Bandit Problems with Random Discounting — Source link

Donald A. Berry

Institutions: University of Minnesota

Published on: 01 Jan 1983

Topics: Discounting

Bandit Problems with Random Discounting

by

Donald A. Berry*

University of Minnesota

Technical Report No. 400

May 1982

*This work was supported by NSF/MCS 8102477.

Bandit Problems with Random Discounting

by Donald A. Berry
University of Minnesota

ABSTRACT

One of k independent stochastic processes with unknown characteristics is observed at each of a possibly infinite number of stages. Future stages are discounted: the m^{th} observation is weighted by α_m. The α_m are random variables. They may be dependent and their distributions unknown; in such a case one can learn about the character of the discounting as well as about the processes. The objective is to maximize the expected sum of the weighted observations. The decision problem is shown to be equivalent to one with nonrandom discounting in some versions. Other versions are intrinsically more complicated than the nonrandom case. Examples are carried out.

Bandit Problems with Random Discounting*

by Donald A. Berry**

1. Introduction.

One of k independent stochastic processes is observed at each of a possibly infinite number of stages. Selecting a process (or arm) to observe is called a pull. The arm pulled at any stage can depend on the pulls and resulting observations at all previous stages.

A strategy is a function that, for each finite history of pulls and observations, assigns an arm to be pulled next. To stress dependence on the strategy, τ_m will denote the observation at stage m when following strategy τ. If τ specifies arm j at stage m then $\tau_m = X_{jm}$. (For notational convenience it is assumed that all k processes are ongoing though only one can be observed at a time.)

Assume for fixed j that the X_{jm}, $m = 1, 2, \ldots,$ are identically distributed and independent given a common parameter θ_j. At least one of the θ_j is unknown, for otherwise the problem would be trivial. The parameters are themselves random variables with given "prior" probability distributions. So if θ_j is unknown, variables X_{jm}, $m = 1, 2, \ldots,$ are exchangeable rather than independent -- learning is possible. The information available about arm j at any time is contained in the current probability distribution on θ_j.

Such decision problems are sometimes called "bandits" in analogy with choosing whether or not to play a slot machine -- colloquially called a "one-armed bandit." Most of the bandit literature treats one

* Paper presented at the conference "Mathematical Learning Models -- Theory and Algorithms," Bad Honnef, W. Germany, May 3-7, 1982.
**Research supported by the National Science Foundation under Grant No. MCS81-02477.

of two objectives:

(i) Finite horizon: for some fixed n, the expected sum of the first n observations is to be maximized.

(ii) Geometric discounting: the m^{th} observation is weighed by a factor α^m, $0 < \alpha < 1$, and the expected weighted sum over the infinite horizon is to be maximized.

Historically important papers concerning these objectives are, respectively, (Bradt, Johnson and Karlin 1956) and (Bellman 1956) -- both papers deal with Bernoulli processes. Very recent papers by participants in this conference, again respectively, are (Bather 1981) and (Gittins 1979).

A general discounting approach, which includes objectives (i) and (ii), is taken in (Berry and Fristedt 1979) -- referred to henceforth as BF79. The m^{th} observation is weighed by a factor α_m and the expected weighted sum over the infinite horizon is to be maximized. So a strategy is optimal if it maximizes expected payoff:

(1.1) $$W(\tau) = E \, \Sigma_{m=1}^{\infty} \alpha_m \tau_m .$$

When the discount factors α_m are known constants, (1.1) becomes

(1.2) $$W(\tau) = \Sigma_{m=1}^{\infty} \alpha_m E\tau_m .$$

Assume $\alpha_m \geq 0$ for all m and $\Sigma_1^{\infty}\alpha_m < \infty$; $A = (\alpha_1, \alpha_2, \ldots)$ is called a discount sequence. That (ii) is a special case is obvious; for (i) take $\alpha_1 = \ldots = \alpha_n = 1$ and $\alpha_{n+1} = \ldots = 0$.

Because the language is so appealing, the arm specified at the first stage by an optimal strategy is called an "optimal arm."

An easy example may underscore some critical issues.

