Electronics Research Laboratory
University of California
Berkeley, California
Internal Technical Memorandum M-70

# A NOTE ON COUNTER-EXAMPLES TO A CONJECTURE CONCERNING THE TWO-ARMED BANDIT

by

B. Gluss

*This research reported herein is made possible through support received from the National Science Foundation grant G-15965/GP 2413.

May 19, 1964

# A NOTE ON COUNTER-EXAMPLES TO A CONJECTURE CONCERNING THE TWO-ARMED BANDIT*

Brian Gluss

University of California, Berkeley

## ABSTRACT

In the case of a two-armed bandit comprising two Bernouilli machines with known parameters $p_1$ and $p_2$, where it is not known which parameter pertains to which machine, Feldman has proved that, to maximize the total expected return over $r$ trials, that machine with the higher _a posteriori_ parameter should be chosen at each trial. It has been conjectured that an analogous result holds when only probability distributions for each machine's parameter are known, selecting that machine with the higher expected _a posteriori_ parameters. It is shown in this note that this conjecture is not generally correct.

*The research reported herein is made possible through support received from the National Science Foundation grant G-15965/GP 2413.

# A NOTE ON COUNTER-EXAMPLES TO A CONJECTURE CONCERNING THE TWO-ARMED BANDIT*

Brian Gluss

Electronics Research Laboratory
University of California, Berkeley

## INTRODUCTION

Consider a two-armed bandit comprising two Bernouilli machines with known parameters $p_1$ and $p_2$, where it is not known which parameter pertains to which machine. Feldman[1] has proved that in order to maximize the total expected return over $r$ trials, $r$ fixed, in which a trial produces scores

$$\left.\begin{array}{ll} 1 & \text{with probability } p_i \\ 0 & \text{with probability } 1 - p_i \end{array}\right\} \qquad (1)$$

according as to which machine $i$ $(= 1, 2)$ is used, the optimal policy is to select at each trial that machine which has the higher *a posteriori* probability $p$ of scoring a 1.

Suppose instead that, as in Gluss,[2] the Bernouilli machines I and II, have *a priori* distributions for their parameters $p$ and $p'$ respectively of the forms

*The research reported herein is made possible through support received from the National Science Foundation grant G-15965/GP 2413.

$$dG(p) = \frac{p^{a-1}(1 - p)^{b-1}}{B(a, b)}\, dp \quad , \tag{2}$$

and

$$dG'(p) = \frac{(p')^{a'-1}(1 - p')^{b'-1}}{B(a', b')}\, dp' \, , \tag{3}$$

so that, after m 1's and n 0's with I, the Bayes a posteriori distribution for p is given by

$$dG_{m,n}(p) = \frac{p^m (1 - p)^n\, dG(p)}{\int_0^1 x^m (1 - x)^n\, dG(x)} \quad ,$$

i.e.,

$$dG_{m,n}(p) = \frac{p^{a+m-1}(1 - p)^{b+n-1}}{B(a+m,\ b+n)}\, dp \quad , \tag{4}$$

with expectation

$$p_{m,n} = \frac{m + a}{(m+a) + (n+b)} \quad , \tag{5}$$

with similar expressions for $dG'_{m',n'}(p')$ and $p'_{m',n'}$. The $B(u, v)$ are beta distributions.

In this case, it has been conjectured by L. A. Zadeh and others that a result analogous to that in Feldman's model holds. That is, in order to maximize the total expected return for an r-stage process, the optimal policy is to choose at each trial that machine with the higher expected a posteriori probability of scoring a 1. That is,

$$\text{choose machine I if and only if } \frac{m + a}{(m+a) + (n+b)} > \frac{m' + a'}{(m'+a') + (n'+b')} \, . \tag{6}$$

We shall show below that this conjecture does not always hold, although it does over large regions of the $(m, n, m', n')$ space.

