A Dynamic Allocation Index for the Discounted Multiarmed Bandit Problem
Author(s): J. C. Gittins and D. M. Jones
Source: *Biometrika*, Vol. 66, No. 3 (Dec., 1979), pp. 561-565
Published by: Oxford University Press on behalf of Biometrika Trust
Stable URL: https://www.jstor.org/stable/2335176
Accessed: 06-08-2019 19:34 UTC

Biometrika (1979), **66**, 3, *pp*. 561–5
Printed in Great Britain

A dynamic allocation index for the discounted multiarmed bandit problem

By J. C. GITTINS
Mathematical Institute, Oxford

and D. M. JONES
Department of Mathematics, Polytechnic of Wales, Llantrisant, Mid-Glamorgan

Summary

Earlier work by the present authors has established the existence of and a characterization of a priority index giving the Bayes rule for the discounted multiarmed bandit problem. The calculation of this index is described and illustrated, and the results obtained briefly discussed.

Some key words: Dynamic allocation index; Sequential decision procedure; Two-armed bandit.

1. Introduction

The two-armed bandit problem is so-called because it models the situation faced by a gambler using a fruit machine with two arms, instead of just one. When an arm is pulled the result is that the gambler either wins a prize or not. All the prizes are of the same value, and for each arm there is a certain constant, and in general unknown, probability of success every time it is pulled, which is different for the two arms. The gambler's problem is to choose a sequence of pulls on the two arms, which depends in a sequential manner on the record of successes and failures, in such a fashion as to maximize his expected total gains. In one version of the problem the gambler is allowed a fixed number of pulls in total. Alternatively we may consider a discounted version of the problem for which the value of a prize received at the tth pull, irrespective of the arms pulled on the previous $(t-1)$ pulls, is multiplied by a^{t-1}, for some $a < 1$. Multiarmed bandit problems are similar, but with more than two arms. Their chief practical motivation comes from clinical trials, though they are also of interest as probably the simplest worthwhile set of problems in the sequential design of experiments.

Recent numerical investigation of Bayesian rules for the two-armed bandit problem with independent arms by P. W. Jones (1975) and Wahrenberger, Antle & Klimko (1977) have shown significant improvements when compared with other rules which have been proposed. However, the calculation of these rules is costly in terms of computer storage and time, and the reported results are all for a total number of trials which does not exceed 50.

The present authors (1974) showed that for the discounted case with an infinite number of trials, first considered by Bellman (1956), the Bayes rule is given by a function which, for each arm, depends on the posterior distribution for the unknown success probability. The Bayes rule is always to pull the arm for which the current value of the function is larger, for which reason the function was termed a dynamic allocation index. Moreover, this result holds for the multiarmed bandit problem. Gittins (1979) obtained a characterization of the dynamic allocation index which actually holds for a range of problems of which the multiarmed bandit is just one example. The present paper reports on a numerical investigation of the dynamic allocation index using this characterization, which shows that some further perfectly manageable computation along similar lines is all that is required for the problem to be, for practical purposes, completely solved.

Section 2 describes the above-mentioned characterization of the dynamic allocation index for the case of the multiarmed bandit, and the computational algorithm to which it leads. Section 3 describes the numerical results and §4 discusses asymptotic properties.

2. The dynamic allocation index and its calculation

Let the prior probability density for the unknown success probability θ for any arm be $\Gamma(\alpha)\,\Gamma(\beta)\,\{\Gamma(\alpha+\beta)\}^{-1}\,\theta^{\alpha-1}(1-\theta)^{\beta-1}$, where $\alpha>0$ and $\beta>0$, that is a beta (α,β) density, and independent of the prior distributions for the other arms. It follows from Bayes's theorem that after n trials on this arm with r successes the posterior distribution for θ is beta $(\alpha+r,\beta+n-r)$, and the state of knowledge of the arm at any stage of the sequence of trials may be summarized by the parameters of the current beta posterior distribution. Although not required by the general theory of dynamic allocation indices, the assumption of a beta distribution is essential for their calculation to be tractable. It allows an arbitrary specification of the mean and variance and, as pointed out by Raiffa & Schlaifer (1961, § 3.1.1), this is for many purposes quite adequate.