Example 1.1. Suppose $A = (1,1,0,\ldots)$; that is, (i) applies with $n = 2$. Each $\{X_{jm}: m = 1,2,\ldots\}$ is a Bernoulli process with $\theta_j = P(X_{jm} = 1)$; assume the k processes are independent. There are k^3 essentially different strategies. This number can be reduced to k^2 by applying the stay-on-a-winner rule (Berry 1972): If an optimal arm is pulled at any stage and yields a success, then it is optimal at the next stage as well. Label the arms so that $E\theta_1 \geq \ldots \geq E\theta_k$. We need only consider strategies that use arm 1 after a failure on the first pull of any arm other than 1 . For, by Cauchy-Schwarz,

$$P(X_{j2} = 1 | X_{j1} = 0) = \frac{E\theta_j - E\theta_j^2}{1 - E\theta_j}$$

$$\leq \frac{E\theta_j (1 - E\theta_j)}{1 - E\theta_j}$$

$$= E\theta_j \leq E\theta_1 .$$

There are two possibilities -- arm 1 and arm 2 -- when arm 1 is used initially and fails.

There are $k + 1$ strategies to consider: $\tau^0, \tau^1, \ldots, \tau^k$. In an evident notation, and using independence,

$$W(\tau^0) = 2E\theta_1 ,$$

$$W(\tau^1) = E\theta_1 + E\theta_1^2 + (1 - E\theta_1)E\theta_2 ,$$

$$W(\tau^j) = E\theta_j + E\theta_j^2 + (1 - E\theta_j)E\theta_1 ,$$

for $j = 2,\ldots,k$. And τ^j is optimal if its expected payoff is greatest.

To illustrate, if the θ_j all have uniform densities on $(0,1)$ then $W(\tau^0) = 1$ and $W(\tau^1) = \ldots = W(\tau^k) = 13/12$. □

The case $k = 2$ is considered in BF79; the characteristics of one arm, say arm 1 for definiteness, are unknown and those of arm 2 are known. So the information concerning arm 1 changes as it is pulled, but that of arm 2 does not. It is well-known in this case for both discount sequences (i) and (ii) that there exists an optimal strategy with the following characteristic: once arm 2 is selected it is thenceforth used exclusively and indefinitely. Such problems are stopping problems: one need only decide when to stop experimenting with arm 1. BF79 shows there are always optimal strategies with this characteristic if the discount sequence (assumed to be monotonic) is <u>regular</u>. Conversely, if it is not regular then there is a distribution on θ_1 for which no optimal strategies have this characteristic (cf. Example 1.2).

<u>Definition 1.1</u>. A discount sequence $A = (\alpha_1, \alpha_2, \ldots)$ is <u>regular</u> if, for each m ,

$$\gamma_m \gamma_{m+2} \leq \gamma_{m+1}^2$$

where $\gamma_r = \Sigma_{i=r}^{\infty} \alpha_i$.

The following are examples.

Regular:

(iii) $(1,\ldots,1,\alpha,\alpha^2,\ldots)$, $0 < \alpha < 1$

(iv) $(4,4,3,3,2,2,1,1,0,\ldots)$

(v) $(2,1,1,0,\ldots)$

Not regular:

(vi) $(2,1,1,1,0,\ldots)$

(vii) $(4,1,1,0,\ldots)$

(viii) $(1/2,5/16,\ldots,(1/2)(3/4)^m + (1/2)(1/4)^m,\ldots)$

That sequence (v) is regular follows from the regularity of (iv); it is listed for easy comparison with (vi).

