J. Appl. Prob. **55,** 318–324 (2018)
*doi:*10.1017/jpr.2018.19

A CONJECTURE ON THE FELDMAN BANDIT PROBLEM

MAHER NOUIEHED * ** AND
SHELDON M. ROSS,* *** *University of Southern California*

Abstract

We consider the Bernoulli bandit problem where one of the arms has win probability α and the others β, with the identity of the α arm specified by initial probabilities. With $u = \max(\alpha, \beta)$, $v = \min(\alpha, \beta)$, call an arm with win probability u a good arm. Whereas it is known that the strategy of always playing the arm with the largest probability of being a good arm maximizes the expected number of wins in the first n games for all n, we conjecture that it also stochastically maximizes the number of wins. That is, we conjecture that this strategy maximizes the probability of at least k wins in the first n games for all k, n. The conjecture is proven when $k = 1$, and $k = n$, and when there are only two arms and $k = n - 1$.

Keywords: Multi-armed bandit; stochastically maximizing

2010 Mathematics Subject Classification: Primary 60G40
Secondary 62L05

1. Introduction

Consider a sequence of n games, where in each game one of m arms is pulled, and either a win or a loss is obtained. We suppose that the set of win probabilities for the m arms is known to equal $\{\alpha, \beta, \ldots, \beta\}$, where $\alpha \neq \beta$ are specified probabilities. That is, it is known that exactly one of the arms has win probability α whereas all the others have win probability β. Although the identity of the arm with win probability α is not known, we assume there are initial probabilities as to which arm it is. This problem was introduced by Feldman [1] for the $m = 2$ case and its study for general m was undertaken by Rodman [2].

Let $u = \max(\alpha, \beta)$, $v = \min(\alpha, \beta)$, and call an arm with win probability u a good arm. With π being the strategy that always plays the arm with the largest posterior probability of being a good arm, it was shown in [1] and [2] that using strategy π maximizes the expected number of wins in the first n games for all n. We believe that the stronger result that π stochastically maximizes the number of wins in the first n games is also valid. That is, if T_n is the total number of wins in the first n games, then we believe that $\mathbb{P}(T_n \geq k)$ is maximized, for all k, n, when π is utilized. In Section 2 we prove this result when $k = 1$ and $k = n$. In Section 3 we show it is also true when $m = 2$ and $k = n - 1$. Final remarks are given in Section 4.

Received 30 November 2016; revision received 23 November 2017.
* Postal address: Department of Industrial and Systems Engineering, University of Southern California, Los Angeles, CA 90089, USA.
** Email address: nouiehed@usc.edu
*** Email address: smross@usc.edu

2. Maximizing the probabilities of at least one win, and of all wins

Let A_i be the event that arm i is a good arm, $i = 1, \ldots, n$, and suppose that p_i, $i = 1, \ldots, n$, are the initial probabilities for these events. (If $u = \alpha$ then $\sum_{i=1}^{n} p_i = 1$; if $u = \beta$ then $\sum_{i=1}^{n}(1 - p_i) = 1$.) Also, let W_j be the event that arm j is used first and results in a win, and let L_j be the event that arm j is used first and results in a loss.

Lemma 2.1. *It holds that*

$$\begin{aligned}
&\mathbb{P}(A_i \mid W_i) \geq p_i, \qquad \mathbb{P}(A_i \mid W_j) \leq p_i, \qquad i \neq j,\\
&\mathbb{P}(A_i \mid L_i) \leq p_i, \qquad \mathbb{P}(A_i \mid L_j) \geq p_i, \qquad i \neq j,\\
&p_i \geq p_r \implies \mathbb{P}(A_i \mid W_j) \geq \mathbb{P}(A_r \mid W_j), \qquad r \neq i \neq j,\\
&p_i \geq p_r \implies \mathbb{P}(A_i \mid L_j) \geq \mathbb{P}(A_r \mid L_j), \qquad r \neq i \neq j.
\end{aligned}$$

Proof. First suppose that $u = \alpha > \beta$. Then

$$\begin{aligned}
\mathbb{P}(A_i \mid W_i) &= \frac{\alpha p_i}{\alpha p_i + \beta(1 - p_i)} \geq p_i,\\
\mathbb{P}(A_i \mid W_j) &= \frac{\beta p_i}{\alpha p_j + \beta(1 - p_j)} \leq p_i, \qquad i \neq j,\\
\mathbb{P}(A_i \mid L_i) &= \frac{(1 - \alpha) p_i}{(1 - \alpha) p_i + (1 - \beta)(1 - p_i)} \leq p_i,\\
\mathbb{P}(A_i \mid L_j) &= \frac{p_i(1 - \beta)}{(1 - \alpha) p_j + (1 - \beta)(1 - p_j)} \geq p_i, \qquad i \neq j.
\end{aligned}$$

