J. Appl. Probab. **61,** 121–136 (2024)
*doi:*10.1017/jpr.2023.24

A CONFIRMATION OF A CONJECTURE ON FELDMAN'S TWO-ARMED BANDIT PROBLEM

ZENGJING CHEN,*
YIWEI LIN,* AND
JICHEN ZHANG ,* ** *Shandong University*

Abstract

The myopic strategy is one of the most important strategies when studying bandit problems. In 2018, Nouiehed and Ross put forward a conjecture about Feldman's bandit problem (*J. Appl. Prob.* (2018) **55**, 318–324). They proposed that for Bernoulli two-armed bandit problems, the myopic strategy stochastically maximizes the number of wins. In this paper we consider the two-armed bandit problem with more general distributions and utility functions. We confirm this conjecture by proving a stronger result: if the agent playing the bandit has a general utility function, the myopic strategy is still optimal if and only if this utility function satisfies reasonable conditions.

Keywords: Myopic strategy; stochastically maximizing; dynamic programming property

2020 Mathematics Subject Classification: Primary 62C10
Secondary 62L05

1. Introduction

The bandit problem is a well-known problem in sequential control under conditions of incomplete information. It involves sequential selections from several options referred to as arms of the bandit. The payoffs of these arms are characterized by parameters which are typically unknown. Agents should learn from past information when deciding which arm to select next, with the aim of maximizing the total payoffs.

This problem can be traced back to Thompson's work [34] related to medical trials. Now it is widely studied and frequently applied as a theoretical framework for many other sequential statistical decision problems in market pricing, medical research, and engineering, which are characterized by the trade-off between exploration and exploitation (see e.g. [15], [23], and [32]).

Here we focus on the two-armed bandit problem, which is studied by Feldman [14]. For a given pair (F_1, F_2) of distributions on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, consider two experiments X and Y (called the X-arm and Y-arm), having distributions under two hypotheses H_1 and H_2 as follows:

$$\begin{array}{rrcc} & & X & Y \\ (\xi_0) & H_1: & F_1 & F_2 \\ (1-\xi_0) & H_2: & F_2 & F_1, \end{array} \tag{1.1}$$

Received 27 July 2022; revision received 23 March 2023.

[*] Postal address: School of Mathematics, Shandong University, Jinan 250100, China.

[**] Email address: jichenzhang@mail.sdu.edu.cn

https://doi.org/10.1017/jpr.2023.24 Published online by Cambridge University Press

where ξ_0 is the *a priori* probability that H_1 is true. In trial i, either the X-arm or Y-arm is selected to generate a random variable X_i or Y_i which describes the payoff, and ξ_i is the posterior probability of H_1 being true after i trials. The aim is to find the optimal strategy that maximizes the total expected payoffs.

Among the many notable strategies such as the myopic strategy, the Gittins strategy, and the play-the-winner strategy, the *myopic strategy* is undoubtedly one of the most appealing. With this strategy, in each trial, agents select the arm with greater immediate expected payoff, i.e. play each time as though there were but one trial remaining. Mathematically, let $\mathbb{E}_{\mathbb{P}}[\cdot \mid H_1]$ be the expectation functional under hypothesis H_1 if

$$\mathbb{E}_{\mathbb{P}}[X_1 \mid H_1] \geq \mathbb{E}_{\mathbb{P}}[Y_1 \mid H_1], \tag{1.2}$$

which is equivalent to

$$\int_{\mathbb{R}} x\mathrm{d}F_1(x) \geq \int_{\mathbb{R}} x\mathrm{d}F_2(x).$$

Then agents select the X-arm in trial i when $\xi_{i-1} \geq \frac{1}{2}$, or the Y-arm otherwise.

When the myopic strategy is optimal, it means that the optimal strategy is time-invariant, i.e. it does not depend on the number of trials remaining. Hence the optimal strategy can be easily implemented. Unfortunately, the myopic strategy is not optimal in general. This is mainly because at each time the myopic strategy only considers the payoff of the next trial; however, to maximize the total payoffs, all the remaining trials should be considered. Kelley [22] and Berry and Fristedt [9] showed counterexamples that the myopic strategy is not optimal. It is an open question to find out under what conditions the myopic strategy is optimal. The optimality of the myopic strategy has always attracted attention. Bradt *et al.* [10] considered model (1.1) when F_1 and F_2 are Bernoulli with expectation α and β respectively. They showed that the myopic strategy is optimal when $\alpha + \beta = 1$. Further, they conjectured that the myopic strategy is also optimal when $\alpha + \beta \neq 1$, and verified this conjecture for $n \leq 8$. For model (1.1), Feldman [14] showed that the myopic strategy is optimal in an arbitrary number of trials. Kelley [22] considered Bernoulli payoffs and asymmetric hypotheses and gave a necessary and sufficient condition for the myopic strategy being optimal. Rodman [31] extended Feldman's result to a multi-armed setting.

The exploration of the conditions under which the myopic strategy is the optimal strategy has not come to an end. This problem remains open under more general settings. Nouiehed and Ross [30] studied a Bernoulli armed bandit problem and posed a conjecture that the myopic strategy also maximizes the probability that no less than k wins occur in the first n trials, for all k, n. They proved this conjecture for $k = 1$ and $k = n$ in an n-armed bandit, and for $k = n - 1$ in the two-armed case, that is, model (1.1) with Bernoulli payoffs.

Why has the question in Nouiehed and Ross's conjecture not been raised for almost 60 years? Nouiehed and Ross [30] explained that this was because in Feldman [14] and similar studies, it was the number of times the better arm was chosen that was maximized, not the total payoff. Although the two approaches are equivalent in [14], they are quite different when we want to study a more general utility function.

This opens up a whole new horizon for us to study this issue. Let x be the total payoff, equal to the sum of the generated values. All works mentioned above considered the utility function $\varphi(x) = x$ (e.g. [9], [14], and [22]) or $\varphi(x) = I_{[k,+\infty)}(x)$ (e.g. [30]). So a natural question is what conditions can guarantee the optimality of the myopic strategy for general utility functions.

In this paper we focus on the optimal strategy for the most typical case of two-armed bandit problems (model (1.1)) proposed in the profound paper of Feldman [14]. With a general utility

https://doi.org/10.1017/jpr.2023.24 Published online by Cambridge University Press

function to be considered, we obtain a necessary and sufficient condition for the optimality of the myopic strategy. As an application, we could solve Nouiehed and Ross's conjecture for the two-armed case.

We consider a situation that the agent playing model (1.1) has a utility function φ and starts with an initial fund of x and a strategy M^n: in trial i, play the X-arm if $\xi_{i-1} \geq \frac{1}{2}$, or the Y-arm otherwise. The innovative aspects of the results obtained in this paper are as follows: first, we take F_1 and F_2 as general distribution functions, continuous or not, rather than Bernoulli distributions; second, we consider general utility functions that are no longer linear. This makes Feldman's proof method invalid and brings some additional difficulties. We shall show that M^n maximizes the expected utility of n trials if and only if the utility function φ and the distributions F_1 and F_2 satisfy

$$\mathbb{E}_{\mathbb{P}}[\varphi(u+X_1) \mid H_1] \geq \mathbb{E}_{\mathbb{P}}[\varphi(u+Y_1) \mid H_1] \quad \text{for any } u \in \mathbb{R}. \tag{1.3}$$

Condition (1.3) means that no matter how much money the agent already has, if only one trial is to be played, playing the arm with distribution F_1 is always better than playing the arm with F_2. In the case that $\varphi(x) = x$, condition (1.3) coincides with condition (1.2).

