \tilde{c}\tau_1 + 1; \gamma_2, \tau_2; A_n^1\bigr)|c\gamma_1, c\tau_1\bigr] && (23)\\
&\quad \ge E\bigl[V\bigl(\tilde{c}\gamma_1 + X, \tilde{c}\tau_1 + 1; \gamma_2, \tau_2; A_n^1\bigr)|\tilde{c}\gamma_1, \tilde{c}\tau_1\bigr]. && (24)
\end{aligned}$$

The inequality (22) holds by the convexity of V as shown by Proposition 2, noting

$$\xi(\tilde{c}\gamma_1 + X) \leq_{\mathrm{cx}} c\gamma_1 + X$$

(see Lemma 3 in Section 2.6, or [22], Theorem 3.A.18). The inequality (23) holds by the induction hypothesis, as $\xi < 1$. The inequality (24) holds by an argument similar to the proof of Proposition 2. Specifically, the prior $(\tilde{c}\gamma_1, \tilde{c}\tau_1)$ is log-concave relative to $(c\gamma_1, c\tau_1)$. Thus the marginal of X increases in the convex order if $(c\gamma_1, c\tau_1)$ replaces $(\tilde{c}\gamma_1, \tilde{c}\tau_1)$ as the prior on θ (the mean of X remains constant). Overall (20) decreases in c, as required. □

Remark. Proposition 2 and Theorem 2 extend naturally to bandits with more than two arms. We present the two-armed version for simplicity. The discount sequence A_n is only required to be nonnegative. By approximation, this can be further extended to the infinite-horizon case assuming $\sum_{i=1}^{\infty} a_i < \infty$. Similar comments apply to Theorem 5 in Section 3.1 and Theorem 6 in Section 3.2, which are our main results in the nonparametric case.

2.4. The one-armed case

In this subsection, we consider the one-armed bandit where we assume arm 2 yields a constant payoff λ at each pull. We shall abuse the notation by calling this a $(\gamma, \tau; \lambda; A_n)$ bandit, where we drop the subscripts on γ_1 and τ_1 for convenience. Results in Section 2.3 are applied to derive monotonicity properties of the break-even value. It is also shown (Proposition 3) that if both arms are optimal initially, then an observation from arm 1 that is less than its prior mean yield would make arm 2 optimal thereafter.

A discount sequence $A_n = (a_1, a_2, \ldots)$ is called *regular* if, letting $b_j = \sum_{i \geq j} a_i$, we have $b_{j+1}^2 \geq b_j b_{j+2}$ for all $j \geq 1$ [3]. For regular discount sequences, our one-armed bandit is an optimal stopping problem, that is, if at any stage the known arm becomes optimal then it remains optimal in all subsequent stages. Moreover, if A_n is regular and $a_1 > 0$, then there exists a break-even value $\Lambda(\gamma, \tau; A_n)$ for the $(\gamma, \tau; \lambda; A_n)$ bandit, such that arm 1 is optimal initially if and only if $\lambda \leq \Lambda(\gamma, \tau; A_n)$ and arm 2 is optimal initially if and only if $\lambda \geq \Lambda(\gamma, \tau; A_n)$. For infinite-horizon geometric discounting, this break-even value is also known as the dynamic allocation index or Gittins index [14]. The following result holds by the optimal stopping characterization.

Lemma 1. *If A_n is regular and $a_1 > 0$, then $\Lambda(\gamma, \tau; A_n)$ is the smallest λ such that*

$$V(\gamma, \tau; \lambda; A_n) \leq \lambda \sum_{i=1}^{n} a_i.$$

Corollary 2 summarizes some monotonicity properties of $\Lambda(\gamma, \tau; A_n)$. It extends to infinite-horizon regular discounting. As special cases, we recover the results of [15] on Bernoulli and normal bandits with geometric discounting.

Corollary 2. *If A_n is regular and $a_1 > 0$, then $\Lambda(c\gamma, c\tau; A_n)$ decreases in $c > 0$ and strictly increases in γ.*

Proof. Monotonicity in c follows from Theorem 2 and Lemma 1. Monotonicity in γ follows from Proposition 2 and Lemma 1. To show strict monotonicity, let us set $c = 1$ and assume that $\gamma, \tilde{\gamma}$ satisfy $\gamma < \tilde{\gamma}$ and

$$\Lambda(\gamma, \tau; A_n) = \Lambda(\tilde{\gamma}, \tau; A_n) \equiv \lambda_*.$$

Then, as in the proof of Proposition 1, we get

$$\begin{aligned} \lambda_* \sum_{i=1}^{n} a_i &= a_1 \frac{\gamma}{\tau} + E\big[V\big(\gamma + X, \tau + 1; \lambda_*; A_n^1\big) | \gamma, \tau\big] \\ &< a_1 \frac{\tilde{\gamma}}{\tau} + E\big[V\big(\gamma + X, \tau + 1; \lambda_*; A_n^1\big) | \gamma, \tau\big] \\ &\le a_1 \frac{\tilde{\gamma}}{\tau} + E\big[V\big(\tilde{\gamma} + X, \tau + 1; \lambda_*; A_n^1\big) | \tilde{\gamma}, \tau\big] \\ &= \lambda_* \sum_{i=1}^{n} a_i, \end{aligned}$$

which is a contradiction. □

For a regular and positive discount sequence A_n, Proposition 3 shows that there exists a break-even observation $b(\gamma, \tau; A_n)$ for the $(\gamma, \tau; \lambda; A_n)$ bandit such that if both arms are optimal initially, and an observation x is taken from arm 1, then arm 1 remains optimal if $x \ge b(\gamma, \tau; A_n)$ and arm 2 becomes optimal if $x \le b(\gamma, \tau; A_n)$. Moreover, this break-even observation is no smaller than γ/τ, the prior mean yield. Note that $b(\gamma, \tau; A_n)$ is a real number in the open interval $\mathcal{K}$ even though an actual observation from arm 1 may be discrete.

Proposition 3. *Suppose A_n is regular, $n \ge 2$, and $a_1, a_2 > 0$. Then there exists a unique $b(\gamma, \tau; A_n) \in \mathcal{K}$ such that $b(\gamma, \tau; A_n) \ge \gamma/\tau$ and*

$$\Lambda(\gamma, \tau; A_n) \ge \Lambda\big(\gamma + x, \tau + 1; A_n^1\big), \qquad \text{if } x \le b(\gamma, \tau; A_n); \tag{25}$$

$$\Lambda(\gamma, \tau; A_n) \le \Lambda\big(\gamma + x, \tau + 1; A_n^1\big), \qquad \text{if } x \ge b(\gamma, \tau; A_n). \tag{26}$$

To prove Proposition 3 we need a continuity lemma. Its proof, taken from [10], is included for completeness.

Lemma 2. *Suppose A_n is regular and $a_1 > 0$. Then $\Lambda(\gamma, \tau; A_n)$ is continuous in γ.*

Proof. Fix γ_0 and note that $\lambda = \Lambda(\gamma, \tau; A_n)$ is the unique root of

$$V^1(\gamma, \tau; \lambda; A_n) - V^2(\gamma, \tau; \lambda; A_n) = 0.$$

By continuity of V^1 and V^2, we have

$$\begin{aligned} 0 &= \lim_{\gamma \uparrow \gamma_0} \big[V^1\big(\gamma, \tau; \Lambda(\gamma, \tau; A_n); A_n\big) - V^2\big(\gamma, \tau; \Lambda(\gamma, \tau; A_n); A_n\big)\big] \\ &= V^1\Big(\gamma_0, \tau; \lim_{\gamma \uparrow \gamma_0} \Lambda(\gamma, \tau; A_n); A_n\Big) - V^2\Big(\gamma_0, \tau; \lim_{\gamma \uparrow \gamma_0} \Lambda(\gamma, \tau; A_n); A_n\Big). \end{aligned}$$

By uniqueness of Λ, we have $\lim_{\gamma\uparrow\gamma_0}\Lambda(\gamma,\tau;A_n)=\Lambda(\gamma_0,\tau;A_n)$. Similarly, the limit holds when $\gamma\downarrow\gamma_0$. □

Proof of Proposition 3. Let U be the upper end point of $\mathcal{K}$. If $U=\infty$ then $\Lambda(\gamma+x,\tau+1;A_n^1)\to\infty$ as $x\to\infty$ (the expected payoff by always selecting arm 1 becomes arbitrarily large). If $U<\infty$ then we can show $\Lambda(\gamma+U,\tau+1;A_n^1)>\Lambda(\gamma,\tau;A_n)$ as follows. Assume the contrary and consider the $(\gamma,\tau;\lambda_*;A_n)$ bandit with $\lambda_*=\Lambda(\gamma+U,\tau+1;A_n^1)$. We have

$$\lambda_*\sum_{i=1}^n a_i\le a_1\frac{\gamma}{\tau}+E\bigl[V\bigl(\gamma+X,\tau+1;\lambda_*;A_n^1\bigr)|\gamma,\tau\bigr].$$