## THE FUNCTIONAL EQUATION

It will be notationally convenient to introduce

$$\left.\begin{array}{lll} M = m+a, & N = n+b, & R = M + N, \\ M' = m'+a', & N' = n'+b', & R' = M' + N'. \end{array}\right\} \quad (7)$$

The policy of Eq. (6) then becomes

$$\text{choose I if and only if} \qquad \frac{M}{R} > \frac{M'}{R'} . \qquad (8)$$

Let $f^{r}_{M, N, M', N'}$ = expected total return for a process with $r$ trials remaining, after $M - a$ 1's and $N - b$ 0's with I and $M' - a'$ 1's and $N' - b'$ 0's with II, using an optimal policy.

Then we have the functional equation

$$f^{r}_{M, N, M', N'} = \text{Max} \left[ \begin{array}{l} \text{I:} \quad \int_0^1 p \left[ 1 + f^{r-1}_{M+1, N, M', N'} \right] dG_{m, n}(p) \\ \qquad + \int_0^1 (1-p)\, f^{r-1}_{M, N+1, M', N'} \, dG_{m, n}(p); \\ \text{II:} \quad \int_0^1 p' \left[ 1 + f^{r-1}_{M, N, M'+1, N'} \right] dG'_{m', n'}(p') \\ \qquad + \int_0^1 (1-p')\, f^{r-1}_{M, N, M', N'+1} \, dG'_{m', n'}(p') \end{array} \right]$$

That is,

$$\phi^{r}_{M,N,M',N'} = \mathrm{Max}\begin{bmatrix} \mathrm{I:}\ \frac{M}{R}\left[1+\phi^{r-1}_{M+1,N,M',N'}\right] + \frac{N}{R}\,\phi^{r-1}_{M,N+1,M',N'} \\ \\ \mathrm{II:}\ \frac{M'}{R'}\left[1+\phi^{r-1}_{M,N,M'+1,N'}\right] + \frac{N'}{R'}\,\phi_{M,N,M',N'+1} \end{bmatrix} \qquad (9)$$

We have the boundary conditions $\phi^{o} = 0$, and the policy of Eq. (8) is evidently optimal for $r = 1$, and

$$\phi^{1}_{M,N,M',N'} = \mathrm{Max}\left[\frac{M}{R}\,,\ \frac{M'}{R'}\right]. \qquad (10)$$

## COUNTER-EXAMPLES FOR $r = 2$

We now consider the two-stage process $(r = 2)$, and assume that initially

$$\frac{M}{R} > \frac{M'}{R'}\ . \qquad (11)$$

It will be shown that the policy of Eq. (8) at the first trial--i.e., choose machine I--is not optimal for all $M, M', N, N'$.

Using Eq. (10), and remembering that inequality (11) implies that we also have

$$\left.\begin{aligned} &\frac{M+1}{R+1} > \frac{M'}{R'} \\ \text{and}\qquad & \\ &\frac{M}{R} > \frac{M'}{R'+1}\ , \end{aligned}\right\} \qquad (12)$$

Eq. (9) reduces to

$$\mathcal{f}^{2}_{M,N,M',N'} = \operatorname{Max}\left[\begin{array}{l} \text{I: } \frac{M}{R}\left[1+\frac{M+1}{R+1}\right] + \frac{N}{R}\operatorname{Max}\left\{\frac{M}{R+1}, \frac{M'}{R'}\right\}; \\ \\ \text{II: } \frac{M'}{R'}\left[1+\operatorname{Max}\left\{\frac{M}{R}, \frac{M'+1}{R'+1}\right\}\right] + \frac{N'}{R'}\left[\frac{M}{R}\right] \end{array}\right] \tag{13}$$

We discuss below the four cases $\frac{M}{R+1} \lessgtr \frac{M'}{R'}$ ; $\frac{M}{R} \lessgtr \frac{M'+1}{R'+1}$, all consistent with inequality (11), determining expressions for I - II, and showing that if

$$\frac{M}{R} \geq \frac{M'+1}{R'+1}, \tag{14}$$

policy (8) is optimal, while otherwise this is not necessarily the case. That is, for

$$\frac{M'}{R'} < \frac{M}{R} < \frac{M'+1}{R'+1}, \tag{15}$$

policy (8) is not necessarily optimal.

