Adv. Appl. Prob. **15,** 221–222 (1983)
Printed in N. Ireland

SOME REWARD–PENALTY RULES FOR THE MULTI-ARMED BANDIT PROBLEM WHICH ARE ASYMPTOTICALLY OPTIMAL

K. D. GLAZEBROOK,* *University of Newcastle upon Tyne*

Abstract

In the mathematical learning literature, reward–penalty rules have been studied in various decision-theoretic and game-theoretic contexts, the multi-armed bandit problem included. Here we propose an elaboration of Bather's randomised allocation indices which yields rules for the multi-armed bandit which are both reward–penalty and asymptotically optimal.

GITTINS INDEX: MATHEMATICAL LEARNING

This note concerns rules for sampling one at a time from $k(\geqq 2)$ Bernoulli populations, population i having unknown probability of success p_i, $1 \leqq i \leqq k$. Our concern is with rules of the reward–penalty type. The central idea of such rules may be stated as follows: Do not decrease (do not increase) the probability of sampling the ith population at time $t+1$ if it was sampled at time t and the outcome was a success (failure). Plainly the 'play the winner' rule introduced by Robbins for the case $k=2$ is reward–penalty. In this rule if a success is observed with p_i then the same p_i is used in the next trial; otherwise we switch to the other one. Reward–penalty rules have been studied in various contexts (this one included) in the mathematical learning literature—see, for example, Meybodi and Lackshmivarahan (1982).

In the class of all reward–penalty rules we seek those which are asymptotically optimal, i.e. which will guarantee that the observed proportion of successes converges to $\max p_i$ when the total number of trials becomes infinite. From Bather (1981) we know that for a special version of the problem with $k=2$ no deterministic stationary rule can be asymptotically optimal and that (excepting the case $p_1 = p_2$) the play-the-winner rule is not asymptotically optimal.

Bather (1981) proposed a class of asymptotically optimal rules based on randomised allocation indices. Although these rules are not in general reward–penalty they can be elaborated in such a way as to make them so while preserving their asymptotic optimality. The new class of rules thus obtained samples one at a time from $k(\geqq 2)$ Bernoulli populations as follows: on the $(t+1)$th occasion sample from population j if and only if j is the smallest integer such that $Q_j(t) = \max_i Q_i(t)$, where

$$Q_i(t) = \eta_i\{s_i(t), f_i(t)\} + \lambda_i\{s_i(t), f_i(t)\}X_i(t) \tag{1}$$

and where for each i

Received 7 October 1982; revision received 21 January 1983.

* Postal address: Department of Statistics, The University, Newcastle upon Tyne, NE1 7RU, U.K.

(*a*) $s_i(t)$ and $f_i(t)$ are, respectively, the number of successes and failures observed in population i on all occasions up to and including the tth. We write $n_i(t) = s_i(t) + f_i(t)$ to denote the number of occasions up to and including the tth on which population i was sampled.

(b) η_i and λ_i are bounded real-valued functions on $\mathbb{Z}^+ \times \mathbb{Z}^+$, both non-decreasing in the first argument and non-increasing in the second. λ_i is also assumed to be positive. We further require that

$$P[\eta_i\{s_i(t), f_i(t)\} \to p_i \quad \text{as} \quad t \to \infty \mid n_i(t) \to \infty \quad \text{as} \quad t \to \infty] = 1 \tag{2}$$

and

$$P[\lambda_i\{s_i(t), f_i(t)\} \to 0 \quad \text{as} \quad t \to \infty \mid n_i(t) \to \infty \quad \text{as} \quad t \to \infty] = 1.$$

Two obvious choices for η_i are, firstly,

$$\eta_i(s_i, f_i) = (s_i + 1)(s_i + f_i + 1)^{-1}$$

and, secondly, a Gittins index for population i (with suitably chosen multiplicative constant to ensure that (2) is achieved). An example of a possible choice for λ_i is

$$\lambda_i(s_i, f_i) = e^{-f_i}(1 - e^{-s_i}).$$

(c) $X_i(t)$, $1 \leqq i \leqq k$, $t \in \mathbb{Z}^+$, are independent and identically distributed positive random variables with absolutely continuous density function.

It is not difficult to show that any rule which samples according to the indices in (1) is (under the conditions in (b) and (c)) both reward–penalty and asymptotically optimal. That the rules are reward–penalty may be established by direct calculation. The proof of asymptotic optimality is a straightforward extension of that given by Bather (1980).

We come to some particularly interesting conclusions if we sample according to (1), choosing the η_i's to be the (suitably modified form of the) Gittins indices. All Gittins indices are defined with respect to a particular Bayesian formulation of the multi-armed bandit problem with discounted rewards. It is not difficult to show that by insisting that

$$\max_{1 \leqq i \leqq k} \sup_{(s_i, f_i)} \{\lambda_i(s_i, f_i)\}$$

be small enough we can in this way obtain rules which are reward–penalty, asymptotically optimal and ε-Bayes (with respect to the same Bayesian formulation with discounted rewards)—a formidable array of properties. This extends the results of Glazebrook (1980).

References

BATHER, J. (1980) Randomised allocation of treatments in sequential trials. *Adv. Appl. Prob.* **12,** 174–182.

BATHER, J. (1981) Randomised allocation of treatments in sequential experiments (with discussion). *J. R. Statist. Soc.* B**43,** 265–292.

GLAZEBROOK, K. D. (1980) On randomized dynamic allocation indices for the sequential design of experiments. *J. R. Statist. Soc.* B**42,** 342–346.

MEYBODI, M. R. AND LACKSHMIVARAHAN, S. (1982) ε-optimality of a general class of learning algorithms. In *Proc. Conf. Mathematical Learning Models—Theory and Applications.* To appear.