Since $\gamma/\tau\in\mathcal{K}$ and $\mathcal{K}$ is open, we have $\lambda_*\ge(\gamma+U)/(\tau+1)>\gamma/\tau$. Thus

$$\begin{aligned}\lambda_*\sum_{i=2}^n a_i&<E\bigl[V\bigl(\gamma+X,\tau+1;\lambda_*;A_n^1\bigr)|\gamma,\tau\bigr]\\&\le V\bigl(\gamma+U,\tau+1;\lambda_*;A_n^1\bigr)\\&=\lambda_*\sum_{i=2}^n a_i,\end{aligned}$$

which is a contradiction. We also have

$$\begin{aligned}\Lambda(\gamma,\tau;A_n)&\ge\Lambda\bigl(\gamma,\tau;A_n^1\bigr)\\&\ge\Lambda\bigl(\gamma+\gamma/\tau,\tau+1;A_n^1\bigr),\end{aligned}$$

where the first inequality holds by the optimal stopping characterization, and the second by Corollary 2.

By Lemma 2 and Corollary 2, $\Lambda(\gamma+x,\tau+1;A_n^1)$ is continuous and strictly increasing in x. Hence there exists a unique $b(\gamma,\tau;A_n)\in[\gamma/\tau,U)$ such that (25) and (26) hold. □

It is tempting to conjecture that $b(\gamma,\tau;A_n)\ge\Lambda(\gamma,\tau;A_n)$, which gives a tighter bound since $\Lambda(\gamma,\tau;A_n)\ge\gamma/\tau$. However, our methods are not yet strong enough to resolve this conjecture. Clayton and Berry [10] conjectured an analogous bound for Dirichlet bandits, which we will prove in Section 3.

2.5. Bernoulli bandits with general priors

As noted earlier, results based on likelihood ratio orders, such as those in Section 2.2, may extend to nonconjugate priors. This section shows that Theorem 2 can also be extended this way, at least in the Bernoulli case.

Given p_i, $i=1,2$, let us assume that observations from arm i are i.i.d. Bernoulli(p_i). Priors on p_i are independent with densities f_i with respect to a σ-finite measure G on [0, 1]. We shall denote the value of this Bernoulli bandit with discount sequence A_n by $V_{\mathrm{B}}(f_1; f_2; A_n)$. Let $\mu(f)$ denote the mean of any prior f, that is, $\mu(f) = \int_{[0,1]} pf(p)\,\mathrm{d}G(p)$.

Theorem 3. *If $f_1 \le_{\mathrm{lc}} \tilde{f}_1$ and $\mu(f_1) \le \mu(\tilde{f}_1)$, then $V_{\mathrm{B}}(f_1; f_2; A_n) \le V_{\mathrm{B}}(\tilde{f}_1; f_2; A_n)$.*

Note that the Beta$(c\alpha, c\beta)$ prior $(c, \alpha, \beta > 0)$ decreases in the relative log-concavity order as c increases. Theorem 3 therefore recovers the Bernoulli case of Theorem 2 for conjugate priors.

Let $\Lambda_{\mathrm{B}}(f; A_n)$ denote the break-even value of a one armed Bernoulli bandit whose unknown arm has prior f. We obtain Corollary 3 as a consequence of Theorem 3 and Lemma 1.

Corollary 3. *Assume A_n is regular and $a_1 > 0$. If $f \le_{\mathrm{lc}} \tilde{f}$ and $\mu(f) \le \mu(\tilde{f})$, then $\Lambda_{\mathrm{B}}(f; A_n) \le \Lambda_{\mathrm{B}}(\tilde{f}; A_n)$.*

Herschkorn [16] posed the problem of identifying a variability ordering between priors so that both V_{B} and Λ_{B} are monotonic with respect to it. Theorem 3 and Corollary 3 show that there is indeed such an ordering, namely the relative log-concavity order (assuming equal means). A conjecture of [16] states that Corollary 3 holds under the weaker assumption $f \le_{\mathrm{cx}} \tilde{f}$. This conjecture remains open.

Proof of Theorem 3. When $\mu(f_1) < \mu(\tilde{f}_1)$, we may define the exponentially tilted density $f^*(p) \propto \tilde{f}_1(p)\exp(\delta p)$ for a suitable $\delta < 0$ so that $\mu(f^*) = \mu(f_1)$. We have $f_1 \le_{\mathrm{lc}} f^* \le_{\mathrm{lr}} \tilde{f}_1$. In view of the likelihood ratio ordering, we can use arguments in the proof of Proposition 1 to show that $V_{\mathrm{B}}(f^*; f_2; A_n) \le V_{\mathrm{B}}(\tilde{f}_1; f_2; A_n)$. Hence, we only need to consider the case with equal means.

Let us assume $\mu(f_1) = \mu(\tilde{f}_1)$ and that $\tilde{f}_1$ is nondegenerate. The $n = 1$ case is easy. For $n \ge 2$ we use induction. The equations (4)–(6) become

$$\begin{aligned}
&V_{\mathrm{B}}(f_1; f_2; A_n) = \max\bigl\{V_{\mathrm{B}}^1(f_1; f_2; A_n), V_{\mathrm{B}}^2(f_1; f_2; A_n)\bigr\};\\
&V_{\mathrm{B}}^1(f_1; f_2; A_n) = \mu(f_1)\bigl(a_1 + V_{\mathrm{B}}(\sigma f_1; f_2; A_n^1)\bigr) + \bigl(1 - \mu(f_1)\bigr)V_{\mathrm{B}}\bigl(\phi f_1; f_2; A_n^1\bigr);\\
&V_{\mathrm{B}}^2(f_1; f_2; A_n) \quad = \mu(f_2)\bigl(a_1 + V_{\mathrm{B}}\bigl(f_1; \sigma f_2; A_n^1\bigr)\bigr) + \bigl(1 - \mu(f_2)\bigr)V_{\mathrm{B}}\bigl(f_1; \phi f_2; A_n^1\bigr).
\end{aligned} \tag{27}$$

We use σf (respectively, ϕf) to denote the posterior density after observing one success (respectively, one failure). That is,

$$(\sigma f)(p) = \frac{f(p)p}{\mu(f)}; \qquad (\phi f)(p) = \frac{f(p)(1-p)}{1-\mu(f)}.$$

Because $f_1 \le_{\mathrm{lc}} \tilde{f}_1$ and $\mu(f_1) = \mu(\tilde{f}_1)$ we have $f_1 \le_{\mathrm{cx}} \tilde{f}_1$ (see, e.g., [28], Theorem 12). Thus

$$\mu(f_1)\mu(\sigma f_1) = \int_{[0,1]} p^2 f_1(p)\,\mathrm{d}G(p) \le \int_{[0,1]} p^2 \tilde{f}_1(p)\,\mathrm{d}G(p) = \mu(\sigma\tilde{f}_1)\mu(\tilde{f}_1),$$

yielding $\mu(\sigma f_1) \le \mu(\sigma \tilde{f}_1)$. Similarly, $\mu(\phi f_1) \ge \mu(\phi \tilde{f}_1)$. Define

$$\varepsilon^* = \frac{\mu(\sigma \tilde{f}_1) - \mu(\sigma f_1)}{\mu(\sigma \tilde{f}_1) - \mu(\phi \tilde{f}_1)}; \qquad \varepsilon_* = \frac{\mu(\phi f_1) - \mu(\phi \tilde{f}_1)}{\mu(\sigma \tilde{f}_1) - \mu(\phi \tilde{f}_1)}.$$

Then $\varepsilon^*, \varepsilon_* \in [0, 1)$. Define

$$g^* = (1 - \varepsilon^*)\sigma \tilde{f}_1 + \varepsilon^* \phi \tilde{f}_1; \qquad g_* = \varepsilon_* \sigma \tilde{f}_1 + (1 - \varepsilon_*)\phi \tilde{f}_1.$$

Convexity of V_{B} with respect to mixtures gives

$$\begin{aligned} V_{\mathrm{B}}\big(g^*; f_2; A_n^1\big) &\le (1 - \varepsilon^*) V_{\mathrm{B}}\big(\sigma \tilde{f}_1; f_2; A_n^1\big) + \varepsilon^* V_{\mathrm{B}}\big(\phi \tilde{f}_1; f_2; A_n^1\big); \\ V_{\mathrm{B}}\big(g_*; f_2; A_n^1\big) &\le \varepsilon_* V_{\mathrm{B}}\big(\sigma \tilde{f}_1; f_2; A_n^1\big) + (1 - \varepsilon_*) V_{\mathrm{B}}\big(\phi \tilde{f}_1; f_2; A_n^1\big). \end{aligned}$$