(1) $\underline{\frac{M}{R} \geq \frac{M'+1}{R'+1}, \quad \frac{M}{R+1} < \frac{M'}{R'}}$

In this case,

$$\begin{aligned} \text{I - II} &= \frac{M}{R} + \frac{M(M+1)}{R(R+1)} + \frac{N}{R}\cdot\frac{M'}{R'} - \frac{M'}{R'} - \frac{M'}{R'}\cdot\frac{M}{R} - \frac{N'}{R'}\cdot\frac{M}{R} \\ &= \frac{M}{R} - \frac{M'}{R'} + \frac{M(M+1)}{R(R+1)} + \frac{N}{R}\cdot\frac{M'}{R'} - \frac{M}{R} \\ &= \frac{M}{R} - \frac{M'}{R'} - \frac{M}{R}\cdot\frac{N}{R+1} + \frac{N}{R}\cdot\frac{M'}{R'} \\ &= \left(\frac{M}{R} - \frac{M'}{R'}\right) + \frac{N}{R}\left(\frac{M'}{R'} - \frac{M}{R+1}\right). \end{aligned}$$

Since $\frac{M'}{R'} > \frac{M}{R+1}$ and $\frac{M}{R} > \frac{M'}{R'}$, here we have I > II.

(2) $\underline{\frac{M}{R} \geq \frac{M'+1}{R'+1}, \quad \frac{M}{R+1} > \frac{M'}{R'}}$

We now have

$$\mathrm{I} - \mathrm{II} = \frac{M}{R} + \frac{M(M+1)}{R(R+1)} + \frac{N}{R} \cdot \frac{M}{R+1} - \frac{M'}{R'} - \frac{M'}{R'} \cdot \frac{M}{R} - \frac{N'}{R'} \cdot \frac{M}{R}$$

$$= \frac{M}{R} - \frac{M'}{R'},$$

which is once again greater than zero. We now reach the two questionable cases.

(3) $\underline{\frac{M}{R} < \frac{M'+1}{R'+1}, \quad \frac{M}{R+1} < \frac{M'}{R'}}$

Here we have

$$\mathrm{I} - \mathrm{II} = \frac{M}{R} + \frac{M(M+1)}{R(R+1)} + \frac{N}{R} \cdot \frac{M'}{R'} - \frac{M'}{R'} - \frac{M'(M'+1)}{R'(R'+1)} - \frac{N'}{R'} \cdot \frac{M}{R}$$

$$= \frac{M(M+1)}{R(R+1)} - \frac{M'(M'+1)}{R'(R'+1)}.$$

Hence, if

$$\left.\begin{array}{ll} & \frac{M(M+1)}{R(R+1)} < \frac{M'(M'+1)}{R'(R'+1)}, \\ & \frac{M}{R} < \frac{M'+1}{R'+1}, \\ & \frac{M}{R+1} < \frac{M'}{R'}, \\ \text{and} & \frac{M}{R} > \frac{M'}{R'}, \end{array}\right\} \qquad (16)$$

which are mutually consistent inequalities, in this situation the conjecture does not hold. Of course, in case (3), the conjecture holds when

$$\frac{M(M+1)}{R(R+1)} > \frac{M'(M'+1)}{R'(R'+1)} .$$

(4) $\frac{M}{R} < \frac{M'+1}{R'+1}$ , $\frac{M}{R+1} > \frac{M'}{R'}$

In this final case,

$$I - II = \frac{M}{R} + \frac{M(M+1)}{R(R+1)} + \frac{N}{R} \cdot \frac{M}{R+1} - \frac{M'}{R'} - \frac{M'(M'+1)}{R'(R'+1)} - \frac{N}{R'} \cdot \frac{M}{R}$$

$$= \left(\frac{M}{R} - \frac{M'}{R'}\right) + \frac{M'}{R'}\left(\frac{M}{R} - \frac{M'+1}{R'+1}\right) .$$

When this expression is greater than zero, the conjecture holds. However, when

$$\left.\begin{array}{l} \frac{M}{R} - \frac{M'}{R'} < \frac{M'}{R'}\left(\frac{M'+1}{R'+1} - \frac{M}{R}\right) , \\ \frac{M}{R} < \frac{M'+1}{R'+1} , \\ \frac{M}{R+1} > \frac{M'}{R'} , \\ \text{and} \quad \frac{M}{R} > \frac{M'}{R'} , \end{array}\right\} \qquad (17)$$

once again the conjecture is false.