It is interesting that if we choose the utility function in condition (1.3) as an indicator function $\varphi(x) = I_{[k,+\infty)}(x)$, and initial fund $u = 0$, we could prove Nouiehed and Ross's conjecture for the two-armed case immediately.

The structure of the paper is as follows. In Section 2 we review several important strategies and compare them with the results of this paper. In Section 3 we describe the two-armed bandit problem and some basic properties. In Section 4 we first introduce a dynamic programming property of the optimal expected utility and then prove the main result. Finally, as a corollary, we derive the validity of Nouiehed and Ross's conjecture in the two-armed bandit case.

2. Related literature

After years of research, many excellent strategies have been produced. Here we review some common strategies and illustrate how the results of this paper relate to and differ from these strategies.

In the multi-armed bandit literature, a celebrated result is the so-called *Gittins index strategy* introduced by Gittins and Jones [17]. Unlike the model studied in this paper, the Gittins index strategy is used in a model with independent arms and an infinite number of trials, and the aim is to maximize the sum of geometrically discounted payoffs. This strategy assigns to each arm an index as a function of its current state, and then activates the arm with the largest index value (breaking the ties arbitrarily). It optimizes the infinite-horizon expected discounted payoffs. If the number of trials N is finite, then the Gittins index strategy can be used to approximate the optimal strategy when N is large. If more than one arm can change its state at every stage, the problem is called 'restless'. Whittle [37] proposed an index rule to solve the restless problem. This index is not necessarily optimal, but Whittle conjectured that it would admit a form of asymptotic optimality as both the number of arms and the number of allocated arms in each period grow to infinity at a fixed proportion, which was eventually proved in Weber and Weiss [36] under some technical assumptions. The restless multi-armed bandit model can be used in many aspects such as clinical trials, sensor management, and capacity management in healthcare; see [3], [13], [19], [27], [28], and [35].

A major drawback of the Gittins index and Whittle index is that they are both difficult to calculate. The current fastest algorithm can only solve the index in cubic time [16, 29]. In contrast, the calculation of the myopic strategy is much easier. A second issue of the Gittins

https://doi.org/10.1017/jpr.2023.24 Published online by Cambridge University Press

index is that the arms must have independent parameters, and the payoffs are geometrically discounted. If these conditions are not met, as in the model studied in this paper, the Gittins index strategy is no longer the optimal strategy [9, 18]. Therefore it is necessary to obtain the condition that the myopic strategy is the optimal strategy when studying the model in this paper.

Another important strategy is the UCB strategy. Lai and Robbins [25] laid out the theory of asymptotically optimal allocation and were the first to actually use the term 'upper confidence bound' (UCB). The UCB strategy is also applied to models with independent arms. Initially the UCB strategy was used for models with an infinite number of trials, but with modifications the UCB can also be used for models with a finite number of trials. When using the UCB strategy, each arm is assigned a UCB for its mean reward, and the arm with the largest bound is to be played. The bound is not the conventional upper limit for a confidence interval. The design of the confidence bound has been successively improved [2, 4–6, 11, 20, 26]. Among these, the kl-UCB strategy [11] and Bayes-UCB strategy [21] are asymptotically optimal for exponential family bandit models. The improved UCB strategy is now easy to compute but is still limited to models with independent arms.

The ϵ-greedy strategy [24] is widely used because it is very simple. At each round $t = 1, 2, \ldots$ the agent selects the arm with the highest empirical mean with probability $1 - \epsilon$, and selects a random arm with probability ϵ. It has poor asymptotic behavior, because it continues to explore long after the optimal solution becomes apparent. Another common strategy, in the case of Bernoulli arms, is play-the-winner. It is a well-known strategy in which arm a is played at time $t + 1$ if it resulted in success at time t. If a failure is observed at time t, then the next arm is either chosen at random or the arms are cycled through deterministically. Play-the-winner can be nearly optimal when the best arm has a very high success rate, but does not perform well in most cases [9, 33].

There are also many variants of the bandit problem, such as the adversarial multi-armed bandit with the EXP3 strategy [7] and online linear optimization with the FTL strategy [1], but these are beyond the scope of this paper.

In fact, compared with the various classical models mentioned above, the biggest feature of the model in this paper is the introduction of the utility function of the agent. All of the above strategies only consider linear utility or discounted returns, but agents may have nonlinear utility functions. For example, in Nouiehed and Ross's conjecture, the utility of an agent will no longer grow after a certain amount of gain. In this case, the effect of strategies such as the Gittins index or UCB will be significantly reduced.

After the introduction of a generalized utility function, the bandit problem becomes very complicated. To simplify the model and clearly demonstrate the ideas in this paper, we study this two-armed bandit model. Nouiehed and Ross's conjecture for the multi-armed bandit machine model can also be studied with the method in this paper.

The multi-armed bandit model in Ross's conjecture can be further generalized. For example, when the arms have independent parameters and the discounting factor is geometric, Banks and Sundaram [8] proved that the myopic strategy is equivalent to the Gittins index strategy for linear utility functions. When the agent has a general utility function, does the Gittins index still exist, and can the myopic strategy still be the optimal strategy under certain conditions? Exploration of these issues will lead us to better understand the nature of the bandit problem. Chen, Epstein, and Zhang [12] studied a multi-armed bandit problem where the agent is loss-averse; in particular, the agent is risk-averse in the domain of gains and risk-loving in the

https://doi.org/10.1017/jpr.2023.24 Published online by Cambridge University Press

domain of losses, and they established the corresponding asymptotically optimal strategy. This is an important advance in this research.

3. Preliminaries

Let us start with the description of the two-armed bandit model (1.1) and the strategies.

Consider the bandit model in equation (1.1). Let $\{X_i\}_{i\geq 1}$ be a sequence of random variables, where X_i describes the payoff of trial i from the X-arm, and let $\{Y_i\}_{i\geq 1}$ be a sequence of random variables selected from the Y-arm; $\{(X_i, Y_i)\}_{i\geq 1}$ are independent under each hypothesis. We define $\mathcal{F}_i := \sigma\{(X_1, Y_1), \ldots, (X_i, Y_i)\}$, which represents all the information that can be obtained until trial i.

Remark 3.1. Note that in practice, regardless of the strategy chosen, the information obtained after the ith trial is a proper subset of $\mathcal{F}_i$, because for experiment j, the agent can only observe one of X_j and Y_j.

We call this model a (ξ_0, n, x)-*bandit* if there are n trials to be played with initial fund x and a prior probability ξ_0. In the following discussion, the distributions F_1 and F_2 of arms are continuous with density f_1 and f_2, respectively.

Remark 3.2. The same results still hold when the distributions of arms are discrete, e.g. Bernoulli. We only need to modify the calculation of expectations in this case.

For each $i \geq 1$, let θ_i be an $\mathcal{F}_{i-1}$-measurable random variable taking values in $\{0, 1\}$, where $\theta_i = 1$ means the X-arm is selected for observation in trial i and $\theta_i = 0$ means the Y-arm is selected for observation in trial i. The payoff that an agent receives using θ_i in trial i is

$$Z_i^\theta := \theta_i X_i + (1 - \theta_i) Y_i.$$

For a (ξ_0, n, x)-bandit and $\theta = \{\theta_1, \ldots, \theta_n\}$, if $\theta_i \in \sigma(Z_1^\theta, \ldots, Z_{i-1}^\theta) \subset \mathcal{F}_{i-1}$, then we call θ a *strategy*. The set of strategies for a (ξ_0, n, x)-bandit is denoted by Θ_n.

For a (ξ_0, n, x)-bandit and a suitable measurable function φ, the *expected utility* obtained by using strategy θ is denoted by

$$W(\xi_0, n, x, \theta) = \mathbb{E}_{\mathbb{P}}\left[\varphi\left(x + \sum_{i=1}^{n} Z_i^\theta\right)\right],$$

where φ is called a utility function.