Noting $\mu(f_1)\varepsilon^* = (1 - \mu(f_1))\varepsilon_*$, we add $\mu(f_1)$ times the first inequality to $1 - \mu(f_1)$ times the second and get

$$\begin{aligned} &\mu(f_1) V_{\mathrm{B}}\big(g^*; f_2; A_n^1\big) + \big(1 - \mu(f_1)\big) V_{\mathrm{B}}\big(g_*; f_2; A_n^1\big) \\ &\quad \le \mu(f_1) V_{\mathrm{B}}\big(\sigma \tilde{f}_1; f_2; A_n^1\big) + \big(1 - \mu(f_1)\big) V_{\mathrm{B}}\big(\phi \tilde{f}_1; f_2; A_n^1\big). \end{aligned} \tag{28}$$

The density g^* is simply

$$g^*(p) = \left[\frac{p(1 - \varepsilon^*)}{\mu(f_1)} + \frac{(1 - p)\varepsilon^*}{1 - \mu(f_1)}\right] \tilde{f}_1(p).$$

It is easy to check $(1 - \varepsilon^*)/\mu(f_1) \ge \varepsilon^*/(1 - \mu(f_1))$, which leads to

$$\sigma f_1 \le_{\mathrm{lc}} \sigma \tilde{f}_1 \le_{\mathrm{lc}} g^*.$$

Moreover, σf_1 and g^* have the same mean. By the induction hypothesis, we have

$$V_{\mathrm{B}}\big(\sigma f_1; f_2; A_n^1\big) \le V_{\mathrm{B}}\big(g^*; f_2; A_n^1\big). \tag{29}$$

Similarly,

$$V_{\mathrm{B}}\big(\phi f_1; f_2; A_n^1\big) \le V_{\mathrm{B}}\big(g_*; f_2; A_n^1\big). \tag{30}$$

We combine (28)–(30) to get

$$\begin{aligned} &\mu(f_1) V_{\mathrm{B}}\big(\sigma f_1; f_2; A_n^1\big) + \big(1 - \mu(f_1)\big) V_{\mathrm{B}}\big(\phi f_1; f_2; A_n^1\big) \\ &\quad \le \mu(f_1) V_{\mathrm{B}}\big(\sigma \tilde{f}_1; f_2; A_n^1\big) + \big(1 - \mu(f_1)\big) V_{\mathrm{B}}\big(\phi \tilde{f}_1; f_2; A_n^1\big). \end{aligned}$$

Applying (27) then yields

$$V_{\mathrm{B}}^1(f_1; f_2; A_n) \le V_{\mathrm{B}}^1(\tilde{f}_1; f_2; A_n).$$

The rest of the proof is standard. □

Remark. Theorem 3 focuses on the parameter p. If we still require equal prior means for p, but impose the log-concavity order on $\theta = \log(p/(1-p))$ rather than p, then V_{B} is ordered by virtually the same proof. This result is distinct from Theorem 3 because the relative log-concavity order is usually not preserved by monotone transformations.

2.6. Normal bandits with general priors

The main result of this subsection (Theorem 4) extends Theorem 2 to general priors for normal bandits. Similar to Theorems 3, 4 is based on the relative log-concavity order, although it is more restrictive because we only compare a general prior with a normal prior.

Given $\theta_i, i = 1, 2$, let us assume that observations from arm i are i.i.d. $\mathrm{N}(\theta_i, 1)$. Priors on θ_i are independent with Lebesgue densities f_i. We shall denote the value of this normal bandit with discount sequence A_n by $V_{\mathrm{N}}(f_1; f_2; A_n)$. Denote the mean of any f by $\mu(f) = \int_{-\infty}^{\infty} \theta f(\theta)\,\mathrm{d}\theta$.

Theorem 4. *Let* $\tilde{f}_1 \equiv \mathrm{N}(\alpha, 1/\tau)$.

1. *If* $f_1 \le_{\mathrm{lc}} \tilde{f}_1$ *and* $\mu(f_1) \le \alpha$, *then* $V_{\mathrm{N}}(f_1; f_2; A_n) \le V_{\mathrm{N}}(\tilde{f}_1; f_2; A_n)$.
2. *If* $\tilde{f}_1 \le_{\mathrm{lc}} f_1$ *and* $\mu(f_1) \ge \alpha$, *then* $V_{\mathrm{N}}(\tilde{f}_1; f_2; A_n) \le V_{\mathrm{N}}(f_1; f_2; A_n)$.

Let $\Lambda_{\mathrm{N}}(f; A_n)$ denote the break-even value of a one-armed normal bandit with prior f for the mean of the unknown arm. We obtain Corollary 4 as a consequence of Theorem 4 and Lemma 1.

Corollary 4. *Assume A_n is regular and $a_1 > 0$. Define* $\tilde{f} \equiv \mathrm{N}(\alpha, 1/\tau)$.

1. *If* $f \le_{\mathrm{lc}} \tilde{f}$ *and* $\mu(f) \le \alpha$, *then* $\Lambda_{\mathrm{N}}(f; A_n) \le \Lambda_{\mathrm{N}}(\tilde{f}; A_n)$.
2. *If* $\tilde{f} \le_{\mathrm{lc}} f$ *and* $\mu(f) \ge \alpha$, *then* $\Lambda_{\mathrm{N}}(\tilde{f}; A_n) \le \Lambda_{\mathrm{N}}(f; A_n)$.

The condition $f \le_{\mathrm{lc}} \mathrm{N}(\alpha, 1/\tau)$ is essentially $\mathrm{d}^2 \log f(\theta)/\mathrm{d}\theta^2 \le -\tau$, which can be regarded as a strong form of information ordering. The appearance of $\le_{\mathrm{lc}}$ is therefore especially intuitive in Theorem 4 and Corollary 4. It is an open problem whether Theorem 4 and Corollary 4 hold without assuming that one of the priors is normal.

The rest of this section proves Theorem 4. We need a technical result (Lemma 3) which may be of independent interest.

Lemma 3. *Let g be a differentiable function on $\mathbb{R}$. Assume X is a random variable satisfying* $Eg(X) = EX$.

1. *If* $0 \le g'(x) \le 1, x \in \mathbb{R}$, *then* $g(X) \le_{\mathrm{cx}} X$.
2. *If* $g'(x) \ge 1, x \in \mathbb{R}$, *then* $X \le_{\mathrm{cx}} g(X)$.

Proof. We prove Part 1 only. Part 2 follows from Part 1 by considering the inverse function of g. As $Eg(X) = EX$, one criterion for $g(X) \le_{\mathrm{cx}} X$ is

$$E \max\bigl\{0, g(X) - b\bigr\} \le E \max\{0, X - b\}, \qquad b \in \mathbb{R}. \tag{31}$$

See, for example, [22], Theorem 3.A.1. Let us assume $0 \le g'(x) \le c$ for some $0 < c < 1$. Otherwise we consider $cg(x) + (1-c)E(X)$ and let $c \uparrow 1$. As $g(x)$ is a contraction, it has a unique fixed point, say x_0. Consider two cases.

Case (i): $b \ge x_0$. If $x \ge x_0$, then $g(x) - g(x_0) \le x - x_0$, that is, $g(x) \le x$, and $\max\{0, g(x) - b\} \le \max\{0, x - b\}$. If $x < x_0$, then $g(x) \le g(x_0) = x_0$ and $\max\{0, g(x) - b\} = 0 = \max\{0, x - b\}$. In either case, $\max\{0, g(x) - b\} \le \max\{0, x - b\}$, which implies (31).

Case (ii): $b < x_0$. Applying the argument of Case (i) to $\tilde{g}(x) \equiv -g(-x)$ and $\tilde{X} \equiv -X$ yields $E \max\{0, b - g(X)\} \le E \max\{0, b - X\}$, which reduces to (31) because $Eg(X) = EX$.

□

Proof of Theorem 4. We only prove Part 1; the second part is similar. If $\mu(f_1) < \alpha$ then by decreasing the mean of $\tilde{f}_1$ from α to $\mu(f_1)$ we preserve the log-concavity ordering and reduce the problem to the case of equal means. Let us assume $\mu(f_1) = \alpha$ throughout the proof.