The preceding formulae also verify that if $p_i \geq p_r$ and $r \neq i \neq j$, then $\mathbb{P}(A_i \mid W_j) \geq \mathbb{P}(A_r \mid W_j)$ and $\mathbb{P}(A_i \mid L_j) \geq \mathbb{P}(A_r \mid L_j)$. The proof in the $u = \beta > \alpha$ case is similar. For instance,

$$\mathbb{P}(A_i^c \mid W_j) = \frac{(1 - p_i)\beta}{(1 - p_j)\alpha + p_j \beta} \geq 1 - p_i. \qquad \square$$

Proposition 2.1. *The policy π, which at every stage chooses the arm with the largest posterior probability of being a good arm, maximizes $\mathbb{P}(T_n \geq j)$ when either $j = 1$ or $j = n$.*

Proof. We prove that π maximizes $\mathbb{P}(T_n \geq 1)$ by induction on n. As it is immediate for $n = 1$, assume it when n games are played, and now assume $n + 1$ games are played. Suppose that $\mathbb{P}(A_1) = \max_i \mathbb{P}(A_i)$. We need to show that there is a policy that initially chooses arm 1 which maximizes $\mathbb{P}(T_{n+1} \geq 1)$.

Consider the best policy that starts with any other arm, say arm 2. If the initial result is a loss then, as Lemma 2.1 shows, arm 1 has the largest posterior probability of being a good arm, and, thus, the induction hypothesis implies that it is optimal to use arm 1 next. If the initial result is a win then it makes no difference which arm is used next (since the goal has been reached). Consequently, the best policy that starts with arm 2 will follow with arm 1. But we can match its probability of at least one win by using a policy that first chooses 1 and then 2. Hence, there is an optimal policy that starts with arm 1, which implies by the induction hypothesis that π is optimal.

The proof that π also maximizes $\mathbb{P}(T_n \geq n)$ is also by induction. As it is immediate for $n = 1$, assume for n, and now suppose that $n + 1$ games are to be played and that we want to maximize the probability of winning all $n + 1$. Again suppose that arm 1 has the largest

probability of being a good arm. First note that if arm 1 is initially used, then if it results in a win it will continue to be the arm with the largest posterior probability of being good, and if it results in a loss then it makes no difference what choices are made next. Consequently, we can conclude that the probability that policy π results in all successes is the probability of all successes when arm 1 is used throughout. Let $\pi_{1,1,\ldots,1}$ be the policy that uses arm 1 for all $n+1$ games. Now consider any policy that starts with arm 2. There are two cases.

Case 1. When a win by arm 2 would result in the posterior probability that arm 2 is good being greater than the posterior probability that arm 1 is good.

Case 2. Where a win with arm 2 would not make its posterior probability larger than the one for arm 1.

If case 1 holds then, by the induction hypothesis, the best policy that starts with arm 2 would have the same probability of obtaining all successes as the one that uses arm 2 throughout. Now if arm i is to be used for all $n+1$ trials then its probability of yielding all successes is $\mathbb{P}(A_i)u^{n+1} + (1-\mathbb{P}(A_i))v^{n+1}$. Since $\mathbb{P}(A_1) \geq \mathbb{P}(A_2)$, it thus follows in case 1 that the best policy that starts with arm 2 is not better than using arm 1 throughout.

If case 2 holds then since it makes no difference what follows arm 2 if it results in a loss, it follows from the induction hypothesis that the best policy that starts with arm 2 will then use arm 1 for the remaining n games. Call this policy $\pi_{2,1,\ldots,1}$. Now the probabilities of success under policies $\pi_{1,1,\ldots,1}$ and $\pi_{2,1,\ldots,1}$ are the same when arms 1 and 2 are either both good or both bad. That is,

$$\mathbb{P}_{\pi_{2,1,\ldots,1}}(T_{n+1} = n+1 \mid A_1A_2 \cup A_1^cA_2^c) = \mathbb{P}_{\pi_{1,1,\ldots,1}}(T_{n+1} = n+1 \mid A_1A_2 \cup A_1^cA_2^c).$$