For each strategy $\theta \in \Theta_n$, let $\{\xi_i^\theta\}_{i\geq 1}$ be the sequence of the posterior probabilities that hypothesis H_1 is true after i trials. The posterior probability ξ_1^θ after trial 1 with payoff s is calculated by

$$\xi_1^\theta(s) = \begin{cases} \dfrac{\xi_0 f_1(s)}{\xi_0 f_1(s) + (1 - \xi_0) f_2(s)} & \text{if } \theta_1 = 1, \text{ i.e. the } X\text{-arm is selected,} \\[2ex] \dfrac{\xi_0 f_2(s)}{\xi_0 f_2(s) + (1 - \xi_0) f_1(s)} & \text{if } \theta_1 = 0, \text{ i.e. the } Y\text{-arm is selected.} \end{cases} \tag{3.1}$$

https://doi.org/10.1017/jpr.2023.24 Published online by Cambridge University Press

We can easily obtain that for any fixed s, $\xi_1^\theta(s)$ is increasing in ξ_0. When the posterior probability ξ_i^θ is known and the payoff of the $i+1$ trial is s, there is a recursive formula

$$\xi_{i+1}^\theta(s) = \begin{cases} \dfrac{\xi_i^\theta f_1(s)}{\xi_i^\theta f_1(s) + (1-\xi_i^\theta)f_2(s)} & \text{if } \theta_{i+1}=1, \text{ i.e. the } X\text{-arm is selected,} \\[2ex] \dfrac{\xi_i^\theta f_2(s)}{\xi_i^\theta f_2(s) + (1-\xi_i^\theta)f_1(s)} & \text{if } \theta_{i+1}=0, \text{ i.e. the } Y\text{-arm is selected.} \end{cases} \tag{3.2}$$

Now we propose the following two-armed bandit problem.

Problem (TAB). For a (ξ_0, n, x)-bandit and a utility function φ, find some strategy in Θ_n to achieve the maximal expected utility

$$V(\xi_0, n, x) := \sup_{\theta\in\Theta_n} W(\xi_0, n, x, \theta) = \sup_{\theta\in\Theta_n} \mathbb{E}_{\mathbb{P}}\left[\varphi\left(x + \sum_{i=1}^n Z_i^\theta\right)\right]. \tag{3.3}$$

Note that the expected utility $\mathbb{E}_{\mathbb{P}}[\cdot]$ depends on hypothesis H_1, H_2, and ξ_0. In fact,

$$\mathbb{E}_{\mathbb{P}}\left[\varphi\left(x + \sum_{i=1}^n Z_i^\theta\right)\right] = \xi_0\mathbb{E}_{\mathbb{P}}\left[\varphi\left(x + \sum_{i=1}^n Z_i^\theta\right) \mid H_1\right] + (1-\xi_0)\mathbb{E}_{\mathbb{P}}\left[\varphi\left(x + \sum_{i=1}^n Z_i^\theta\right) \mid H_2\right],$$

where $\mathbb{E}_{\mathbb{P}}[\cdot \mid H_i]$ is the expectation under hypothesis H_i $(i=1, 2)$.

To simplify the notation, we write $\mathbb{E}_{\mathbb{P}}[\cdot \mid H_1]$ as $\mathbb{E}_1[\cdot]$, $\mathbb{E}_{\mathbb{P}}[\cdot \mid H_2]$ as $\mathbb{E}_2[\cdot]$, and $\mathbb{E}_{\mathbb{P}}[\cdot]$ as $\mathbb{E}_{\xi_0}[\cdot]$. Then the expected utility can be written as

$$W(\xi_0, n, x, \theta) = \mathbb{E}_{\xi_0}\left[\varphi\left(x + \sum_{i=1}^n Z_i^\theta\right)\right],$$

where

$$\mathbb{E}_{\xi_0}\left[\varphi\left(x + \sum_{i=1}^n Z_i^\theta\right)\right] = \xi_0\mathbb{E}_1\left[\varphi\left(x + \sum_{i=1}^n Z_i^\theta\right)\right] + (1-\xi_0)\mathbb{E}_2\left[\varphi\left(x + \sum_{i=1}^n Z_i^\theta\right)\right].$$

Immediately, equality (3.3) can be written as follows:

$$V(\xi_0, n, x) = \sup_{\theta\in\Theta_n} W(\xi_0, n, x, \theta) = \sup_{\theta\in\Theta_n} \mathbb{E}_{\xi_0}\left[\varphi\left(x + \sum_{i=1}^n Z_i^\theta\right)\right].$$

Consider a strategy M^n: in trial i, play the X-arm if $\xi_{i-1} \geq \frac{1}{2}$, or the Y-arm otherwise. Our main result is to find conditions under which M^n could solve Problem (TAB).

The following lemma shows that when calculating the expected utility, we can temporarily fix the payoff of the first trial, calculate the expected utility as a function of the first payoff, and then take expectation while seeing the first payoff as a random variable. This is an extension of equation (2) in Feldman [14].

Lemma 3.1. *For each integer $n \geq 2$ and strategy $\theta = \{\theta_1, \ldots, \theta_n\} \in \Theta_n$, we have*

$$\mathbb{E}_{\xi_0}\left[\varphi\left(x + \sum_{i=1}^n Z_i^\theta\right)\right] = \mathbb{E}_{\xi_0}\big[h\big(x, Z_1^\theta\big)\big] \quad \textit{for all } x \in \mathbb{R}, \tag{3.4}$$

https://doi.org/10.1017/jpr.2023.24 Published online by Cambridge University Press

where

$$h(x, u) = \mathbb{E}_{\xi_1^\theta(u)}\left[\varphi\left(x + u + \sum_{i=2}^{n} Z_i^{\theta[u]}\right)\right],$$

ξ_1^θ *is defined by* (3.1)*, and* $\theta[u]$ *is the strategy obtained from* θ *by fixing the payoff of the first trial to be u.*

Remark 3.3. $\mathbb{E}_{\xi_1^\theta(u)}[\cdot]$ is the expected utility $\mathbb{E}_{\xi_0}[\cdot]$ replacing ξ_0 with $\xi_1^\theta(u)$. In integral form, equation (3.4) is

$$\mathbb{E}_{\xi_0}\left[\varphi\left(x + \sum_{i=1}^{n} Z_i^\theta\right)\right] = \int_{\mathbb{R}} \mathbb{E}_{\xi_1^\theta(u)}\left[\varphi\left(x + u + \sum_{i=2}^{n} Z_i^{\theta[u]}\right)\right](\xi_0 f_1(u) + (1 - \xi_0) f_2(u))\,\mathrm{d}u.$$

Remark 3.4. Here $h\big(x, Z_1^\theta\big)$ can be seen as a conditional expectation of $\varphi\big(x + \sum_{i=1}^n Z_i^\theta\big)$ given $\mathcal{F}_1$, and this lemma shows that it has the same expectation as $\varphi\big(x + \sum_{i=1}^n Z_i^\theta\big)$. This is in fact the tower property for conditional expectations, so we omit the proof.

We can see that the form of $h(x, u)$ is very similar to the expected utility of the $(\xi_1^\theta(u), n - 1, x + u)$-bandit with some strategy. There is indeed such a strategy to make the value of $h(x, u)$ equal to the expected utility of the $(\xi_1^\theta(u), n - 1, x + u)$-bandit.