We have thus found that, while it holds under all other conditions, the conjecture fails under the two sets of conditions (16) and (17).

## TWO NUMERICAL EXAMPLES

(a) For $M = 17$, $N = 11$, $M' = 3$, $N' = 2$, the inequalities (16) hold, since

$$\frac{17.18}{28.29} < \frac{3.4}{5.6}, \quad \frac{17}{28} < \frac{4}{6}, \quad \frac{17}{29} < \frac{3}{5}, \text{ and } \frac{17}{28} > \frac{3}{5}.$$

In this case

$$I = \frac{17}{28}\left(1 + \frac{18}{29}\right) + \frac{11}{28} \cdot \frac{3}{5} = \frac{1}{70}\left(85\,\frac{11}{29}\right).$$

$$II = \frac{3}{5}\left(1 + \frac{4}{6}\right) + \frac{3}{5} \cdot \frac{17}{28} = \frac{1}{70}\,(87),$$

so that although $\frac{17}{28} > \frac{3}{5}$, $I < II$.

(b) Let $M = 50$, $N = 31$, $M' = 3$, $N' = 2$. Then the inequalities (17) hold, since

$$\frac{50}{81} - \frac{3}{5} < \frac{3}{5}\left(\frac{4}{6} - \frac{50}{81}\right), \quad \frac{50}{81} < \frac{4}{6}, \quad \frac{50}{82} > \frac{3}{5}, \text{ and } \frac{50}{81} > \frac{3}{5}.$$

Here we have

$$I = \frac{50}{81}\left(1 + \frac{51}{82}\right) + \frac{31}{81} \cdot \frac{50}{82} = \frac{100}{81}.$$

$$II = \frac{3}{5}\left(1 + \frac{4}{6}\right) + \frac{2}{5} \cdot \frac{50}{81} = \frac{101}{81}.$$

## GENERAL COMMENTS

For general $r$, one might hope for simple conditions under which the maximum-expected-a-posteriori-policy holds, and correspondingly simple conditions under which it does not. However, as $r$ increases these conditions become progressively more complicated, comprising larger and larger numbers of inequalities on M, R, M', R'. Moreover, the sum of these regions does not appear to tend to a null region as $r \to \infty$ (when we introduce a discount factor into the model), so that even in the asymptotic case of an infinite number of trials, the conjecture appears false.

It might be of interest to give--for brevity, without proofs--some simple regions for which the conjecture holds.

$$\text{(i)} \qquad \frac{M}{R+r-1} > \frac{M'+r-1}{R'+r-1} \, . \tag{18}$$

This condition is intuitively obvious, since the a posteriori probabilities of machine I are then always greater than those of machine II, whichever machine is used at any trial and whatever the result of its use.

$$\text{(ii)} \qquad \frac{M}{R+r-1} > \frac{M'}{R'} \quad \text{and} \quad N + r - 2 < M' . \tag{19}$$

$$\text{(iii)} \qquad \frac{M}{R} > \frac{M'+r-1}{R'+r-1} \quad \text{and} \quad N > M' + r - 2 \, . \tag{20}$$

In both these cases, (ii) and (iii), the inequalities imply that

$$\frac{M}{R+r-1-i} > \frac{M'+i}{R'+i} \, , \qquad i = 0, 1, \ldots, r-1, \tag{21}$$

so that losing at each stage with machine I still leaves us with higher a posteriori probabilities than winning at each stage with machine II.

## REFERENCES


1. FELDMAN, D., (1962). Contributions to the two-armed bandit problem. Annals of Math. Stat. 33, 847-856.

2. GLUSS, B., (1962). A note on a computational approximation to the two-machine problem. Inf. and Control 5, 268-275.