Hence, with $D \equiv \mathbb{P}_{\pi_{1,1,\ldots,1}}(T_{n+1} = n+1) - \mathbb{P}_{\pi_{2,1,\ldots,1}}(T_{n+1} = n+1)$, we have

$$\begin{aligned} D &= (u^{n+1} - vu^n)\mathbb{P}(A_1A_2^c) + (v^{n+1} - uv^n)\mathbb{P}(A_1^cA_2) \\ &\geq (u^{n+1} - vu^n)\mathbb{P}(A_1^cA_2) + (v^{n+1} - uv^n)\mathbb{P}(A_1^cA_2) \\ &= (u^n - v^n)(u-v)\mathbb{P}(A_1^cA_2) \\ &\geq 0, \end{aligned}$$

where the preceding equation used the fact that $u > v$ and

$$\mathbb{P}(A_1A_2^c) = \mathbb{P}(A_1) - \mathbb{P}(A_1A_2) \geq \mathbb{P}(A_2) - \mathbb{P}(A_1A_2) = \mathbb{P}(A_1^cA_2).$$

Hence, in both cases we have shown that the best policy starting with arm 2 is not better than using arm 1 throughout. As arm 2 was arbitrary, the result follows. □

3. The case of two arms

Suppose that $m = 2$ and $\alpha > \beta$. Thus, the arm with win probability α is the good arm. The dynamic programming state of the system at any time is given by the triplet (i, n, p), where p is the posterior probability that arm 1 is the good arm, a total of n games remain, and our objective is to win at least i of these games. Let $V_{i,n}(p)$ be the maximal probability of winning at least i of the n games when p is the probability that arm 1 is the good arm. Note that, from the proof of Proposition 2.1, it follows that

$$V_{n,n}(p) = p\alpha^n + (1-p)\beta^n = \beta^n + p(\alpha^n - \beta^n), \qquad p \geq \tfrac{1}{2}.$$

Lemma 3.1. *It holds that $V_{i,n}(p)$ is a convex function of p that is symmetric about $p = \frac{1}{2}$. It is also a nondecreasing function of p for $p \geq \frac{1}{2}$.*

Proof. Fix i. It is immediate from its definition that $V_{i,n}(p)$ is symmetric about $p = \frac{1}{2}$. Let a deterministic policy be one whose initial choice is specified, and whose choice at any later time depends on all earlier choices and the results of those choices. That is, a deterministic policy is one whose decisions are not influenced by the initial value of p. Now, for a given initial probability p, any policy whose decisions are a function of the current state can be implemented by using a deterministic policy (with the appropriate deterministic policy depending on p). Consequently, it follows that

$$V_{i,n}(p) = \max_{d \in D} \mathbb{P}_{d \mid p}(T_n \geq i), \tag{3.1}$$

where D is the set of deterministic policies, and $\mathbb{P}_{d \mid p}(T_n \geq i)$ is the probability, when arm 1 is the good arm with probability p, that using policy d will result in at least i wins in the next n games. Since

$$\mathbb{P}_{d \mid p}(T_n \geq i) = p\mathbb{P}_{d \mid 1}(T_n \geq i) + (1 - p)\mathbb{P}_{d \mid 0}(T_n \geq i),$$

it follows that $\mathbb{P}_{d \mid p}(T_n \geq i)$ is a linear and, thus, convex function of p, and the convexity of $V_{i,n}(p)$ follows from (3.1) since the maximum of convex functions is convex. That $V_{i,n}(p)$ is nondecreasing in p for $p \geq \frac{1}{2}$ follows from its convexity, and the fact that $V_{i,n}(p)$ is symmetric about $\frac{1}{2}$. □

Then $V_{i,n}(p)$ satisfies the optimality equation

$$V_{i,n}(p) = \max\{H^1_{i,n}(p), H^2_{i,n}(p)\},$$

where $H^j_{i,n}(p)$ is the maximal probability of winning at least i of the following n games when p is the probability that arm 1 is the good arm and arm j is initially played, $j = 1, 2$. In addition, the policy that uses arm 1 in state (i, n, p) if $H^1_{i,n}(p) \geq H^2_{i,n}(p)$, and arm 2 otherwise, is an optimal policy.

3.1. Maximizing the probability of winning at least n out of $n + 1$ games

With p equal to the probability that arm 1 is the α arm, let $w_2(p) = \beta p/(\beta p + \alpha(1 - p))$ be the conditional probability that arm 1 is the α arm given that a pull of arm 2 results in a win.

Proposition 3.1. *The policy π, which at every stage chooses the arm with the largest posterior probability of being a good arm, maximizes $\mathbb{P}(T_{n+1} \geq n)$.*

Proof. We prove that π maximizes $\mathbb{P}(T_{n+1} \geq n)$ by induction on n. As it is immediate for $n = 0$, assume that it maximizes $\mathbb{P}(T_n \geq n - 1)$ and now consider the problem of maximizing $\mathbb{P}(T_{n+1} \geq n)$. Suppose that $p \geq \frac{1}{2}$. We will argue that there is an optimal policy that starts with arm 1. Since Proposition 2.1 along with the induction hypothesis imply that whatever the result of the first arm pulled, the optimal continuation is to use policy π, it follows that showing that there is an optimal policy that starts with arm 1 proves the theorem.