The $n = 1$ case is easy. For $n \ge 2$ we use induction. The equations (4)–(6) become

$$\begin{aligned} V_{\mathrm{N}}(f_1; f_2; A_n) &= \max\bigl\{V_{\mathrm{N}}^1(f_1; f_2; A_n), V_{\mathrm{N}}^2(f_1; f_2; A_n)\bigr\}; \\ V_{\mathrm{N}}^1(f_1; f_2; A_n) &= a_1\mu(f_1) + E\bigl[V_{\mathrm{N}}\bigl(f_1^X; f_2; A_n^1\bigr)|\Phi f_1\bigr]; \\ V_{\mathrm{N}}^2(f_1; f_2; A_n) &= a_1\mu(f_2) + E\bigl[V_{\mathrm{N}}\bigl(f_1; f_2^Y; A_n^1\bigr)|\Phi f_2\bigr]. \end{aligned} \tag{32}$$

We denote the posterior $f_1^x(\theta) \propto f_1(\theta)\exp[-(x-\theta)^2/2]$; similarly for f_2^y. In $E[g(X)|\Phi f]$, the density of X, denoted by Φf, is the convolution of f with the standard normal. (Note the difference from the notation in Section 2.1.) Let $m(x; f)$ denote the posterior mean of θ when x is observed and the prior is f, that is, $m(x; f) = \int_{-\infty}^{\infty} \theta f^x(\theta)\,\mathrm{d}\theta$. Direct calculation yields

$$\frac{\mathrm{d}m(x; f)}{\mathrm{d}x} = \mathrm{Var}\bigl(\theta|f^x\bigr). \tag{33}$$

That is, the derivative of $m(x; f)$ is simply the posterior variance of θ.

Suppose $f_1 \le_{\mathrm{lc}} \tilde{f}_1 \equiv \mathrm{N}(\alpha, 1/\tau)$ and $\mu(f_1) = \alpha$. Then

$$f_1^x \le_{\mathrm{lc}} \mathrm{N}\left(m(x; f_1), \frac{1}{\tau+1}\right). \tag{34}$$

It can be shown that (i) if X is distributed as Φf_1, then $m(X; f_1) \le_{\mathrm{cx}} (X + \tau\alpha)/(\tau+1)$; (ii) Φf_1 is smaller than $\Phi \tilde{f}_1 \equiv \mathrm{N}(\alpha, 1 + 1/\tau)$ in the convex order. To prove (i), note that (34) holds with $\le_{\mathrm{lc}}$ replaced by $\le_{\mathrm{cx}}$ as the two sides have equal means. By (33), we have

$$0 \le \frac{\mathrm{d}m(x; f_1)}{\mathrm{d}x} \le \frac{1}{\tau+1}, \qquad x \in \mathbb{R}.$$

If X is distributed as Φf_1 then both $(X + \tau\alpha)/(\tau+1)$ and $m(X; f_1)$ have mean $\mu(f_1) = \alpha$. Thus claim (i) holds by Lemma 3. Claim (ii) holds because $f_1 \le_{\mathrm{cx}} \tilde{f}_1$ and the convex order is closed under convolution.

We have

$$\begin{aligned}
E\big[V_{\mathrm{N}}\big(f_1^X; f_2; A_n^1\big)|\Phi f_1\big] &\le E\left[V_{\mathrm{N}}\left(\mathrm{N}\left(m(X; f_1), \frac{1}{\tau+1}\right); f_2; A_n^1\right)\Big|\Phi f_1\right] \\
&\le E\big[V_{\mathrm{N}}\big(\tilde{f}_1^X; f_2; A_n^1\big)\big|\Phi f_1\big] \\
&\le E\big[V_{\mathrm{N}}\big(\tilde{f}_1^X; f_2; A_n^1\big)\big|\Phi \tilde{f}_1\big],
\end{aligned}$$

where the first inequality holds by (34) and the induction hypothesis, the second by claim (i), noting

$$\tilde{f}_1^X = \mathrm{N}\left(\frac{X+\tau\alpha}{\tau+1}, \frac{1}{\tau+1}\right),$$

and the third by claim (ii). The last two inequalities also use the convexity of V_{N} with respect to the mean of a normal prior, that is, Proposition 2. (Although Proposition 2 assumes normal priors for both arms, this can be relaxed.) It follows from (32) that

$$V_{\mathrm{N}}^1(f_1; f_2; A_n) \le V_{\mathrm{N}}^1(\tilde{f}_1; f_2; A_n).$$

The rest of the proof is standard. □

Remark. We mention some related results of [25]. Given a density function $g(\theta)$, consider the location-scale family $f_{a,b}(\theta) = b^{-1}g((\theta-a)/b), -\infty < a < \infty, b > 0$. For the normal bandit, assume that observations from an arm with parameter θ are conditionally iid $\mathrm{N}(\theta, \sigma^2)$ with known noise variance σ^2. It follows from the location-scale structure that the break-even value satisfies

$$\Lambda_{(\mathrm{N},\sigma^2)}(a, b; A_n) = a + b\Lambda_{(\mathrm{N},\sigma^2/b^2)}(0, 1; A_n),$$

where (a, b) refers to the prior $f_{a,b}$ and the subscript (N, σ^2) refers to the normal bandit with noise variance σ^2. Furthermore, while Lemma 1 and Theorem 1 of [25] consider only normal priors with geometric discounting, the method works more generally, yielding that $\Lambda_{(\mathrm{N},\sigma^2)}(a, b; A_n)$ is nonincreasing in σ^2 and that $\Lambda_{(\mathrm{N},\sigma^2)}(0, b; A_n)/b$ is nondecreasing in b. In particular, $\Lambda_{(\mathrm{N},\sigma^2)}(a, b; A_n)$ is nondecreasing in a and b. It seems difficult to extend these results to non-normal bandits.

2.7. Discussion

Results in previous sections suggest the following conjecture. Consider a two-armed bandit in the general exponential family setting with conjugate priors. Suppose the prior expected yield of one pull from each arm is the same, but the prior weight of arm 1 is larger. Then it seems reasonable that arm 2 is optimal at the first stage, that is, in the notation of Section 2.2,

$$\frac{\gamma_1}{\tau_1} = \frac{\gamma_2}{\tau_2} \quad \text{and} \quad \tau_1 > \tau_2 \quad \Longrightarrow \quad \Delta(\gamma_1, \tau_1; \gamma_2, \tau_2; A_n) \le 0.$$

This holds if the discount sequence is infinite-horizon geometric. Indeed, it is optimal to pull arm 2 because, according to Corollary 2, arm 2 has a larger Gittins index. For non-geometric discounting, we cannot apply Corollary 2 due to the lack of an index policy. In fact, [2] proposed this conjecture for Bernoulli bandits with uniform discounting, and this special case is still open.

3. The nonparametric case with Dirichlet process priors

In the nonparametric case, associated with arms 1 and 2 are probability measures $P_i, i = 1, 2$, respectively. Observations from arm i are independent samples given P_i; observations from different arms are independent. The P_i's themselves are treated as random, with independent Dirichlet process priors. Specifically, $P_i \sim \mathrm{DP}(\alpha_{\mathrm{i}})$, where α_i is a finite nonnull measure with a finite first moment. It is often helpful to write $\alpha_i = M_i F_i$ where $M_i = \alpha_i(\mathbb{R})$ so that F_i is a probability distribution. We refer to F_i and M_i as the prior mean distribution and prior weight of the Dirichlet process, respectively. We use $(\alpha_1, \alpha_2; A_n)$ to denote such a Dirichlet bandit with discount sequence A_n.

3.1. Prior mean monotonicity

Let us denote the maximum expected payoff of a two-armed Dirichlet bandit $(\alpha_1, \alpha_2; A_n)$ by $W(\alpha_1, \alpha_2; A_n)$. Let $W^i(\alpha_1, \alpha_2; A_n)$ be the expected payoff when selecting arm i initially and using an optimal strategy thereafter. Then

$$W(\alpha_1, \alpha_2; A_n) = \max\bigl\{W^1(\alpha_1, \alpha_2; A_n), W^2(\alpha_1, \alpha_2; A_n)\bigr\}. \tag{35}$$

Suppose arm 1 is selected initially, resulting in an observation X. Because the prior on P_1 is a Dirichlet process, the posterior is again a Dirichlet process $\mathrm{DP}(\alpha_1 + \delta_X)$, where δ_x denotes a point mass at x. Thus we have

$$W^1(\alpha_1, \alpha_2; A_n) = a_1\mu_1 + E\bigl[W\bigl(\alpha_1 + \delta_X, \alpha_2; A_n^1\bigr)|\alpha_1\bigr], \tag{36}$$

$$W^2(\alpha_1, \alpha_2; A_n) = a_1\mu_2 + E\bigl[W\bigl(\alpha_1, \alpha_2 + \delta_Y; A_n^1\bigr)|\alpha_2\bigr], \tag{37}$$

where $A_n^1 = (a_2, a_3, \dots, a_n)$ and μ_i denotes the first moment of α_i, which is also the expected value of an observation from arm i. In $E[g(X)|\alpha]$, the distribution of X is α/M with $M = \alpha(\mathbb{R})$. The quantities W, W^1 and W^2 are well defined and finite as long as $\alpha_i, i = 1, 2$, have finite first moments, which we assume throughout Section 3.