Lemma 3.2. *For any strategy* $\theta \in \Theta_n$*, let*

$$h(x, u) = \mathbb{E}_{\xi_1^\theta(u)}\left[\varphi\left(x + u + \sum_{i=2}^{n} Z_i^{\theta[u]}\right)\right].$$

Then, for any u, there exists a strategy $\rho \in \Theta_{n-1}$*, such that the value of h(x,u) is equal to the expected utility of the* $(\xi_1^\theta(u), n - 1, x + u)$*-bandit with strategy* ρ*, that is,*

$$\begin{aligned} h(x, u) &= \mathbb{E}_{\xi_1^\theta(u)}\left[\varphi\left(x + u + \sum_{i=2}^{n} Z_i^{\theta[u]}\right)\right] \\ &= \mathbb{E}_{\xi_1^\theta(u)}\left[\varphi\left(x + u + \sum_{i=1}^{n-1} Z_i^{\rho}\right)\right] \\ &= W(\xi_1^\theta(u), n - 1, x + u, \rho). \end{aligned}$$

Proof. We know that $\theta[u]$ is obtained by fixing the payoff u of the first trial, so for any $\theta_i' \in \theta[u]$ it is $\sigma(X_2, Y_2, \ldots, X_{i-1}, Y_{i-1})$-measurable. Then there are measurable functions π_i, $i \geq 2$, such that

$$\theta_i' = \pi_i(X_2, Y_2, \ldots, X_{i-1}, Y_{i-1}), \qquad i \geq 2.$$

Define a new strategy $\rho \in \Theta_{n-1}$ by

$$\rho_i = \pi_{i+1}(X_1, Y_1, \ldots, X_{i-1}, Y_{i-1}), \qquad 1 \leq i \leq n - 1.$$

The fact that $\rho_{i+1} \in \sigma(Z_1^\rho, \ldots, Z_i^\rho)$ can be easily verified.

https://doi.org/10.1017/jpr.2023.24 Published online by Cambridge University Press

By the definition of ρ, we know that ρ_i has the same distribution as θ'_{i+1} in both hypotheses, so $Z_i^{\theta[u]}$ and Z_i^{ρ} have the same distribution. Using this fact, we can easily verify that

$$\mathbb{E}_{\xi_1^{\theta}(u)}\left[\varphi\left(x+u+\sum_{i=2}^{n} Z_i^{\theta[u]}\right)\right]=\mathbb{E}_{\xi_1^{\theta}(u)}\left[\varphi\left(x+u+\sum_{i=1}^{n-1} Z_i^{\rho}\right)\right]. \qquad \square$$

Now we introduce the dynamic programming property of the expected utility, which plays an important role in the subsequent arguments. Similar results are found in many works on bandit problems, but only for the case of $\varphi(x)=x$ (e.g. [9], [14]). Our result extends the classical ones.

Theorem 3.1. *For each $\xi_0\in[0, 1]$, $n\geq 1$ and $x\in\mathbb{R}$, consider the (ξ_0, n, x)-bandit. The optimal strategy $\theta^{[n]}\in\Theta_n$ exists. Then there are measurable functions*

$$\pi_i^{[n]}:[0, 1]\times\mathbb{R}\times\mathbb{R}^{2i-2}\mapsto\{0, 1\}, \quad 1\leq i\leq n,$$

such that for any $\xi_0\in[0, 1]$ and $x\in\mathbb{R}$, the optimal strategy $\theta^{[n]}$ for the (ξ_0, n, x)-bandit satisfies

$$\begin{aligned}\theta_1^{[n]}&=\pi_1^{[n]}(\xi_0, x),\\ \theta_i^{[n]}&=\pi_i^{[n]}(\xi_0, x, X_1, Y_1, \ldots, X_{i-1}, Y_{i-1}), \quad i\geq 2.\end{aligned}$$

Further, the optimal expected utility satisfies the following dynamic programming property:

$$\begin{aligned}V(\xi_0, n, x)&=\sup_{\theta\in\Theta_n}\mathbb{E}_{\xi_0}\left[V\left(\xi_1^{\theta}, n-1, x+Z_1^{\theta}\right)\right]\\ &=\max\left\{\mathbb{E}_{\xi_0}\left[V\left(\frac{\xi_0 f_1(X_1)}{\xi_0 f_1(X_1)+(1-\xi_0)f_2(X_1)}, n-1, x+X_1\right)\right],\right.\\ &\qquad\left.\mathbb{E}_{\xi_0}\left[V\left(\frac{\xi_0 f_2(Y_1)}{\xi_0 f_2(Y_1)+(1-\xi_0)f_1(Y_1)}, n-1, x+Y_1\right)\right]\right\}.\end{aligned}$$

Theorem 3.1 is mostly standard for a finite action and horizon dynamic programming problem, so we omit the proof.

4. Main results

Now we are going to study the specific form of the optimal strategy, for the finite two-armed bandit (1.1) with utility function φ. It is obvious that different utility functions may lead to different optimal strategies, but we will show that when a reasonable condition is satisfied, the optimal strategy is independent of the specific form of φ.

Recall that the myopic strategy for (ξ_0, n, x)-bandits is $\mathsf{M}^n=\{\mathsf{m}_1^n, \ldots, \mathsf{m}_n^n\}$: in trial i, play the X-arm if the posterior probability $\xi_{i-1}^{\mathsf{M}^n}\geq\frac{1}{2}$, or the Y-arm if $\xi_{i-1}^{\mathsf{M}^n}<\frac{1}{2}$. Note that $\mathsf{M}^n\in\Theta_n$. In fact, M^n can be denoted by

$$\begin{aligned}\mathsf{M}^n&=\left\{\mathsf{m}_1^n, \ldots, \mathsf{m}_n^n\right\}\\ &=\left\{g(\xi_0), g\left(\xi_1^{\mathsf{M}^n}\right), \ldots, g\left(\xi_{n-1}^{\mathsf{M}^n}\right)\right\}\in\Theta_n,\end{aligned} \tag{4.1}$$

https://doi.org/10.1017/jpr.2023.24 Published online by Cambridge University Press

where $g(x)=1$ if $x \geq \frac{1}{2}$, or $g(x)=0$ if $x<\frac{1}{2}$. From the definition of $\xi_{i-1}^{\mathsf{M}^n}$ and m_i^n, we know that they are both independent of n and x. Hence we can write $\xi_{i-1}^{\mathsf{M}^n}$ for short as ξ_{i-1}^{M}, and write m_i^n as m_i. Now the myopic strategy M^n is denoted by

$$\begin{aligned} \mathsf{M}^n &= \{\mathsf{m}_1, \ldots, \mathsf{m}_n\} \\ &= \left\{g(\xi_0), g\left(\xi_1^{\mathsf{M}}\right), \ldots, g\left(\xi_{n-1}^{\mathsf{M}}\right)\right\}. \end{aligned}$$

Next we will give a condition on φ which is necessary and sufficient for M^n being the optimal strategy.

Theorem 4.1. *For any integer $n \geq 1$, the myopic strategy M^n is the optimal strategy of the (ξ_0, n, x)-bandit for all $\xi_0 \in [0, 1]$, $x \in \mathbb{R}$, if and only if*

$$\mathbb{E}_1[\varphi(u+X_1)] \geq \mathbb{E}_1[\varphi(u+Y_1)] \quad \textit{for all } u \in \mathbb{R}. \tag{I}$$

Remark 4.1. When $\varphi(x)=x$, condition (I) is in fact

$$\mathbb{E}_1[X_1] \geq \mathbb{E}_1[Y_1],$$

and Theorem 4.1 in this case is exactly Theorem 2.1 of Feldman [14].

Remark 4.2. When the two distributions F_1, F_2 are Bernoulli distributions, say Bernoulli (α) and Bernoulli (β), $\alpha, \beta \in (0, 1)$, then condition (I) is written as

$$(\varphi(x+1)-\varphi(x))(\alpha-\beta) \geq 0 \quad \text{for any } x \in \mathbb{R}.$$

If $\varphi(x)$ is an increasing function of x, and $\alpha \geq \beta$, then condition (I) holds.