Sequence (viii) is the average of two geometrics, which, of course, are themselves regular. But geometrics are barely regular: $\gamma_{m+1}^2 = \gamma_m\gamma_{m+2}$ for all m . So the slightest tampering destroys regularity. In particular, means of nondegenerate mixtures of geometrics are never regular, as the following calculation shows. Consider the sequence $(EV,EV^2,EV^3,\ldots)$ where V is a random variable on $[0,1]$. Then, for $m = 1,2,\ldots,$

$$\gamma_m = E(\frac{V^m}{1-V}) .$$

We have

$$\gamma_2{}^2 - \gamma_1\gamma_3 = E^2(\frac{V^2}{1-V}) - E(\frac{V^2}{1-V} + V)E(\frac{V^2}{1-V} - V)$$

$$= E(\frac{VEV^2 - V^2EV}{1-V}) .$$

The function $(xEV^2 - x^2EV)/(1-x)$ is concave in x on $[0,1]$ -- strictly concave unless $V=0$ or $V=1$ with probability one. Therefore, Jensen's inequality applies to show that

$$\gamma_2^2 - \gamma_1\gamma_3 \leq 0$$

with strict inequality provided V is not concentrated at one point.

Example 1.2. Suppose $k=2$. As in Example 1.1, the processes are Bernoulli with, for $j=1,2$, $\theta_j = P(X_{jm} = 1)$. Suppose θ_2 is known and θ_1 is either 0 or 1 with probabilities 1/2 each. This assumption makes the problem relatively easy because a single observation on arm 1 reveals θ_1. If the discount sequence is regular then the problem is trivial because only two strategies need be considered. Namely, τ': pull arm 1, if $\tau_1' = 1$ (success) pull arm 1 forever and if $\tau_1' = 0$ (failure) pull arm 2 forever; and τ'' : pull arm 2 forever.

Consider discount sequence (viii). Since it is not regular we must allow for switches to arm 1 from arm 2. The optimal strategy depends on θ_2; a complete list is given in the Table 1. The notation "2221," for example, means arm 2 is pulled at the first three stages and arm 1 at the fourth stage -- naturally, arm 1 is continued if it is successful and dropped otherwise.

TABLE 1

Interval for θ_2 (rounded to four decimals)	Optimal Strategy
(0, 0.7273)	1 (or τ')
(0.7273, 0.7692)	21
(0.7692, 0.7887)	221
(0.7887, 0.7961)	2221
(0.7961, 0.7987)	22221
(0.7987, 0.7996)	222221
(0.7996, 0.7999)	2222221
(0.7999, 1)	222... (or τ'')

Even though the structure is otherwise simple, the fact that the discount sequence is not regular makes the solution complicated. □

The possibility that the discount factors are unknown is introduced in the next section. Allowing for randomness in the discount sequence is natural enough, but it seems not to be considered in the literature -- not in the bandit literature anyway. Two versions are considered depending on whether the discount factors are observable. When they are not, or when they must be ignored, the problem is shown to be equivalent to one with nonrandom discounting. When they are, it sometimes reduces to a nonrandom problem and sometimes does not.

2. Preliminaries.

Suppose the discount sequence is not completely known. In economics, for example, the inflation rates in future years would not be known. In a medical trial the size of the patient pool may itself be random. Or, a new arm may be discovered -- one that is obviously better than the arms in the trial. This would likely end the trial prematurely; the discount factors become 0 from some stage on, and that stage is random.

One way to allow a discount sequence to be random is to place a measure on the space of nonrandom sequences. A random discount sequence is the corresponding mixture of nonrandom ones. However, specifying a measure with a large support is difficult. The bulk of this article takes a narrower approach, but one that is natural and seems easy to apply. Mixtures will be discussed again in Section 5.

Let $U_1, U_2, \ldots$ be nonnegative random variables. Set $\alpha_1 = U_1$ and for $m = 2,3,\ldots,$ recursively define

$$\alpha_m = \alpha_{m-1} U_m .$$

The distribution measures of $U_1, U_2, \ldots,$ call them $F_1, F_2, \ldots,$ may themselves be unknown. Given $F_1, F_2, \ldots,$ variables $U_1, U_2, \ldots$ are assumed to be independent. However, if the F_i are dependent random distributions then the U_i are not generally independent.

It will be assumed throughout that the U_i are independent of the X_{jm} .

There is now some ambiguity in the use of the term "strategy." This will be resolved momentarily. In any case definition (1.1) of expected payoff of a strategy τ continues to apply with

$$\alpha_m = \Pi_1^m U_i \ .$$

The expectation in (1.1) is now with respect to the distribution of the U_i as well as that of the τ_m .