Among all policies that start with arm 2, let π' be one that maximizes the probability of reaching the goal of at least n successes in the $n + 1$ trials. We now argue that there is a policy that starts with arm 1 and has at least as large a probability of resulting in at least n wins as does π'. To show this, suppose that we start by pulling arm 2. There are two cases, depending on whether $w_2(p)$ is greater or less than $\frac{1}{2}$.

When $w_2(p) \geq \frac{1}{2}$, it follows by Proposition 2.1 that arm 1 will be pulled next if there is a loss and by the induction hypothesis that arm 1 will be pulled next if there is a win. Hence, whether or not the first play results in a win or a loss, arm 1 will be used on the second pull. But we can then match the probability of reaching the goal when π' is used by first using arm 1 and then arm 2. Thus, there is an optimal policy that starts with arm 1 in this case.

Now suppose that $w_2(p) < \frac{1}{2}$. If the initial pull of arm 2 results in a loss, then by Proposition 2.1 it is optimal to use arm 1 from then on. On the other hand, if arm 2 results in a win, then by the induction hypothesis it is optimal to use arm 2 again, and, in fact, to continue to use arm 2 until it results in a loss or $n+1$ successes have been obtained. If a loss does occur before $n+1$ consecutive successes, it follows by Proposition 2.1 that it is optimal after the loss to use from then on whichever arm has the higher probability of being the best. Since $w_2(p) < p$, it follows that the probability that arm 2 is the best arm is an increasing function of the number of wins obtained before the first loss of arm 2. Using this, it follows that there is a value $k \leq n$, such that the best policy that starts with arm 2 uses arm 2 until a loss occurs, and then either uses arm 1 from then on if the first loss occurs within the first k pulls or uses arm 2 from then on if the loss occurs after there have been at least k wins. Letting $\mathbb{P}_2^k$ be the probability that this policy results in at least n successes in $n+1$ trials, we have

$$\mathbb{P}_2^k = pB_k + (1-p)A_k,$$

where

$$A_k = \alpha^{n+1} + (1-\alpha)\sum_{i=1}^{k} \alpha^{i-1}\beta^{n+1-i} + (1-\alpha)(n+1-k)\alpha^n, \tag{3.2}$$

$$B_k = \beta^{n+1} + (1-\beta)\sum_{i=1}^{k} \beta^{i-1}\alpha^{n+1-i} + (1-\beta)(n+1-k)\beta^n. \tag{3.3}$$

But now consider the policy that starts with arm 1 and uses it until a loss, and then either uses arm 2 from then on if the first loss occurs within the first k pulls or uses arm 1 from then on if the loss occurs after there have been at least k wins. Letting $\mathbb{P}_1^k$ be the probability this policy results in at least n successes in $n+1$ trials, we have

$$\mathbb{P}_1^k = pA_k + (1-p)B_k.$$

Since $p \geq 1-p$, it follows that $\mathbb{P}_1^k \geq \mathbb{P}_2^k$ is equivalent to $A_k \geq B_k$. Thus, if we can show that with $F_k \equiv A_k - B_k \geq 0$, then it follows that there is an optimal policy that starts with arm 1, which proves the theorem. To show that $F_k \geq 0$, first note that

$$A_{k+1} - A_k = (1-\alpha)\alpha^k\beta^{n-k} - (1-\alpha)\alpha^n, \qquad B_{k+1} - B_k = (1-\beta)\beta^k\alpha^{n-k} - (1-\beta)\beta^n.$$

Hence, for $k < n$,

$$\begin{aligned} F_{k+1} - F_k &= (1-\alpha)\alpha^k\beta^{n-k} - (1-\alpha)\alpha^n - (1-\beta)\beta^k\alpha^{n-k} + (1-\beta)\beta^n \\ &= (\beta^{n-k} - \alpha^{n-k})((1-\beta)\beta^k + (1-\alpha)\alpha^k) \\ &\leq 0. \end{aligned}$$

Thus, $F_{k+1} \leq F_k$, and because, as will be shown in the following lemma, $F_n = 0$ it follows that $F_k \geq 0$. □

Lemma 3.2. *With A_k and B_k as given by (3.2) and (3.3), $A_n = B_n$.*

Proof. We need to show that $A_n = B_n$, where

$$A_n = \alpha^n + (1-\alpha)\sum_{i=1}^{n} \alpha^{i-1}\beta^{n+1-i}, \qquad B_n = \beta^n + (1-\beta)\sum_{i=1}^{n} \beta^{i-1}\alpha^{n+1-i}.$$