Remark 4.3. Note that condition (I) here is a necessary and sufficient condition to make M^n the optimal strategy of the (ξ_0, n, x)-bandit model for any $x \in \mathbb{R}$ and any $\xi_0 \in [0, 1]$. However, when condition (I) is not satisfied, it is still possible that there is a specific triple $(\overline{\xi_0}, \overline{n}, \overline{x})$ that makes $\mathsf{M}^{\overline{n}}$ the optimal strategy of the $(\overline{\xi_0}, \overline{n}, \overline{x})$-bandit, but the optimality of M^n does not hold for general (ξ_0, n, x) triples.

To achieve our goal, we need to formulate some properties of the expected utility of M^n, which can be considered as extensions of *properties of R_N*, and Lemma 2.1 in Feldman's paper [14].

Lemma 4.1. *For each $n \in \mathbb{Z}^+$, we have the following.*

(1) *The expected utility with strategy M^n is symmetric about $\xi_0 = \frac{1}{2}$, that is,*

$$W(\xi_0, n, x, \mathsf{M}^n) = W(1-\xi_0, n, x, \mathsf{M}^n) \quad \textit{for all } \xi_0 \in [0, 1], x \in \mathbb{R}. \tag{4.2}$$

(2) *Let us define the strategies*

$$\mathsf{L}^n = \left\{1, g\left(\xi_1^{\mathsf{L}^n}\right), \ldots, g\left(\xi_{n-1}^{\mathsf{L}^n}\right)\right\} \quad \textit{and} \quad \mathsf{R}^n = \left\{0, g\left(\xi_1^{\mathsf{R}^n}\right), \ldots, g\left(\xi_{n-1}^{\mathsf{R}^n}\right)\right\},$$

where $g(\cdot)$ is the function defined in (4.1). Then

$$W(\xi_0, n, x, \mathsf{L}^n) = W(1-\xi_0, n, x, \mathsf{R}^n) \quad \textit{for all } \xi_0 \in [0, 1], x \in \mathbb{R}. \tag{4.3}$$

https://doi.org/10.1017/jpr.2023.24 Published online by Cambridge University Press

Proof. An important feature of model (1.1) is symmetry. Swapping the X-arm with the Y-arm and exchanging hypothesis H_1 with H_2, we can obtain equation (4.3). Equation (4.2) can be obtained from (4.3) and the definition of M^n. Lemma 4.1 can also be proved using mathematical induction and Lemmas 3.1 and 3.2. □

Lemma 4.2. *Consider the following strategies.*

For $n=2$, let $\mathsf{U}^2=\{1, 0\}$ and $\mathsf{V}^2=\{0, 1\}$.

For each $n \geq 3$, let

$$\mathsf{U}^n=\{1, 0, g(\xi_2^{\mathsf{U}^n}), \ldots, g(\xi_{n-1}^{\mathsf{U}^n})\},$$

$$\mathsf{V}^n=\{0, 1, g(\xi_2^{\mathsf{V}^n}), \ldots, g(\xi_{n-1}^{\mathsf{V}^n})\},$$

where $g(\cdot)$ is the function defined in (4.1). Then, for each $n \geq 2$, $\xi_0 \in [0, 1]$ and $x \in \mathbb{R}$, the expected utilities obtained by using U^n and V^n satisfy the relation

$$W(\xi_0, n, x, \mathsf{U}^n)=W(\xi_0, n, x, \mathsf{V}^n).$$

Proof. For $n=2$ the result is obvious. We only need to consider cases where $n \geq 3$. Let $\xi_2^{\mathsf{U}^n}(u, s)$ be the posterior probability given the first payoff u and second payoff s, using strategy U^n, and let $\xi_2^{\mathsf{V}^n}(s, u)$ be the posterior probability given the first payoff s and second payoff u, using strategy V^n. According to (3.2), there is

$$\xi_2^{\mathsf{U}^n}(u, s)=\frac{\xi_0 f_1(u) f_2(s)}{\xi_0 f_1(u) f_2(s)+(1-\xi_0) f_2(u) f_1(s)}=\xi_2^{\mathsf{V}^n}(s, u). \tag{4.4}$$

Let the conditional strategy $\mathsf{U}^n[u, s]$ denote the strategy U^n given the first payoff u and second payoff s, and let the conditional strategy $\mathsf{V}^n[s, u]$ denote the strategy V^n given the first payoff s and second payoff u. Then, by (4.4) and the definitions of $\mathsf{U}^n[u, s]$ and $\mathsf{V}^n[s, u]$, we can easily get that these two conditional strategies are the same for $i \geq 3$ trials.

Using the same techniques as in Lemma 3.1, we obtain

$$\begin{aligned}
& W(\xi_0, n, x, \mathsf{U}^n) \\
& \quad =\mathbb{E}_{\xi_0}\left[\varphi\left(x+\sum_{i=1}^n Z_i^{\mathsf{U}^n}\right)\right] \\
& \quad =\int_{\mathbb{R}} \int_{\mathbb{R}} \mathbb{E}_{\xi_2^{\mathsf{U}^n}(u,s)}\left[\varphi\left(x+u+s+\sum_{j=3}^n Z_j^{\mathsf{U}^n[u,s]}\right)\right](\xi_0 f_1(u) f_2(s)+(1-\xi_0) f_2(u) f_1(s)) \,\mathrm{d}s \,\mathrm{d}u
\end{aligned}$$

and

$$\begin{aligned}
& W(\xi_0, n, x, \mathsf{V}^n) \\
& \quad =\mathbb{E}_{\xi_0}\left[\varphi\left(x+\sum_{i=1}^n Z_i^{\mathsf{V}^n}\right)\right] \\
& \quad =\int_{\mathbb{R}} \int_{\mathbb{R}} \mathbb{E}_{\xi_2^{\mathsf{V}^n}(s,u)}\left[\varphi\left(x+s+u+\sum_{j=3}^n Z_j^{\mathsf{V}^n[s,u]}\right)\right](\xi_0 f_2(s) f_1(u)+(1-\xi_0) f_1(s) f_2(u)) \,\mathrm{d}u \,\mathrm{d}s.
\end{aligned}$$

Then the desired result is obtained by using (4.4) and the fact that $\mathsf{U}^n[u, s] = \mathsf{V}^n[s, u]$ for $i \geq 3$ trials. □

For each $n \in \mathbb{Z}^+$, let L^n and R^n be the strategies defined in Lemma 4.1. For $x \in \mathbb{R}$, $\xi_0 \in [0, 1]$, define the difference of expected utilities by

$$\begin{aligned}\Delta_n(x, \xi_0) :=\ & W(\xi_0, n, x, \mathsf{L}^n) - W(\xi_0, n, x, \mathsf{R}^n) \\ &= \mathbb{E}_{\xi_0}\left[\varphi\left(x + \sum_{i=1}^n Z_i^{\mathsf{L}^n}\right)\right] - \mathbb{E}_{\xi_0}\left[\varphi\left(x + \sum_{i=1}^n Z_i^{\mathsf{R}^n}\right)\right].\end{aligned}$$

By Lemma 4.1 we can obtain $\Delta_n(x, \xi_0) = -\Delta_n(x, 1 - \xi_0)$ and $\Delta_n(x, 0.5) = 0$.

In the case $\varphi(x) = x$, there is an important recurrence formula for $\Delta_n(x, \xi_0)$; see equations (12), (13), and (14) in [14] and equation (7.1.1) in [9]. The following lemma shows that this recurrence formula still holds for a general utility function.