Lemma 4 reveals a convexity property of W which we shall use repeatedly.

Lemma 4. *Let α be a finite measure on $\mathbb{R}$ with a finite mean. Then, for $u, v \in \mathbb{R}$ and $r > 0$, the function $W(\alpha + \rho\delta_u + (r - \rho)\delta_v, \alpha_2; A_n)$ is convex in $\rho \in [0, r]$.*

Proof. Let us use induction on n. It is easy to check that the claim holds for $n = 1$. For $n \geq 2$, we note that by (35) it suffices to show that each of $W^i(\alpha + \rho\delta_u + (r - \rho)\delta_v, \alpha_2; A_n), i = 1, 2$, is convex in $\rho \in [0, r]$. Since the mean of $\alpha + \rho\delta_u + (r-\rho)\delta_v$ is linear in ρ, by (36) and (37), we only need to show that both

$$E\big[W\big(\alpha + \rho\delta_u + (r-\rho)\delta_v + \delta_X, \alpha_2; A_n^1\big)|\alpha + \rho\delta_u + (r-\rho)\delta_v\big] \quad \text{and} \tag{38}$$

$$E\big[W\big(\alpha + \rho\delta_u + (r-\rho)\delta_v, \alpha_2 + \delta_Y; A_n^1\big)|\alpha_2\big] \tag{39}$$

are convex in ρ. Convexity of (39) follows from the induction hypothesis. To deal with (38), we directly compute

$$E\big[W\big(\alpha + \rho\delta_u + (r-\rho)\delta_v + \delta_X, \alpha_2; A_n^1\big)|\alpha + \rho\delta_u + (r-\rho)\delta_v\big]$$

$$= \frac{M}{M+r} E\big[W\big(\alpha + \rho\delta_u + (r-\rho)\delta_v + \delta_X, \alpha_2; A_n^1\big)|\alpha\big] \tag{40}$$

$$+ \frac{\rho\phi(\rho+1) + (r-\rho)\phi(\rho)}{M+r}, \tag{41}$$

where $M = \alpha(\mathbb{R})$ and

$$\phi(\rho) = W\big(\alpha + \rho\delta_u + (r+1-\rho)\delta_v, \alpha_2; A_n^1\big).$$

By the induction hypothesis, $\phi(\rho)$ is convex in $\rho \in [0, r+1]$. We claim that this implies that $\psi(\rho) \equiv \rho\phi(\rho+1) + (r-\rho)\phi(\rho)$ is convex in $\rho \in [0, r]$. In fact, if $\phi(\rho)$ is twice differentiable, then we have

$$\psi''(\rho) = 2\big(\phi'(\rho+1) - \phi'(\rho)\big) + \rho\phi''(\rho+1) + (r-\rho)\phi''(\rho) \geq 0, \qquad \rho \in [0, r],$$

by the convexity of ϕ. A standard limiting argument shows that $\psi(\rho)$ is convex in $\rho \in [0, r]$ as long as $\phi(\rho)$ is convex in $\rho \in [0, r+1]$ without assuming differentiability. Hence the second term (41) is convex. The first term (40) is convex in $\rho \in [0, r]$ by the induction hypothesis, since in this expectation X is distributed according to α/M independently of ρ. Thus the convexity of (38) is established. □

Theorem 5 says that the value of the bandit increases as the mean of the Dirichlet process prior for any arm becomes stochastically larger and more dispersed. This strengthens Proposition 2.2 of [10] who consider the usual stochastic order rather than the increasing convex order.

Theorem 5. *If $M > 0$ and $F \leq_{\text{icx}} \tilde{F}$, both with finite means, then*

$$W(MF, \alpha_2; A_n) \leq W(M\tilde{F}, \alpha_2; A_n).$$

Proof. Let us use induction. The claim obviously holds for $n = 1$. For $n \geq 2$ we have $W^2(MF, \alpha_2; A_n) \leq W^2(M\tilde{F}, \alpha_2; A_n)$ by (37) and the induction hypothesis. Moreover,

$$W^1(MF, \alpha_2; A_n) = a_1 E(X|F) + E\big[W\big(MF + \delta_X, \alpha_2; A_n^1\big)|F\big]$$

$$\leq a_1 E(X|\tilde{F}) + E\big[W\big(M\tilde{F} + \delta_X, \alpha_2; A_n^1\big)|F\big]$$

$$\begin{aligned}&\le a_1 E(X|\tilde{F}) + E\big[W\big(M\tilde{F} + \delta_X, \alpha_2; A_n^1\big)|\tilde{F}\big]\\ &= W^1(M\tilde{F}, \alpha_2; A_n),\end{aligned}$$

where the first inequality follows from $F \le_{\mathrm{icx}} \tilde{F}$ and the induction hypothesis, noting that $(MF + \delta_x)/(M+1) \le_{\mathrm{icx}} (M\tilde{F} + \delta_x)/(M+1)$ for any x; the second inequality holds by the definition of $\le_{\mathrm{icx}}$, because $W(M\tilde{F} + \delta_x, \alpha_2; A_n^1)$ is an increasing, convex function of x. To show this, fix $-\infty < u < v < \infty$. It is easy to show $(M\tilde{F} + \delta_u)/(M+1) \le_{\mathrm{icx}} (M\tilde{F} + \delta_v)/(M+1)$, which, by the induction hypothesis, implies $W(M\tilde{F} + \delta_u, \alpha_2; A_n^1) \le W(M\tilde{F} + \delta_v, \alpha_2; A_n^1)$. Moreover, for fixed $\rho \in (0, 1)$, we have

$$\begin{aligned}&\rho W\big(M\tilde{F} + \delta_u, \alpha_2; A_n^1\big) + (1-\rho) W\big(M\tilde{F} + \delta_v, \alpha_2; A_n^1\big)\\ &\quad\ge W\big(M\tilde{F} + \rho\delta_u + (1-\rho)\delta_v, \alpha_2; A_n^1\big)\\ &\quad\ge W\big(M\tilde{F} + \delta_{\rho u + (1-\rho)v}, \alpha_2; A_n^1\big),\end{aligned}$$

where the first inequality follows from Lemma 4, and the second inequality holds by the induction hypothesis, noting that

$$\frac{M\tilde{F} + \delta_{\rho u + (1-\rho)v}}{M+1} \le_{\mathrm{icx}} \frac{M\tilde{F} + \rho\delta_u + (1-\rho)\delta_v}{M+1}.$$

Hence $W(M\tilde{F} + \delta_x, \alpha_2; A_n^1)$ is convex in x as needed. □

When arm 2 has a known distribution P_2 with mean λ, the problem reduces to a one-armed bandit. Without loss of generality we may assume the known arm yields a constant payoff λ at each stage, i.e., we consider the $(\alpha, \delta_\lambda; A_n)$ bandit (the subscript on α_1 is dropped for convenience). Similar to the parametric case, assuming the discount sequence is regular, this one-armed bandit is an optimal stopping problem. If A_n is regular and $a_1 > 0$, then there exists a break-even value $\Lambda(\alpha; A_n)$ for the $(\alpha, \delta_\lambda; A_n)$ bandit, such that arm 1 is optimal initially if and only if $\lambda \le \Lambda(\alpha; A_n)$ and arm 2 is optimal initially if and only if $\lambda \ge \Lambda(\alpha; A_n)$. The following result is the nonparametric counterpart of Lemma 1, and is stated for uniform discounting as Lemma 2.1 in [10].

Lemma 5. *If A_n is regular and $a_1 > 0$, then $\Lambda(\alpha; A_n)$ is the smallest λ such that $W(\alpha, \delta_\lambda; A_n) \le \lambda \sum_{i=1}^n a_i$.*

Lemma 5 and Theorem 5 yield the following result comparing $\Lambda(\alpha; A_n)$.