Consider now a sequence of τ consecutive trials on a single arm A, starting when its state is (α,β). Let $\tau = \min\,[t>0\colon \{\alpha(t),\beta(t)\}\in\Theta_0(t)]$, where $\{\alpha(t),\beta(t)\}$ is the state of the arm after t trials and $\Theta_0(t)$ is a subset of the set $\{(\gamma,\delta)\colon \gamma+\delta=\alpha+\beta+t\}$ $(t=1,2,\ldots)$; thus τ may be regarded as the stopping time for the sequence of trials on A defined by the stopping set

$$\bigcup_{t=1}^{\infty}\Theta_0(t).$$

Let $R_\tau(\alpha,\beta,a)$ be the expected total discounted reward from the τ trials on A, when each successful trial yields a reward of 1, and a reward at the tth trial is discounted by the factor a^{t-1}. Let

$$W_\tau(\alpha,\beta,a) = E\left(\sum_{t=0}^{\tau-1}a^t\right)$$

and finally

$$\nu(\alpha,\beta,a) = \sup_{\tau} R_\tau(\alpha,\beta,a)/W_\tau(\alpha,\beta,a), \tag{1}$$

where the supremum is over all stopping times τ defined as above. The quantity $\nu(\alpha,\beta,a)$ turns out (Gittins, 1979) to be the dynamic allocation index for any arm of a multiarmed bandit in state (α,β) when a is the discount factor.

Thus in a multiarmed bandit for which a success at the tth trial yields a discounted reward a^{t-1}, and with n arms which at any given stage are in states (α_i,β_i) $(i=1,\ldots,n)$, under a Bayes rule the arm j which is tried next must be such that

$$\nu(\alpha_j,\beta_j,a) = \max_{1\leqslant i\leqslant n}\nu(\alpha_i,\beta_i,a).$$

It is important, to avoid confusion, to realize that the τ trials mentioned in the previous paragraph do not necessarily take place under a Bayes rule. Their purpose is simply the conceptual one of enabling us to define the form of the dynamic allocation index.

Equation (1) expresses $\nu(\alpha,\beta,a)$ as the supremum over all stopping times τ of a kind of average reward per trial up to the τth trial. For an arm for which θ is known to take the value p this average must be equal to p, which is therefore the dynamic allocation index for such an arm. In general, the probability of success at the first trial for an arm in state (α,β) is $\alpha/(\alpha+\beta)$, that is the mean of a beta (α,β) distribution, and $\nu(\alpha,\beta,a)$ is greater than this, since $\alpha/(\alpha+\beta)$ is the value of the average just mentioned when $\text{pr}(\tau = 1) = 1$.

The basic idea of the method of calculation is to obtain successive approximations to $\nu(\alpha,\beta,a)$ by evaluating the right-hand side of (1) for stopping times τ which are restricted to be less than or equal to T, and then increasing T. Thus let

$$\nu^T(\alpha,\beta,a) = \sup_{\tau\leqslant T} R_\tau(\alpha,\beta,a)/W_\tau(\alpha,\beta,a).$$

Clearly

$$\nu(\alpha,\beta,a) = \lim_{T\to\infty} \nu^T(\alpha,\beta,a).$$

Gittins (1979) gives the details of the algorithm.

3. Results

The calculations described in the previous section were carried out for a discount rate $a = 0{\cdot}75$, leading to values of the function $\nu(\alpha,\beta,0{\cdot}75)$ correct to four decimal places for all integer values of $\alpha+\beta$ up to 140. The results for values of α and β between one and ten are shown in Table 1.

Table 1. *Values of* $\nu(\alpha,\beta,0{\cdot}75)$

β	$\alpha = 1$	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$	$\alpha = 5$	$\alpha = 6$	$\alpha = 7$	$\alpha = 8$	$\alpha = 9$	$\alpha = 10$
1	0·6211	0·7465	0·8062	0·8419	0·8659	0·8833	0·8965	0·9069	0·9153	0·9223
2	0·4256	0·5760	0·6607	0·7159	0·7548	0·7841	0·8068	0·8251	0·8401	0·8526
3	0·3182	0·4641	0·5554	0·6191	0·6659	0·7023	0·7312	0·7548	0·7745	0·7912
4	0·2519	0·3871	0·4773	0·5436	0·5946	0·6348	0·6673	0·6946	0·7176	0·7372
5	0·2073	0·3307	0·4182	0·4838	0·5360	0·5784	0·6134	0·6428	0·6678	0·6896
6	0·1755	0·2883	0·3713	0·4359	0·4875	0·5306	0·5669	0·5978	0·6244	0·6476
7	0·1518	0·2550	0·3334	0·3961	0·4473	0·4899	0·5266	0·5584	0·5860	0·6102
8	0·1335	0·2285	0·3025	0·3627	0·4129	0·4553	0·4916	0·5236	0·5518	0·5767
9	0·1190	0·2067	0·2767	0·3343	0·3832	0·4249	0·4611	0·4928	0·5212	0·5465
10	0·1072	0·1886	0·2547	0·3100	0·3573	0·3983	0·4341	0·4657	0·4937	0·5192