Lemma 4.3. *For $n \geq 2$, and any $x \in \mathbb{R}$, $\xi_0 \in [0, 1]$, there is*

$$\begin{aligned}\Delta_n(x, \xi_0) = & \int_{\mathbb{R}} I_{\left\{\xi_1^X(u) \geq 0.5\right\}} \Delta_{n-1}\big(x + u, \xi_1^X(u)\big)\big(\xi_0 f_1(u) + (1 - \xi_0) f_2(u)\big)\, \mathrm{d}u \\ & + \int_{\mathbb{R}} I_{\left\{\xi_1^Y(u) < 0.5\right\}} \Delta_{n-1}\big(x + u, \xi_1^Y(u)\big)\big(\xi_0 f_2(u) + (1 - \xi_0) f_1(u)\big)\, \mathrm{d}u,\end{aligned}$$

where

$$\xi_1^X(u) = \frac{\xi_0 f_1(u)}{\xi_0 f_1(u) + (1 - \xi_0) f_2(u)} \quad \textit{and} \quad \xi_1^Y(u) = \frac{\xi_0 f_2(u)}{\xi_0 f_2(u) + (1 - \xi_0) f_1(u)}.$$

Proof. Consider strategies U^n, V^n defined in Lemma 4.2. Using Lemmas 3.1 and 3.2, we can make the following two differences:

$$\begin{aligned}& W(\xi_0, n, x, \mathsf{L}^n) - W(\xi_0, n, x, \mathsf{U}^n) \\ &\quad = \mathbb{E}_{\xi_0}\left[\varphi\left(x + \sum_{i=1}^n Z_i^{\mathsf{L}^n}\right)\right] - \mathbb{E}_{\xi_0}\left[\varphi\left(x + \sum_{i=1}^n Z_i^{\mathsf{U}^n}\right)\right] \\ &\quad = \int_{\mathbb{R}} I_{\left\{\xi_1^X(u) \geq 0.5\right\}} \Delta_{n-1}\big(x + u, \xi_1^X(u)\big)\big(\xi_0 f_1(u) + (1 - \xi_0) f_2(u)\big)\, \mathrm{d}u, \\ & W(\xi_0, n, x, \mathsf{V}^n) - W(\xi_0, n, x, \mathsf{R}^n) \\ &\quad = \mathbb{E}_{\xi_0}\left[\varphi\left(x + \sum_{i=1}^n Z_i^{\mathsf{V}^n}\right)\right] - \mathbb{E}_{\xi_0}\left[\varphi\left(x + \sum_{i=1}^n Z_i^{\mathsf{R}^n}\right)\right] \\ &\quad = \int_{\mathbb{R}} I_{\left\{\xi_1^Y(u) < 0.5\right\}} \Delta_{n-1}\big(x + u, \xi_1^Y(u)\big)\big(\xi_0 f_2(u) + (1 - \xi_0) f_1(u)\big)\, \mathrm{d}u.\end{aligned}$$

The computational details are similar to the proofs of Lemmas 4.1 and 4.2, so we omit them. The desired formula is obtained by adding the above two equations and using Lemma 4.2. □

https://doi.org/10.1017/jpr.2023.24 Published online by Cambridge University Press

The key to achieving our main theorem is to prove that for any fixed $x \in \mathbb{R}$ and $n \in \mathbb{Z}^+$, the above difference $\Delta_n(x, \xi_0)$ is an increasing function of ξ_0. To prove this assertion, we use a method similar to that used by Rodman [31].

Define functions $D_n(x, t_X, t_Y)$, $n = 1, 2, \ldots$, for every tuple (x, t_X, t_Y) of numbers, such that $x \in \mathbb{R}$ and $t_X, t_Y \geq 0$:

$$D_n(x, t_X, t_Y) = \begin{cases} (t_X + t_Y)\Delta_n\left(x, \dfrac{t_X}{t_X + t_Y}\right) & \text{if } t_X + t_Y > 0, \\ 0 & \text{otherwise.} \end{cases}$$

From this definition we obtain immediately that

$$D_n(x, t_X, t_Y) = -D_n(x, t_Y, t_X) \quad \text{and} \quad D_n(x, t_X, t_Y) = 0 \quad \text{if } t_X = t_Y.$$

Lemma 4.4. *If condition* (I) *holds, then $D_n(x, t_X, t_Y)$ is an increasing function of t_X, when x, t_Y are kept fixed.*

Proof. The proof is by induction. When $n = 1$,

$$D_1(x, t_X, t_Y) = (t_X - t_Y) \int_{\mathbb{R}} \varphi(x + u)(f_1(u) - f_2(u))\, \mathrm{d}u$$

is clearly an increasing function of t_X when x, t_Y are kept fixed.

Now suppose Lemma 4.4 is proved for $n = k$. From Lemma 4.3 we know that

$$\begin{aligned} D_{k+1}(x, t_X, t_Y) = &\int_{\mathbb{R}} I_{\{t_X f_1(u) \geq t_Y f_2(u)\}} D_k(x + u, t_X f_1(u), t_Y f_2(u))\, \mathrm{d}u \\ &+ \int_{\mathbb{R}} I_{\{t_X f_2(u) < t_Y f_1(u)\}} D_k(x + u, t_X f_2(u), t_Y f_1(u))\, \mathrm{d}u. \end{aligned}$$

When u, x, t_Y are fixed,

$$D_k(x + u, t_X f_1(u), t_Y f_2(u)) \quad \text{and} \quad D_k(x + u, t_X f_2(u), t_Y f_1(u))$$

are increasing functions of t_X. Moreover,

$$\begin{aligned} D_k(x + u, t_X f_1(u), t_Y f_2(u)) &\geq 0 \quad \text{when } t_X f_1(u) \geq t_Y f_2(u), \\ D_k(x + u, t_X f_2(u), t_Y f_1(u)) &\leq 0 \quad \text{when } t_X f_2(u) < t_Y f_1(u). \end{aligned}$$

So the two integrands in the above two integrals are both increasing functions of t_X.

Hence we obtain that $D_{k+1}(x, t_X, t_Y)$ is an increasing function of t_X when x, t_Y are fixed, and complete the proof. □

Now let $t_X = \xi_0$, $t_Y = 1 - \xi_0$. Then

$$D_n(x, t_X, t_Y) = \Delta_n(x, \xi_0).$$

By Lemma 4.4 and $D_n(x, t_X, t_Y) = -D_n(x, t_Y, t_X)$, we get the desired fact.

https://doi.org/10.1017/jpr.2023.24 Published online by Cambridge University Press

Corollary 4.1. *For any fixed $x \in \mathbb{R}$ and $n \in \mathbb{Z}^+$, $\Delta_n(x, \xi_0)$ is an increasing function of ξ_0.*

Now we are ready to prove Theorem 4.1.

Proof of Theorem 4.1. Firstly, we assume that condition (I) holds and prove the optimality of M^n. This can be easily obtained by Theorem 3.1 and Corollary 4.1. We now use mathematical induction.

When $n = 1$, by Corollary 4.1,

$$\Delta_1(x, \xi_0) = \mathbb{E}_{\xi_0}[\varphi(x + X_1)] - \mathbb{E}_{\xi_0}[\varphi(x + Y_1)]$$

is an increasing function of ξ_0 for any fixed x. Since $\Delta_1(x, \frac{1}{2}) = 0$, then the optimal strategy should choose X first if $\xi_0 \geq \frac{1}{2}$ and choose Y first if $\xi < \frac{1}{2}$. This means that M^1 is the optimal strategy of the $(\xi_0, 1, x)$-bandit, for any $x \in \mathbb{R}$ and $\xi_0 \in [0, 1]$.