We shall consider two sets of ground rules:

Version 1. The random variables U_i are not observable. So while the τ_m are observed, the discounted payoff at stage m , $\alpha_m \tau_m$, is not. The set of available strategies in this version, call it T_1 , is as defined in Section 1 for the nonrandom case.

Version 2. The random variables U_i are observable. The decision at stage $m+1$ can depend on $(U_1,\ldots,U_m)$ as well as on τ and $(\tau_1,\ldots,\tau_m)$. Let T_2 denote the corresponding set of available strategies.

A third possibility -- one not considered here -- is that the product $\alpha_m \tau_m$ is observed at stage m , but not α_m and τ_m individually.

Version 2 seems more realistic than Version 1. But one can imagine circumstances in which a strategy can be programmed to depend only on the results of the pulls. Strategies in T_1 are simpler than typical strategies in T_2 . Actually, each $\tau \in T_1$ has a version in T_2: there is a strategy in T_2 which duplicates the decisions specified by any $\tau \in T_1$. Therefore, Version 1 provides a bound for Version 2: The maximal expected payoff in Version 2 is no smaller than in Version 1. Typically, it is greater. But, as will be seen, there are numerous circumstances in which they are equal, when the ability to observe the U_i provides no advantage.

3. Version 1: Nonobservable Discount Factors.

For all $\tau \in T_1$, $(\tau_1, \tau_2, \ldots)$ is independent of $(U_1, U_2, \ldots)$. Therefore, (1.1) becomes

(3.1) $$W(\tau) = \Sigma_1^\infty E\alpha_m E\tau_m$$

for all $\tau \in T_1$, where

(3.2) $$E\alpha_m = E\Pi_1^m U_i \ .$$

So (1.2) applies with α_m replaced by $E\alpha_m$. And the problem considered here is no more general than that considered in BF79 (except that the number of arms is now arbitrary and the possibility $E\alpha_{m+1} > E\alpha_m$ is not ruled out).

In the special case in which the U_i are independent, (3.2) becomes

(3.3) $$E\alpha_m = \Pi_{i=1}^m EU_i \ .$$

Example 3.1. Suppose the U_i are independent with

$$EU_1 = \ldots = EU_n = 1, \ EU_{n+1} = \ldots = 0$$

(F_{n+1} concentrates its mass at 0 and the F_i for $i > n+1$ are immaterial). Then the discount sequence relevant for choosing a strategy is (i), finite horizon: $EA = (1,1,\ldots,1,0,\ldots)$. This is not to say the choice is easy. But backward induction is available for finding optimal strategies just as in the usual, nonrandom finite horizon setting. □

Example 3.2. Suppose the U_i are independent with $EU_i = \alpha$ for $i = 1,2,\ldots$; α is known and $0 < \alpha < 1$. It may be, for example, that the trial terminates at stage m with conditional probability $1 - U_m$. Then $E\alpha_m = \alpha^m$ and the problem is the same as (ii), geometric discounting. In particular, the results of (Gittins 1979) apply. □

The nonrandom discount sequences in the previous two examples are regular. The resultant sequence in the next example is not regular. It will be referred to again in Example 4.1.

Example 3.3. Discount sequence (viii) considered in Example 1.2 is $(1/2,5/16,7/32,\ldots)$. This can arise as the mean of $A = \{\alpha_m\}$ in a number of ways. For example, the U_i may be independent (so (3.3) applies) with

$$EU_i = \frac{1}{4}\,\frac{3^i + 1}{3^{i-1} + 1}$$

for $i = 2,3,\ldots$. Or, $P(F_1 = F_2 = \ldots = F) = 1$ where F is an equal mixture of two one-point distributions; one at $3/4$ and one at $1/4$. In the latter interpretation $P(U_1 = U_2 = \ldots = 3/4) = P(U_1 = U_2 = \ldots = 1/4) = 1/2$. This is consistent with viewing A as the average of two geometrics. Regardless of how the sequence arises, an optimal strategy is as given in a nonrandom setting with discount sequence $EA = (1/2,5/16,7/32,\ldots)$; for a special case see Example 1.2. □