For values of α and β such that the probability $\alpha/(\alpha+\beta) = \lambda$, say, of success on the next trial for an arm in state (α,β) is fixed, it was found that $\nu(\alpha,\beta,0{\cdot}75)$ may be closely approximated by a function of the form

$$\lambda+\{A(\lambda)+B(\lambda)(\alpha+\beta)\}^{-1}. \tag{2}$$

The values of $A(\lambda)$ and $B(\lambda)$ in Table 2 were obtained. For fixed values of $\alpha+\beta$, values of $\nu(\alpha,\beta,0{\cdot}75)$ were interpolated between those given by (2), assuming $\nu(\alpha,\beta,0{\cdot}75)$ to be a linear function of $\lambda(1-\lambda)$. For integer-valued α and β these values agreed with those calculated by the algorithm to within 2×10^{-4} for $\alpha+\beta\geqslant 10$, and to within 2×10^{-3} for $\alpha+\beta\geqslant 2$. Thus it seems reasonable to interpolate in this way even when α and β are not integers.

Table 2. *Approximation* (2) *for* $\nu(\alpha,\beta,0{\cdot}75)$

λ	0·0125	0·025	0·05	0·1	0·2	0·3	0·4	0·5	0·6	0·7	0·8	0·9	0·95	0·975	0·9875
$A(\lambda)$	40	20	10	6	3·5	3·4	3·35	3·5	4·2	5·6	8·3	18	40	80	160
$B(\lambda)$	30	15	8	4·8	3·16	2·65	2·45	2·44	2·45	2·65	3·16	4·8	8	15	30

Curves of constant dynamic allocation index, iso-DAI's, were calculated by this method for $\nu(\alpha,\beta,0{\cdot}75) = (0{\cdot}1, 0{\cdot}2, \ldots, 0{\cdot}9)$. Each iso-DAI is asymptotic to a straight line whose slope is the corresponding value of the dynamic allocation index, as is easily deduced from (2). If we define $\Delta(\alpha,\beta)$ by the equation

$$\nu(\alpha,\beta,0{\cdot}75) = \{\alpha+\Delta(\alpha,\beta)\}/(\alpha+\beta),$$

then $\Delta(\alpha,\beta)$ tends to the limit $\{B(\lambda)\}^{-1}$ as $\alpha+\beta$ tends to infinity along the iso-DAI on which $\nu(\alpha,\beta,0{\cdot}75) = \lambda$.

The quantity $\Delta(\alpha,\beta)$ has an interesting interpretation. It is, for a fixed value of $\alpha+\beta$, the increase in α which changes the probability of success $\alpha/(\alpha+\beta)$ on the next trial in state (α,β) by as much as the difference between $\nu(\alpha,\beta,0{\cdot}75)$ and $\alpha/(\alpha+\beta)$. As pointed out in § 2, for an arm whose success probability is known, this probability, which clearly is the probability of success at the next trial at any stage, is equal to the dynamic allocation index. Thus $\Delta(\alpha,\beta)$ is a measure of the effect of uncertainty about the unknown success probability θ on the Bayes rule. Some examples of the values which it takes on the iso-DAI's are shown in Table 3. It is perhaps surprising that the effect of uncertainty is so small, indicating that a rule which used $\alpha/(\alpha+\beta)$ in place of $\nu(\alpha,\beta,0{\cdot}75)$ would be almost as good as the Bayes rule. For values of a nearer to one, the function $\nu(\alpha,\beta,a)$ turns out to be of similar form, though, as is to be expected, $\Delta(\alpha,\beta)$ increases with a. However, even for $a = 0{\cdot}99$, which is nearing the limits of computational feasibility, the largest value of $\Delta(\alpha,\beta)$ is only slightly greater than 3·0.

Table 3. *The uncertainty effect* $\Delta(\alpha,\beta)$

	$\nu(\alpha,\beta,0{\cdot}75)$								
$\alpha+\beta$	0·1	0·2	0·3	0·4	0·5	0·6	0·7	0·8	0·9
10	0·162	0·264	0·322	0·354	0·361	0·356	0·328	0·271	0·172
25	0·188	0·294	0·353	0·384	0·388	0·384	0·354	0·295	0·191
∞	0·208	0·316	0·377	0·408	0·410	0·408	0·377	0·316	0·208

This important conclusion parallels a finding by P. W. Jones (1975) for the undiscounted two-armed bandit problem. His calculations show that selecting the arm with the higher current value of $\alpha/(\alpha+\beta)$ is more than 99% efficient when compared with the Bayes rule, for total numbers of trials up to 15, the largest number which he considers for the case when the success probabilities are unknown for both arms.