Assume that for fixed $k \geq 1$, the myopic strategy M^k is optimal for (ξ_0, k, x)-bandits, for any $x \in \mathbb{R}$ and $\xi_0 \in [0, 1]$.

Now consider a bandit problem with $k + 1$ trials. Note that by definition of M^{k+1}, L^{k+1}, and R^{k+1}, there is

$$W(\xi_0, k+1, x, \mathsf{M}^{k+1}) = \begin{cases} W(\xi_0, k+1, x, \mathsf{L}^{k+1}) & \text{if } \xi_0 \geq \frac{1}{2}, \\ W(\xi_0, k+1, x, \mathsf{R}^{k+1}) & \text{otherwise.} \end{cases}$$

By Lemmas 3.1 and 3.2 we know that

$$\begin{aligned} W(\xi_0, k+1, x, \mathsf{L}^{k+1}) &= \mathbb{E}_{\xi_0}\left[\varphi\left(x + \sum_{i=1}^{k+1} Z_i^{\mathsf{L}^{k+1}}\right)\right] \\ &= \mathbb{E}_{\xi_0}\left[W\left(\frac{\xi_0 f_1(X_1)}{\xi_0 f_1(X_1) + (1-\xi_0) f_2(X_1)}, k, x + X_1, \mathsf{M}^k\right)\right], \\ W(\xi_0, k+1, x, \mathsf{R}^{k+1}) &= \mathbb{E}_{\xi_0}\left[\varphi\left(x + \sum_{i=1}^{k+1} Z_i^{\mathsf{R}^{k+1}}\right)\right] \\ &= \mathbb{E}_{\xi_0}\left[W\left(\frac{\xi_0 f_2(Y_1)}{\xi_0 f_2(Y_1) + (1-\xi_0) f_1(Y_1)}, k, x + Y_1, \mathsf{M}^k\right)\right]. \end{aligned}$$

By Corollary 4.1, $\Delta_{k+1}(x, \xi_0)$ is an increasing function of ξ_0, $\Delta_{k+1}(x, \frac{1}{2}) = 0$. By the induction hypothesis, M^k is the optimal strategy of $\left(\frac{\xi_0 f_1(u)}{\xi_0 f_1(u) + (1-\xi_0) f_2(u)}, k, x + u\right)$-bandits and $\left(\frac{\xi_0 f_2(u)}{\xi_0 f_2(u) + (1-\xi_0) f_1(u)}, k, x + u\right)$-bandits, for $u \in \mathbb{R}$. Then we have

$$\begin{aligned} & W\big(\xi_0, k+1, x, \mathsf{M}^{k+1}\big) \\ &\quad = \max\big\{W\big(\xi_0, k+1, x, \mathsf{L}^{k+1}\big), W\big(\xi_0, k+1, x, \mathsf{R}^{k+1}\big)\big\} \\ &\quad = \max\left\{\mathbb{E}_{\xi_0}\left[V\left(\frac{\xi_0 f_1(X_1)}{\xi_0 f_1(X_1) + (1-\xi_0) f_2(X_1)}, k, x + X_1\right)\right],\right. \\ &\qquad\qquad \left.\mathbb{E}_{\xi_0}\left[V\left(\frac{\xi_0 f_2(Y_1)}{\xi_0 f_2(Y_1) + (1-\xi_0) f_1(Y_1)}, k, x + Y_1\right)\right]\right\}. \end{aligned}$$

https://doi.org/10.1017/jpr.2023.24 Published online by Cambridge University Press

Therefore, by Theorem 3.1,

$$W(\xi_0, k+1, x, \mathsf{M}^{k+1}) = V(\xi_0, k+1, x).$$

Hence M^{k+1} is the optimal strategy of $(\xi_0, k+1, x)$-bandits, for any $x \in \mathbb{R}$ and $\xi_0 \in [0, 1]$. The first part of the main theorem is proved.

Now we prove that if the strategy M^n is optimal for (ξ_0, n, x)-bandits, for any integer $n \geq 1$, for all $\xi_0 \in [0, 1]$, $x \in \mathbb{R}$, then condition (I) holds.

Indeed, we only need to consider the case when $n = 1$. For any fixed $x \in \mathbb{R}$, we have

$$\mathbb{E}_{\xi_0}[\varphi(x+X_1)] = \xi_0 \int_{\mathbb{R}} \varphi(x+u) f_1(u)\, \mathrm{d}u + (1-\xi_0) \int_{\mathbb{R}} \varphi(x+u) f_2(u)\, \mathrm{d}u,$$

$$\mathbb{E}_{\xi_0}[\varphi(x+Y_1)] = \xi_0 \int_{\mathbb{R}} \varphi(x+u) f_2(u)\, \mathrm{d}u + (1-\xi_0) \int_{\mathbb{R}} \varphi(x+u) f_1(u)\, \mathrm{d}u.$$

Since the strategy M^1 which chooses X if and only if $\xi_0 \geq \frac{1}{2}$ is the optimal strategy, we know that

$$\mathbb{E}_{\xi_0}[\varphi(x+X_1)] - \mathbb{E}_{\xi_0}[\varphi(x+Y_1)] \geq 0 \quad \text{if } \xi_0 \geq \tfrac{1}{2}.$$

Then, for any fixed $x \in \mathbb{R}$, we have

$$(2\xi_0 - 1)\left[\int_{\mathbb{R}} \varphi(x+u) f_1(u)\, \mathrm{d}u - \int_{\mathbb{R}} \varphi(x+u) f_2(u)\, \mathrm{d}u\right] \geq 0 \quad \text{if } \xi_0 \geq \tfrac{1}{2},$$

which leads to

$$\int_{\mathbb{R}} \varphi(x+u) f_1(u)\, \mathrm{d}u - \int_{\mathbb{R}} \varphi(x+u) f_2(u)\, \mathrm{d}u \geq 0.$$

This is clearly condition (I). □

Theorem 4.1 gives a reasonable condition on φ that is necessary and sufficient for the myopic strategy M^n being the optimal strategy. With Theorem 4.1 in hand, we can immediately obtain the following two corollaries. The first corollary is the same as the result of Feldman [14], obtained by applying Theorem 4.1 on the $\varphi(x) = x$ case. The second corollary answers Nouiehed and Ross's conjecture for the two-armed case.

Corollary 4.2. (Feldman.) *The myopic strategy* M^n *is optimal for* (ξ_0, n, x)*-bandits with utility function* $\varphi(x) = x$*, for any integer* $n \geq 1$*, for all* $\xi_0 \in [0, 1]$*,* $x \in \mathbb{R}$*, if and only if*

$$\mathbb{E}_{\mathbb{P}}[X_1 \mid H_1] \geq \mathbb{E}_{\mathbb{P}}[Y_1 \mid H_1].$$

Corollary 4.3. (Nouiehed and Ross's conjecture.) *Consider the two-armed Bernoulli bandits. Let the distributions* F_1 *and* F_2 *be* Bernoulli (α) *and* Bernoulli (β)*, and* $\alpha > \beta$*. Under hypothesis* H_1*, experiment X obeys* Bernoulli (α) *and Y obeys* Bernoulli (β)*; under hypothesis* H_2*, X obeys* Bernoulli (β) *and Y obeys* Bernoulli (α)*. Then, for any fixed positive integers k and n, Feldman's strategy* M^n *maximizes* $\mathbb{P}(S_n \geq k) = \mathbb{E}_{\mathbb{P}}[I_A]$*, where* I_A *is the indicator function of set* $A = \{S_n \geq k\}$*.*

Proof. Note that the proof of Theorem 4.1 also holds in the case of the distributions being Bernoulli distributions. In this case we only need to modify the calculation of expectations.

https://doi.org/10.1017/jpr.2023.24 Published online by Cambridge University Press

For any fixed k, let the utility function be $\varphi(x) = I_{[k,+\infty)}(x)$. By Remark 2, this φ satisfies condition (I), so Theorem 4.1 can be applied, and hence M^n maximizes the expectation $\mathbb{P}(S_n \geq k) = \mathbb{E}_{\mathbb{P}}[I_A] = \mathbb{E}_{\xi_0}[\varphi(S_n)]$. □

Remark 4.4. An anonymous reviewer reminded us that if the utility φ is increasing, then Nouiehed and Ross's conjecture for model (1.1) is in fact equivalent to proving that the myopic strategy maximizes $\mathbb{E}_{\xi_0}[\varphi(x + S_n)]$.