Example 3.4. Suppose $F_1 = F_2 = \dots F$ where $F(\{1\}) = q = 1 - F(\{0\})$; q is unknown and has a uniform distribution on $(0,1)$. This seems to be a harmless assumption. However,

$$E\alpha_m = P(U_1 = \dots = U_m = 1)$$

$$= \int_0^1 q^m dq = \frac{1}{m + 1} .$$

So $\Sigma E\alpha_m = \infty$ and EA is not a discount sequence. (If $\Sigma E\alpha_m = \infty$ were allowed then EA would not be regular. For such a sequence one would ignore immediate gain and sample only to obtain information that might help in the long run. Optimal strategies would be similar to Kelly's (1981) "least-failures rule.") □

4. Version 2: Observable Discount Factors.

Strategies in T_1 do not depend on the U_i. Strategies in T_2 depend on the U_i as well as the observed X_{jm}. This section treats the latter possibility.

There is an important distinction in Version 2 between independent and dependent U_i. These cases are considered separately.

4.1 Independent U_i.

Suppose for $i = 1,2,\dots$ that F_i is a random distribution with measure μ_i on the space of distributions. F_i is known if μ_i is a one-point measure. For the purposes of this section assume the F_i are independent. Then so are the U_i. In making a decision at stage $m+1$,

$U_1,\ldots,U_m$ are known. Since the U_i are independent the conditional distribution of U_{m+1} given $U_1,\ldots,U_m$ is the same as the unconditional. Therefore, (3.1) and (3.3) apply. The mean of U_i can be expressed as

$$EU_i = \int E(U_i | F_i)\mu_i(dF_i) .$$

The above argument is complete but brief. The following discussion may be helpful. The initial selection depends on later possibilities. Consider stage $j+1$ assuming $U_1 = u_1 , \ldots , U_j = u_j$. The current decision problem is to maximize

$$(4.1) \qquad W_{j+1}(\tau) = \Sigma_{m=j+1}^{\infty}(\Pi_{i=1}^{j} u_i)(\Pi_{i=j+1}^{m} EU_i)E\tau_m .$$

But two problems with proportional discount sequences are equivalent -- (4.1) can be written

$$W_{j+1}(\tau) = K\Sigma_{m=j+1}^{\infty}(\Pi_{i=1}^{m} EU_i)E\tau_m ,$$

where

$$K = \Pi_{i=1}^{j}(u_i/EU_i) .$$

Therefore an optimal selection at stage $j+1$ can be made without observing the U_i; equivalently, each U_i can be assumed equal to its mean.

So when the F_i are independent the problem is the same whether or not the discount factors are observable. And in turn both random discounting versions are equivalent to nonrandom discounting.

Moreover, the expected payoff of any strategy is the same in all three cases. Of course, the expected payoff of the continuation of a strategy changes depending on the u_i .

Examples 3.1, 3.2 and 3.3 apply also for the case considered here. Take Example 3.1. The mean of the discount sequence relevant at stage 2, given $U_1 = u_1$, is u_1 times the (n-1)-horizon: (1,1,...,1,0,...) . Each new stage gives a problem identical with the corresponding one in Example 3.1.

4.2. Dependent U_i.

Some additional notation is helpful for this case. The ideas apply generally but for convenience the development is restricted to the Bernoulli case: every pull results in a 0 or a 1 . The $j^{\underline{th}}$ arm gives 1 with probability θ_j .

The (initial) random discount sequence is

$$A = (U_1, U_1U_2, U_1U_2U_3, \ldots) .$$

At stage 2, after observing U_1 , the relevant discount sequence is

$$\begin{aligned}(A^{(1)}|U_1) &= (U_1U_2|U_1, U_1U_2U_3|U_1, \ldots) \\ &= U_1(U_2|U_1, U_2U_3|U_1, \ldots) ;\end{aligned}$$

this and subsequent notation is consistent with BF79.