For undiscounted two-armed bandit problems with one arm known to have a probability p of success at each trial, optimal policies may be described by functions $\nu_N(\alpha,\beta)$ and $\Delta_N(\alpha,\beta)$ with the following properties: N is the number of further trials allowed; (α,β) is the current beta parameter for the arm with unknown success probability; it is optimal to pull the arm with success probability p if and only if $p \geqslant \nu_N(\alpha,\beta)$; and for $\Delta_N(\alpha,\beta) > 0$

$$\nu_N(\alpha,\beta) = \{\alpha+\Delta_N(\alpha,\beta)\}/(\alpha+\beta).$$

Thus $\nu_N(\alpha,\beta)$ is analogous to $\nu(\alpha,\beta,a)$ and $\Delta_N(\alpha,\beta)$ is the corresponding uncertainty effect. Unpublished calculations by P. W. Jones show that $\Delta_N(\alpha,\beta)$ is less than 3·0 for all values of N, α and β, such that $N+\alpha+\beta \leqslant 50$, $\alpha > 0$ and $\beta > 0$, and depends on α and β in a similar way to $\Delta(\alpha,\beta)$. However, it should be noted that, unlike the dynamic allocation index, $\nu_N(\alpha,\beta)$ does not lead directly to the optimal policy when both success probabilities are unknown, or when there are more than two arms.

4. Asymptotic behaviour

J. C. Gittins has shown in an as yet unpublished paper that for all α and β, $\nu(\alpha, \beta, a)$ is an increasing function of a, tends to $\alpha/(\alpha+\beta)$ as a tends to zero, and to one as a tends to one.

As noted in § 2, $\nu(\alpha, \beta, a)$ is the supremum over all stopping times τ of a kind of average reward per trial up to the τth trial. If $\text{pr}(\tau = 1) = 1$ this average is equal to $\alpha/(\alpha+\beta)$. If $\text{pr}(\tau > 1) > 0$, the average reward per trial on trials after the second and up to the τth trial may be greater than $\alpha/(\alpha+\beta)$, bringing the overall average up to the τth trial above $\alpha/(\alpha+\beta)$, so that, in turn, $\nu(\alpha, \beta, a) > \alpha/(\alpha+\beta)$. The reason why $\nu(\alpha, \beta, a)$ increases with a is that the effect of later trials is more important for large values of a, so the effect of this averaging up is greater.

That $\nu(\alpha, \beta, a)$ should actually tend to one as a tends to one is perhaps rather more surprising. This implies that for values of a sufficiently close to one, and sufficiently large values of α and β, the uncertainty effect $\Delta(\alpha, \beta)$ must take arbitrarily large values, even though it is never greater than 4·0 when $a = 0{\cdot}99$. The behaviour of the iso-DAI's in the limit as a tends to one is an intriguing open question. However, it can be shown that on any iso-DAI, $\alpha/(\alpha+\beta)$ tends to the dynamic allocation index as $\alpha+\beta$ tends to infinity.

We have already noted that the dynamic allocation index coincides with θ when θ is known. As $\alpha+\beta$ tends to infinity along an iso-DAI, the posterior probability that θ is close to $\alpha/(\alpha+\beta)$ increases. The result at the end of the last paragraph is the consequence for the dynamic allocation index of this convergence in probability. It also follows from the empirical result that $\Delta(\alpha, \beta)$ tends to a limit as $\alpha+\beta$ tends to infinity along an iso-DAI. However, an analytic proof of this result itself does not seem easy to find.

References

Bellman, R. E. (1956). A problem in the sequential design of experiments. *Sankhyā* **16**, 221–9.

Gittins, J. C. (1979). Bandit processes and dynamic allocation indicies (with discussion). *J. R. Statist. Soc.* B **41**, 148–77.

Gittins, J. C. & Jones, D. (1974). A dynamic allocation index for the sequential design of experiments. *Progress in Statistics*, Ed. J. Gani, pp. 241–66. Amsterdam: North Holland.

Jones, P. W. (1975). The two-armed bandit. *Biometrika* **62**, 523–4.

Raiffa, H. & Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Cambridge, Mass: Harvard University Press.

Wahrenberger, D. L., Antle, C. E. & Klimko, L. A. (1977). Bayesian rules for the two-armed bandit problem. *Biometrika* **64**, 172–4.

[*Received October* 1978. *Revised June* 1979]