Corollary 5. *For $M > 0$ and $F \le_{\mathrm{icx}} \tilde{F}$, both with finite means, we have $\Lambda(MF; A_n) \le \Lambda(M\tilde{F}; A_n)$, assuming A_n is regular and $a_1 > 0$.*

Suppose A_n is regular and $a_1 > 0$. Analogous to the parametric case, one can show (see [10]) that, for the $(\alpha, \delta_\lambda; A_n)$ bandit there exists a break-even observation $b(\alpha; A_n)$ such that if both arms are optimal initially, and an observation x is taken from arm 1, then arm 1 remains optimal if $x \geq b(\alpha; A_n)$ and arm 2 becomes optimal if $x \leq b(\alpha; A_n)$. Note that $b(\alpha; A_n)$ is a real number not necessarily in the support of α. That is,

$$\begin{aligned} \Lambda(\alpha; A_n) &\geq \Lambda\big(\alpha + \delta_x; A_n^1\big), \qquad && \text{if } x \leq b(\alpha; A_n); \\ \Lambda(\alpha; A_n) &\leq \Lambda\big(\alpha + \delta_x; A_n^1\big), \qquad && \text{if } x \geq b(\alpha; A_n). \end{aligned}$$

Calculating this break-even observation is nontrivial. In the case of uniform discounting, [10] prove an upper bound for $b(\alpha; A_n)$ and conjecture that $b(\alpha; A_n) \geq \Lambda(\alpha; A_n)$ based on numerical evidence. We confirm this in Proposition 4.

Proposition 4. *Suppose A_n is regular, $n \geq 2$, and $a_1, a_2 > 0$. Then $b(\alpha; A_n) \geq \Lambda(\alpha; A_n)$.*

As noted by [4]; page 131, Proposition 4 has an intuitive interpretation. Suppose both arms are optimal initially, and arm 1 is selected. If the initial pull on arm 1 yields no more than $\Lambda(\alpha; A_n)$, which is the yield of arm 2 per pull, the hope of getting higher payoff fades. Not surprisingly, arm 2 becomes optimal afterwards. This suggests that the break-even observation is at least $\Lambda(\alpha; A_n)$.

To prove Proposition 4, we need a lemma.

Lemma 6. *For $c > 0$, $\lambda \in \mathbb{R}$ and an arbitrary discount sequence A_n, we have*

$$W(\alpha + c\delta_\lambda, \delta_\lambda; A_n) \leq W(\alpha, \delta_\lambda; A_n).$$

Proof. We use induction on n. The $n = 1$ case is easy. Suppose $n \geq 2$. Let us write $M = \alpha(\mathbb{R})$ and let μ be the first moment of α. Direct calculation using (35)–(37) yields

$$W(\alpha + c\delta_\lambda, \delta_\lambda; A_n) = \max\left\{\frac{M\phi_0 + c\phi_1}{M + c}, \phi_2\right\}, \tag{42}$$

where

$$\begin{aligned} \phi_0 &= a_1\mu + E\big[W\big(\alpha + c\delta_\lambda + \delta_X, \delta_\lambda; A_n^1\big)|\alpha\big]; \\ \phi_1 &= a_1\lambda + W\big(\alpha + (c+1)\delta_\lambda, \delta_\lambda; A_n^1\big); \\ \phi_2 &= a_1\lambda + W\big(\alpha + c\delta_\lambda, \delta_\lambda; A_n^1\big). \end{aligned}$$

Applying the induction hypothesis, and then (35) and (36), we get

$$\begin{aligned} \phi_0 &\leq a_1\mu + E\big[W\big(\alpha + \delta_X, \delta_\lambda; A_n^1\big)|\alpha\big] \\ &\leq W(\alpha, \delta_\lambda; A_n). \end{aligned}$$

Applying the induction hypothesis, and then (35) and (37), we get

$$\phi_1 \leq \phi_2 \leq a_1\lambda + W\big(\alpha, \delta_\lambda; A_n^1\big) \leq W(\alpha, \delta_\lambda; A_n).$$

That is, $\phi_i \leq W(\alpha, \delta_\lambda; A_n)$ for $i = 0, 1, 2$. Hence the claim holds by (42). □

Proof of Proposition 4. Suppose $\lambda = \Lambda(\alpha; A_n)$. By the optimal stopping characterization, we have $W(\alpha, \delta_\lambda; A_n^1) = \lambda \sum_{i=2}^n a_i$. Lemma 6 yields $W(\alpha + \delta_\lambda, \delta_\lambda; A_n^1) \le \lambda \sum_{i=2}^n a_i$. It follows from Lemma 5 that $\lambda \ge \Lambda(\alpha + \delta_\lambda; A_n^1)$. That is, $\Lambda(\alpha; A_n) \ge \Lambda(\alpha + \delta_\lambda; A_n^1)$, which implies $\lambda \le b(\alpha; A_n)$ (under the assumptions one can show that $b(\alpha; A_n)$ is unique). □

3.2. Prior weight monotonicity

The main result of this subsection (Theorem 6) shows that the maximum expected payoff of a bandit decreases as the prior weight for the Dirichlet process prior of an arm increases. When arm 2 is known and the discount sequence is regular, this shows that the break-even value $\Lambda(M_1 F_1; A_n)$ decreases as M_1 (the prior weight associated with arm 1) increases. That is, given the same immediate payoff, arm 1 becomes less desirable as the amount of information about it increases. Theorem 6 is the nonparametric counterpart of Theorem 2.

Theorem 6. *Let F be a probability distribution on $\mathbb{R}$ with a finite mean. If $0 < M < \tilde{M}$ then*

$$W(MF, \alpha_2; A_n) \ge W(\tilde{M}F, \alpha_2; A_n). \tag{43}$$

Lemma 5 and Theorem 6 yield the following result concerning the break-even value $\Lambda(\alpha; A_n)$ for the one armed bandit $(\alpha, \delta_\lambda; A_n)$, as conjectured by [10] in the case of uniform discounting.

Corollary 6. *For $0 < M < \tilde{M}$ we have $\Lambda(MF; A_n) \ge \Lambda(\tilde{M}F; A_n)$, assuming A_n is regular and $a_1 > 0$.*

When F has only two support points, Corollary 2 says that for a Bernoulli one-armed bandit with a Beta(Mu, Mv) prior, $u, v > 0$, for the unknown arm, the break-even value decreases in M. This Bernoulli case was proved by [15] for infinite-horizon geometric discounting.

The rest of this section gives a proof of Theorem 6. We assume F has finite, and then bounded, and finally arbitrary, support. The key step is summarized as Lemma 7.

Lemma 7. *Assume $n \ge 2$, $L > 0$. Assume α is a finite measure on $\mathbb{R}$ with a finite mean and F is a probability distribution on $\mathbb{R}$ with $s < \infty$ support points. Then $E[W(\alpha + \theta F + (L - \theta)\delta_X, \alpha_2; A_n^1)|F]$ decreases in $\theta \in [0, L]$.*

Proof. We use induction on s. Although the induction may start at the trivial case $s = 1$, we present the $s = 2$ case to illustrate the convexity arguments. Write $F = p\delta_1 + (1 - p)\delta_0$ where $p \in (0, 1)$ and $\{0, 1\}$ are the support points without loss of generality. For fixed $0 \le \theta_1 < \theta_2 \le L$, let $Z \sim$ Bernoulli(p) and define

$$Z_i = \theta_i p + (L - \theta_i)Z, \qquad i = 1, 2.$$

Then $EZ_1 = EZ_2 = pL$, and it is easy to verify $Z_2 \le_{\mathrm{cx}} Z_1$ as $\theta_1 < \theta_2$ (see, e.g., [22], Theorem 3.A.18). Let us define

$$\phi(u) = W\big(\alpha + u\delta_1 + (L - u)\delta_0, \alpha_2; A_n^1\big).$$

By direct calculation

$$\begin{aligned}
E\big[W\big(\alpha+\theta_1 F+(L-\theta_1)\delta_X,\alpha_2;A_n^1\big)|F\big] &= p\phi(\theta_1 p+L-\theta_1)+(1-p)\phi(\theta_1 p)\\
&= E\phi(Z_1)\\
&\ge E\phi(Z_2)\\
&= E\big[W\big(\alpha+\theta_2 F+(L-\theta_2)\delta_X,\alpha_2;A_n^1\big)|F\big],
\end{aligned}$$

where the inequality holds because $Z_2 \le_{\mathrm{cx}} Z_1$ and, by Lemma 4, $\phi(u)$ is convex in $u\in[0,L]$.