Let X_1 and X_2 be independent geometric random variables with respective parameters $1-\alpha$ and $1-\beta$. Conditioning on X_1 yields that $\mathbb{P}(X_1 + X_2 \geq n+2) = A_n$, and conditioning on X_2 yields that $\mathbb{P}(X_1 + X_2 \geq n+2) = B_n$. □

Thus, when $p \geq \frac{1}{2}$, there is an optimal policy that uses arm 1 until its first loss. If that loss occurs on pull k then the conditional probability that arm 1 is the α arm, call it $\mathbb{P}(1 \mid k)$, is

$$\mathbb{P}(1 \mid k) = \frac{\alpha^{k-1}(1-\alpha)p}{\alpha^{k-1}(1-\alpha)p + \beta^{k-1}(1-\beta)(1-p)}.$$

Letting

$$k(p) = \begin{cases} 0 & \text{if } \mathbb{P}(1 \mid 1) \geq \frac{1}{2}, \\ \max\{k\colon \mathbb{P}(1 \mid k) < \frac{1}{2}\} & \text{otherwise,} \end{cases}$$

it follows that if the first loss occurs before there have been $k(p)$ wins, then after that loss it is optimal to pull arm 2 from then on. (If a loss occurs when using arm 2 then, as it would be the second loss, there is no need to change back to arm 1.) Thus, if $k(p) = n$ then the optimal policy would use arm 1 until its first loss and then switch to using arm 2 from then on. However, by Lemma 3.2, one would obtain the same return by starting with arm 2 and using it until its first loss and then switching to arm 1 from then on. Consequently, it need not be uniquely optimal to use arm 1 when $p > \frac{1}{2}$. For example, if $n = 1$, $\alpha = 0.8$, and $\beta = 0.3$, then it is easy to check that $k(p) = 1$ for any $p \in [\frac{1}{2}, \frac{7}{9}]$. Since it is irrelevant which arm is used on the second play when the first results in a win (since the goal has been reached), it is optimal to use each arm once, showing that $V_{1,2}(p)$ is constant for $p \in [\frac{1}{2}, \frac{7}{9}]$. Thus, perhaps surprisingly, we see that it need not be uniquely optimal to use arm 1 when $p > \frac{1}{2}$ and $V_{n,n+1}(p)$ need not be a strictly increasing function of p when $p > \frac{1}{2}$.

4. Final remarks

As Feldman's original paper was published in 1962, one might wonder why the issue raised in this paper has never been considered before. We believe it is because of how the results of [1] and [2] were proven. It is easy to see, say in the two arm case when $\alpha > \beta$, that if N_n is equal to the number of times in the first n plays that the α arm is used, then

$$\mathbb{E}[T_n] = \alpha\mathbb{E}[N_n] + \beta(n - \mathbb{E}[N_n]) = (\alpha - \beta)\mathbb{E}[N_n] + n\beta.$$

Consequently, π maximizing $\mathbb{E}[T_n]$ is equivalent to its use maximizing $\mathbb{E}[N_n]$, and it is the latter that was actually proven in [1] and [2]. Thus, in thinking about π being stochastically optimal, our initial thought might be towards thinking about whether it stochastically maximizes N_n. Since this is clearly not the case (for instance, if $n = 2$ then the probability of using the best arm at least once in the two games is maximized by the strategy that plays each arm once), this may be a reason why the stochastic optimality possibility has not been considered.

We have been able to prove in the case of two bandits that the conjectured optimal policy maximizes the probability of at least two wins in the first four trials when $\beta = 1 - \alpha$.

A continuous-time version of our conjecture appeared in [3], where a proof, based on Pontryagin's maximum principle, was given in the case of two arms. The proof is, however, very difficult to read and assess.

Acknowledgements

We thank the anonymous referees for their many helpful comments, including a simplified proof of Proposition 3.1 and the existence of reference [3]. This material is based upon work supported by, or in part by, the National Science Foundation under contract/grant number CMMI1662442.

References

[1] Feldman, D. (1962). Contributions to the "two-armed bandit" problem. *Ann. Math. Statist.* **33,** 847–856.
[2] Rodman, L. (1978). On the many-armed bandit problem. *Ann. Prob.* **6,** 491–498.
[3] Presman, É. L. and Sonin, I. N. (1990). *Sequential Control With Incomplete Information: The Bayesian Approach to Multi-Armed Bandit Problems*. Academic Press, San Diego, CA.