Acknowledgements

We sincerely thank Professor Jordan Stoyanov for his valuable suggestions and amendments to this paper. And we sincerely thank the anonymous reviewers for their help in improving this paper.

Funding information

Chen gratefully acknowledges the support of the National Key R&D Program of China (grant 2018YFA0703900), Shandong Provincial Natural Science Foundation, China (grant ZR2019ZD41), and Taishan Scholars Project. Lin gratefully acknowledges the support of the National Science Foundation of China (grant 12101367).

Competing interests

There were no competing interests to declare which arose during the preparation or publication process of this article.

References

[1] Abernethy, J., Hazan, E. and Rakhlin, A. (2008). Competing in the dark: an efficient algorithm for bandit linear optimization. In *21st Annual Conference on Learning Theory (COLT 2008)*, pp. 263–274.

[2] Agrawal, R. (1995). Sample mean based index policies by $o(\log n)$ regret for the multi-armed bandit problem. *Adv. Appl. Prob.* **27**, 1054–1078.

[3] Ahmad, S. H. A., Liu, M., Javidi, T., Zhao, Q. and Krishnamachari, B. (2009). Optimality of myopic sensing in multichannel opportunistic access. *IEEE Trans. Inform. Theory* **55**, 4040–4050.

[4] Audibert, J.-Y. and Bubeck, S. (2010). Regret bounds and minimax policies under partial monitoring. *J. Mach. Learn. Res.* **11**, 2785–2836.

[5] Audibert, J.-Y., Munos, R. and Szepesvári, C. (2009). Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoret. Comput. Sci.* **410**, 1876–1902.

[6] Auer, P., Cesa-Bianchi, N. and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.* **47**, 235–256.

[7] Auer, P., Cesa-Bianchi, N., Freund, Y. and Schapire, R. (1995). Gambling in a rigged casino: the adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th Annual Foundations of Computer Science*, pp. 322–331. IEEE Computer Society Press.

[8] Banks, J. S. and Sundaram, R. K. (1992). A class of bandit problems yielding myopic optimal strategies. *J. Appl. Prob.* **29**, 625–632.

[9] Berry, D. A. and Fristedt, B. (1985). *Bandit Problems: Sequential Allocation of Experiments*. Springer, Netherlands.

[10] 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.

[11] Cappé, O., Garivier, A., Maillard, O.-A., Munos, R. and Stoltz, G. (2013). Kullback–Leibler upper confidence bounds for optimal sequential allocation. *Ann. Statist.* **41**, 1516–1541.

[12] Chen, Z., Epstein, L. G. and Zhang, G. (2021). A central limit theorem, loss aversion and multi-armed bandits. Available at arXiv:2106.05472.

[13] Deo, S., Iravani, S., Jiang, T., Smilowitz, K. and Samuelson, S. (2013). Improving health outcomes through better capacity allocation in a community-based chronic care model. *Operat. Res.* **61**, 1277–1294.

[14] Feldman, D. (1962). Contributions to the 'two-armed bandit' problem. *Ann. Math. Statist.* **33**, 847–856.

https://doi.org/10.1017/jpr.2023.24 Published online by Cambridge University Press

[15] GARBE, R. AND GLAZEBROOK, K. D. (1998). Stochastic scheduling with priority classes. *Math. Operat. Res.* **23**, 119–144.

[16] GAST, N., GAUJAL, B. AND KHUN, K. (2022). Testing indexability and computing Whittle and Gittins index in subcubic time. Available at arXiv:2203.05207.

[17] GITTINS, J. AND JONES, D. (1974). A dynamic allocation index for the sequential design of experiments. In *Progress in Statistics*, ed. J. Gani, pp. 241–266. North-Holland, Amsterdam.

[18] GITTINS, J. AND WANG, Y.-G. (1992). The learning component of dynamic allocation indices. *Ann. Statist.* **20**, 1625–1636.

[19] GITTINS, J., GLAZEBROOK, K. AND WEBER, R. (2011). *Multi-Armed Bandit Allocation Indices*. John Wiley.

[20] HONDA, J. AND TAKEMURA, A. (2010). An asymptotically optimal bandit algorithm for bounded support models. In *23rd Conference on Learning Theory (COLT 2010)*, pp. 67–79.

[21] KAUFMANN, E. (2018). On Bayesian index policies for sequential resource allocation. *Ann. Statist.* **46**, 842–865.

[22] KELLEY, T. A. (1974). A note on the Bernoulli two-armed bandit problem. *Ann. Statist.* **2**, 1056–1062.

[23] KIM, M. J. AND LIM, A. E. (2015). Robust multiarmed bandit problems. *Manag. Sci* **62**, 264–285.

[24] KULESHOV, V. AND PRECUP, D. (2014). Algorithms for multi-armed bandit problems. Available at arXiv:1402.6028.

[25] LAI, T. AND ROBBINS, H. (1985). Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.* **6**, 4–22.

[26] LAI, T. L. (1987). Adaptive treatment allocation and the multi-armed bandit problem. *Ann. Statist.* **15**, 1091–1114.

[27] LEE, E., LAVIERI, M. S. AND VOLK, M. (2019). Optimal screening for hepatocellular carcinoma: a restless bandit model. *Manuf. Serv. Oper. Manag.* **21**, 198–212.

[28] MAHAJAN, A. AND TENEKETZIS, D. (2008). Multi-armed bandit problems. In *Foundations and Applications of Sensor Management*, ed. A. O. Hero *et al.*, pp. 121–151. Springer, Boston.

[29] NIÑO-MORA, J. (2007). A $(2/3)n^3$ fast-pivoting algorithm for the Gittins index and optimal stopping of a Markov chain. *INFORMS J. Computing* **19**, 596–606.

[30] NOUIEHED, M. AND ROSS, S. M. (2018). A conjecture on the Feldman bandit problem. *J. Appl. Prob.* **55**, 318–324.

[31] RODMAN, L. (1978). On the many-armed bandit problem. *Ann. Prob.* **6**, 491–498.

[32] RUSMEVICHIENTONG, P., SHEN, Z.-J. M. AND SHMOYS, D. B. (2010). Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Operat. Res.* **58**, 1666–1680.

[33] SCOTT, S. L. (2010). A modern Bayesian look at the multi-armed bandit. *Appl. Stoch. Models Business Industry* **26**, 639–658.

[34] THOMPSON, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* **25**, 285.

[35] WASHBURN, R. B. (2008). Application of multi-armed bandits to sensor management. In *Foundations and Applications of Sensor Management*, ed. A. O. Hero *et al.*, pp. 153–175. Springer, Boston.

[36] WEBER, R. R. AND WEISS, G. (1990). On an index policy for restless bandits. *J. Appl. Prob.* **27**, 637–648.

[37] WHITTLE, P. (1988). Restless bandits: activity allocation in a changing world. *J. Appl. Prob.* **25**, 287–298.

https://doi.org/10.1017/jpr.2023.24 Published online by Cambridge University Press