For $s\ge 3$, write $F=\sum_{j=1}^s p_j\delta_{x_j}$, where $\{x_j, j=1,\dots,s\}$ are the support points, $p_j>0$ and $\sum_{j=1}^s p_j=1$. Consider the leave-one-out distributions

$$F^k=\sum_{j\neq k}\frac{p_j}{1-p_k}\delta_{x_j},\qquad k=1,\dots,s.$$

Denote $W(\gamma)=W(\gamma,\alpha_2;A_n^1)$ for convenience. For fixed $0\le\theta_1<\theta_2\le L$, we have

$$\begin{aligned}
&(s-1)E\big[W\big(\alpha+\theta_1 F+(L-\theta_1)\delta_X\big)|F\big]\\
&=\sum_{k=1}^s(1-p_k)E\big[W\big(\alpha+\theta_1 F+(L-\theta_1)\delta_X\big)|F^k\big]\\
&=\sum_{k=1}^s(1-p_k)E\big[W\big(\alpha+\theta_1 p_k\delta_{x_k}+\theta_1(1-p_k)F^k+(L-\theta_1)\delta_X\big)|F^k\big] \qquad (44)\\
&\ge\sum_{k=1}^s(1-p_k)E\big[W\big(\alpha+\theta_1 p_k\delta_{x_k}+\theta_2(1-p_k)F^k+\big(L-\theta_2(1-p_k)-\theta_1 p_k\big)\delta_X\big)|F^k\big]\\
&=\sum_{k=1}^s\sum_{j\neq k}p_j V_{jk},
\end{aligned}$$

where

$$\begin{aligned}
V_{jk}&=W\big(\alpha+\theta_2\gamma^{jk}+\theta_1 p_k\delta_{x_k}+\big(L-\theta_2(1-p_k-p_j)-\theta_1 p_k\big)\delta_{x_j}\big),\\
\gamma^{jk}&=\sum_{l\neq j,k}p_l\delta_{x_l},\qquad j\neq k.
\end{aligned}$$

The inequality (44) follows from the induction hypothesis; other steps are algebraic manipulations.

For fixed $j\neq k$, let $Z\sim\mathrm{Bernoulli}(p_k/(p_j+p_k))$ and define

$$\begin{aligned}
Z_1&=\theta_1 p_k+Z\big(L-\theta_2+(\theta_2-\theta_1)(p_j+p_k)\big);\\
Z_2&=\theta_2 p_k+Z(L-\theta_2).
\end{aligned}$$

It is easy to verify that

$$EZ_1 = EZ_2; \qquad Z_2 \le_{\mathrm{cx}} Z_1.$$

We have

$$\begin{aligned} p_j V_{jk} + p_k V_{kj} &= (p_j + p_k) EW\bigl(\alpha + \theta_2 \gamma^{jk} + Z_1 \delta_{x_k} + \bigl(L - \theta_2(1 - p_k - p_j) - Z_1\bigr)\delta_{x_j}\bigr) \\ &\ge (p_j + p_k) EW\bigl(\alpha + \theta_2 \gamma^{jk} + Z_2 \delta_{x_k} + \bigl(L - \theta_2(1 - p_k - p_j) - Z_2\bigr)\delta_{x_j}\bigr) \\ &= p_j W\bigl(\alpha + \theta_2 F + (L - \theta_2)\delta_{x_j}\bigr) + p_k W\bigl(\alpha + \theta_2 F + (L - \theta_2)\delta_{x_k}\bigr), \end{aligned}$$

where the inequality holds by Lemma 4 as $Z_2 \le_{\mathrm{cx}} Z_1$. Hence,

$$\begin{aligned} \sum_{k=1}^{s} \sum_{j \ne k} p_j V_{jk} &= \sum_{1 \le j < k \le s} (p_j V_{jk} + p_k V_{kj}) \\ &\ge \sum_{1 \le j < k \le s} \bigl[p_j W\bigl(\alpha + \theta_2 F + (L - \theta_2)\delta_{x_j}\bigr) + p_k W\bigl(\alpha + \theta_2 F + (L - \theta_2)\delta_{x_k}\bigr)\bigr] \\ &= (s-1) \sum_{j=1}^{s} p_j W\bigl(\alpha + \theta_2 F + (L - \theta_2)\delta_{x_j}\bigr) \\ &= (s-1) E\bigl[W\bigl(\alpha + \theta_2 F + (L - \theta_2)\delta_X\bigr) | F\bigr]. \end{aligned}$$

Thus, we have shown that $E[W(\alpha + \theta F + (L - \theta)\delta_X)|F]$ decreases in $\theta \in [0, L]$. □

Proof of Theorem 6. (i) Assume F has finite support. The claim obviously holds for $n = 1$. For $n \ge 2$ we use induction. In view of (35)–(37), we only need to show

$$E\bigl[W\bigl(MF + \delta_X, \alpha_2; A_n^1\bigr)|F\bigr] \ge E\bigl[W\bigl(\tilde{M}F + \delta_X, \alpha_2; A_n^1\bigr)|F\bigr] \quad \text{and} \tag{45}$$

$$E\bigl[W\bigl(MF, \alpha_2 + \delta_Y; A_n^1\bigr)|\alpha_2\bigr] \ge E\bigl[W\bigl(\tilde{M}F, \alpha_2 + \delta_Y; A_n^1\bigr)|\alpha_2\bigr]. \tag{46}$$

By the induction hypothesis, (46) holds. Define $\eta = (\tilde{M} + 1)/(M + 1)$ and $\theta = \tilde{M}/\eta$. Noting $M < \theta < M + 1$, we may apply Lemma 7 and get

$$\begin{aligned} E\bigl[W\bigl(MF + \delta_X, \alpha_2; A_n^1\bigr)|F\bigr] &\ge E\bigl[W\bigl(\theta F + (M + 1 - \theta)\delta_X, \alpha_2; A_n^1\bigr)|F\bigr] \\ &\ge E\bigl[W\bigl(\eta\bigl(\theta F + (M + 1 - \theta)\delta_X\bigr), \alpha_2; A_n^1\bigr)|F\bigr] \qquad (47) \\ &= E\bigl[W\bigl(\tilde{M}F + \delta_X, \alpha_2; A_n^1\bigr)|F\bigr], \end{aligned}$$

where (47) holds by the induction hypothesis, as $\eta > 1$. Thus (45) holds as required.

(ii) Assume F has bounded support. Then for arbitrary $\varepsilon > 0$ we can construct two distributions F^* and F_* supported on $\{x_1, \dots, x_s\}$ and $\{x_0, \dots, x_{s-1}\}$ respectively, where $x_j = x_0 + j\varepsilon$, such that $F(x_0)=0$, $F(x_s)=1$ and $F_*(x_{j-1})=F^*(x_j)=F(x_j)$, $j=1,\dots,s$. By construction, $F_* \le_{\mathrm{st}} F \le_{\mathrm{st}} F^*$. Theorem 5 yields

$$W(MF_*,\alpha_2;A_n)\le W(MF,\alpha_2;A_n)\le W\bigl(MF^*,\alpha_2;A_n\bigr).$$

Note that if $X\sim F^*$ then $X-\varepsilon\sim F_*$. Therefore the bandits $(MF^*,\alpha_2;A_n)$ and $(MF_*,\alpha_2;A_n)$ can be coupled in an obvious way such that, for every strategy of $(MF^*,\alpha_2;A_n)$, there exists a strategy of $(MF_*,\alpha_2;A_n)$ under which the payoff at each stage is either the same (when arm 2 is selected), or exactly ε less (when arm 1 is selected). Thus, we have shown

$$W\bigl(MF^*,\alpha_2;A_n\bigr)-W(MF_*,\alpha_2;A_n)\le\varepsilon\sum_{i=1}^n a_i.$$

Hence, $W(MF^*,\alpha_2;A_n)\to W(MF,\alpha_2;A_n)$ as $\varepsilon\to 0$, and the monotonicity of $W(MF^*,\alpha_2;A_n)$ with respect to M implies the corresponding monotonicity of $W(MF,\alpha_2;A_n)$.

(iii) Finally, assume F is an arbitrary distribution with a finite mean. Suppose $X\sim F$. For $L>0$ let $x_0\in(-L,L)$ be such that $F\{x_0\}=0$ and let F^* be the distribution of X^*, defined as $1_{|X|\le L}X+1_{|X|>L}x_0$. We construct a coupling between $(MF,\alpha_2;A_n)$ and $(MF^*,\alpha_2;A_n)$. Let X_k be the resulting observation when arm 1 of $(MF,\alpha_2;A_n)$ is pulled for the kth time. If $|X_1|\le L$ then let $X_1^*=X_1$, otherwise $X_1^*=x_0$, yielding $X_1^*\sim F^*$. For general $k\ge 1$, suppose $|X_i|\le L, i=1,\dots,k$, then let $X_{k+1}^*=X_{k+1}$ if $|X_{k+1}|\le L$ and $X_{k+1}^*=x_0$ otherwise. In this case the conditional distribution of X_{k+1} given $X_i, i=1,\dots,k$, is $(MF+\sum_{i=1}^k\delta_{X_i})/(M+k)$. Since $|X_i|\le L, i=1,\dots,k$, we have $X_i^*=X_i, i=1,\dots,k$, and the conditional distribution of X_{k+1}^* given $X_i^*, i=1,\dots,k$, is precisely $(MF^*+\sum_{i=1}^k\delta_{X_i^*})/(M+k)$. That is, $X_i^*, i=1,\dots,k+1$, can be regarded as successive pulls from arm 1 of $(MF^*,\alpha_2;A_n)$ as long as $|X_i|\le L, i=1,\dots,k$. Let the kth pull from arm 2 be Y_k for both bandits. In the event that all $|X_i|\le L, i=1,\dots,n$, the optimal strategy for $(MF,\alpha_2;A_n)$ can be adopted for $(MF^*,\alpha_2;A_n)$ throughout, yielding identical pulls (not all $X_i, i=1,\dots,n$, are realized). By considering a trivial upper (respectively, lower) bound for the payoff of $(MF,\alpha_2;A_n)$ (respectively, $(MF^*,\alpha_2;A_n)$) when at least one $|X_i|>L$, we have

$$\begin{aligned}W(MF,\alpha_2;A_n)-W\bigl(MF^*,\alpha_2;A_n\bigr)&\le E\Biggl[1_{\bigcup_{i=1}^n\{|X_i|>L\}}\sum_{i=1}^n\bigl(a_i\bigl(|Y_i|+|X_i|\bigr)-a_i\bigl(-|Y_i|-L\bigr)\bigr)\Biggr]\\&\le E\Biggl[1_{\bigcup_{i=1}^n\{|X_i|>L\}}\sum_{i=1}^n a^*\bigl(2|Y_i|+|X_i|+L\bigr)\Biggr]\\&\le E\Biggl[\Biggl(\sum_{i=1}^n 1_{\{|X_i|>L\}}\Biggr)\sum_{i=1}^n a^*\bigl(2|Y_i|+|X_i|+L\bigr)\Biggr]\\&\equiv a^*h(L),\end{aligned}$$

where $a^*\equiv\max_{i=1}^n a_i$. Direct calculation using exchangeability yields

$$h(L)=n^2\Pr\bigl(|X_1|>L\bigr)\bigl(2E|Y_1|+L\bigr)+nE\bigl[1_{|X_1|>L}|X_1|\bigr]+n(n-1)E\bigl[1_{|X_1|>L}|X_2|\bigr].$$

The first two terms tend to zero as $L \to \infty$ by dominated convergence since $E|X_1| < \infty$. For the last term, by conditioning on X_1 we have

$$E\big[1_{|X_1|>L}|X_2|\big] = E\left[1_{|X_1|>L}\left(\frac{M}{M+1}E|X| + \frac{1}{M+1}|X_1|\right)\right],$$

which also vanishes as $L \to \infty$. Thus

$$\limsup_{L\to\infty}\big[W(MF, \alpha_2; A_n) - W\big(MF^*, \alpha_2; A_n\big)\big] \le 0.$$

By a parallel argument, we get $\liminf_{L\to\infty}[W(MF, \alpha_2; A_n) - W(MF^*, \alpha_2; A_n)] \ge 0$. Thus $W(MF^*, \alpha_2; A_n)$ tends to $W(MF, \alpha_2; A_n)$ as $L \to \infty$, and the monotonicity of $W(MF, \alpha_2; A_n)$ with respect to M is proved as before. □

Remark. Clayton and Berry [10] also conjecture that the monotonicity in Corollary 6 is strict if $n \ge 2$, $A_n = (1, 1, \dots, 1)$, and F is nondegenerate. This can be confirmed by a careful analysis of the above results. Some modifications are needed. Using arguments similar to steps (ii) and (iii) in the proof of Theorem 6, we can first establish that Lemma 7 holds without the finite support restriction. Directly applying this strengthened Lemma 7 shows that (43) holds with strict inequality assuming $n \ge 2$, $A_n = (1, 1, \dots, 1)$, F is nondegenerate, and arm 1 is optimal initially in $(\tilde{M}F, \alpha_2; A_n)$. Under such conditions, the strictness of the inequality holds by induction as one key step (47) holds with strict inequality. It follows that Corollary 6 can be strengthened to strict monotonicity assuming uniform discounting, $n \ge 2$, and a nondegenerate F.

References

[1] Bellman, R. (1956). A problem in the sequential design of experiments. *Sankhyā* **16** 221–229. MR0079386

[2] Berry, D.A. (1972). A Bernoulli two-armed bandit. *Ann. Math. Statist.* **43** 871–897. MR0305531

[3] Berry, D.A. and Fristedt, B. (1979). Bernoulli one-armed bandits—arbitrary discount sequences. *Ann. Statist.* **7** 1086–1105. MR0536512

[4] Berry, D.A. and Fristedt, B. (1985). *Bandit Problems: Sequential Allocation of Experiments. Monographs on Statistics and Applied Probability*. London: Chapman & Hall. MR0813698

[5] Bradt, R.N., Johnson, S.M. and Karlin, S. (1956). On sequential designs for maximizing the sum of *n* observations. *Ann. Math. Statist.* **27** 1060–1074. MR0087288

[6] Brown, L.D. (1986). *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory. Institute of Mathematical Statistics Lecture Notes—Monograph Series*, 9. Hayward, CA: IMS. MR0882001

[7] Chattopadhyay, M.K. (1994). Two-armed Dirichlet bandits with discounting. *Ann. Statist.* **22** 1212–1221. MR1311973

[8] Chernoff, H. (1968). Optimal stochastic control. *Sankhyā Ser. A* **30** 221–252. MR0241149

[9] Chernoff, H. and Petkau, A.J. (1986). Numerical solutions for Bayes sequential decision problems. *SIAM J. Sci. Statist. Comput.* **7** 46–59. MR0819456

[10] Clayton, M.K. and Berry, D.A. (1985). Bayesian nonparametric bandits. *Ann. Statist.* **13** 1523–1534. MR0811507

[11] Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230. MR0350949

[12] Gittins, J.C. (1979). Bandit processes and dynamic allocation indices. *J. Roy. Statist. Soc. Ser. B* **41** 148–177. MR0547241

[13] Gittins, J.C., Glazebrook, K.D. and Weber, R.R. (2011). *Multi-Armed Bandit Allocation Indices*, 2nd ed. New York: Wiley.

[14] Gittins, J.C. and Jones, D.M. (1974). A dynamic allocation index for the sequential design of experiments. In *Progress in Statistics* (*European Meeting Statisticians*, *Budapest*, 1972) (J. Gani, ed.) 241–266. North-Holland: Amsterdam. MR0370964

[15] Gittins, J. and Wang, Y.-G. (1992). The learning component of dynamic allocation indices. *Ann. Statist.* **20** 1625–1636. MR1186269

[16] Herschkorn, S.J. (1997). Bandit bounds from stochastic variability extrema. *Statist. Probab. Lett.* **35** 283–288. MR1484965

[17] Karlin, S. (1968). *Total Positivity. Vol. I*. Stanford, CA: Stanford Univ. Press. MR0230102

[18] Kaspi, H. and Mandelbaum, A. (1998). Multi-armed bandits in discrete and continuous time. *Ann. Appl. Probab.* **8** 1270–1290. MR1661180

[19] Marshall, A.W. and Olkin, I. (1979). *Inequalities*: *Theory of Majorization and Its Applications*. *Mathematics in Science and Engineering* **143**. New York: Academic Press. MR0552278

[20] Müller, A. and Stoyan, D. (2002). *Comparison Methods for Stochastic Models and Risks*. *Wiley Series in Probability and Statistics*. Chichester: Wiley. MR1889865

[21] Rieder, U. and Wagner, H. (1991). Structured policies in the sequential design of experiments. *Ann. Oper. Res.* **32** 165–188. MR1128177

[22] Shaked, M. and Shanthikumar, J.G. (2007). *Stochastic Orders*. *Springer Series in Statistics*. New York: Springer. MR2265633

[23] Whitt, W. (1985). Uniform conditional variability ordering of probability distributions. *J. Appl. Probab.* **22** 619–633. MR0799285

[24] Whittle, P. (1980). Multi-armed bandits and the Gittins index. *J. Roy. Statist. Soc. Ser. B* **42** 143–149. MR0583348

[25] Yao, Y.-C. (2006). Some results on the Gittins index for a normal reward process. In *Time Series and Related Topics*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **52** 284–294. Beachwood, OH: IMS. MR2427855

[26] Yu, Y. (2009). On the entropy of compound distributions on nonnegative integers. *IEEE Trans. Inform. Theory* **55** 3645–3650. MR2598065

[27] Yu, Y. (2009). Monotonic convergence in an information-theoretic law of small numbers. *IEEE Trans. Inform. Theory* **55** 5412–5422. MR2597172

[28] Yu, Y. (2010). Relative log-concavity and a pair of triangle inequalities. *Bernoulli* **16** 459–470. MR2668910

Received April 2011 and revised May 2016