Let G denote the initial joint distribution of $(\theta_1, \ldots, \theta_k)$. If arm j is pulled and results in success, $X_{j1} = 1$, then G is changed via Bayes theorem to $\sigma_j G$, say. Similarly, a failure on arm j changes G to $\varphi_j G$.

Let V_j denote the expected payoff of pulling arm j initially and then following an optimal strategy (in T_2) . Define

$$V = \max\{V_1,\ldots,V_k\} .$$

The relevant standard functional equations are

$$(4.2) \qquad V_j(A,G) = E\theta_1 EU_1 + E\theta_1 E\big[U_1 V((A^{(1)}|U_1),\sigma_j G)\big] \\ + (1-E\theta_1)E\big[U_1 V((A^{(1)}|U_1),\varphi_j G)\big] ,$$

for $j = 1,\ldots,k$. The problem can be solved, or at least the solution approximated, by repeated application of (4.2). But the calculations can be forbidding. In particular, the posterior distribution of $(U_{m+1},U_{m+2},\ldots)$ given $U_1,\ldots,U_m$ can be arbitrarily difficult unless a simple structure is imposed.

To make the calculations manageable, assume the unknown F_i have a special kind of dependence: for all i , $F_i = F$ which is a random distribution with measure μ . When a discount factor α_m -- and therefore U_m -- is observed, the current measure of F is updated. Updating is easiest if F is known up to some real-valued parameter η . For then Bayes theorem applies to modify a prior distribution on η .

A useful alternate approach due to Ferguson (1973) is to give F a "Dirichlet process prior." For each real u , $F(u)$ has a beta distribution with parameters $MF_0(u)$ and $M(1 - F_0(u))$; F_0 is the prior mean of F and M is a measure of prior precision. After observing $U_1 = u_1,\ldots,U_m = u_m$, the posterior of F is also a

Dirichlet process. The new M is $M+m$ and MF_0 becomes $MF_0 + \Sigma_1^m I_{u_i}$; here, $I_x(u) = 1$ if $u \leq x$ and 0 otherwise. This approach has promise for two reasons: (1) As is clear from the above comments, calculations are manageable. (2) The support of a Dirichlet process (in the topology of pointwise convergence) contains all probability measures absolutely continuous with respect to F_0 (Ferguson 1973).

Neither of the above-mentioned possibilities for updating the distribution of F are carried forward in the present paper. (I plan more work on this problem.) Instead, an example is given in which updating is quite simple.

Example 4.1. Consider the setting of Example 1.2: there are two Bernoulli arms, θ_2 is known, and θ_1 is either 0 or 1 , with equal probabilities under G . Distribution F is unknown; it is one of two one-point distributions with equal probabilities, one point is $3/4$ and the other is $1/4$. Therefore the U_i are either all $3/4$ or all $1/4$; which one will be revealed at the first stage.

In Version 1 (see Example 3.3) the relevant discount sequence, $EA = (1/2, 5/16, 7/32, \ldots)$, is not regular. When F is unknown regularity of EA is not a consideration. However, F becomes known after stage 1. And, for $u = 3/4$ or $u = 1/4$,

$$(A^{(1)} | U_1 = u) = u(u, u^2, u^3, \ldots) ,$$

with probability one. Since both these sequences are geometric, and therefore regular, the number of strategies in T_2 that must be considered is sharply reduced.

A further reduction is possible. Example 4.4 of BF79 shows that the "break-even value" of θ_2 when $U_1 = 3/4$ is $\theta_2 = 4/5$; when $U_1 = 1/4$ it is $4/7$. We need consider only three strategies -- τ': pull arm 1, pulling it indefinitely if it is successful and switching to arm 2 (permanently) otherwise; τ'': pull arm 2 indefinitely; τ''': pull arm 2, then follow τ' if $U_1 = 3/4$ and τ'' if $U_1 = 1/4$. Easy calculations show:

$$W(\tau') = \frac{7}{12}\theta_2 + \frac{5}{6}$$

$$W(\tau'') = \frac{5}{3}\theta_2$$

$$W(\tau''') = \frac{185}{192}\theta_2 + \frac{9}{16}$$

So $V_1(A,G) = W(\tau')$ and $V_2(A,G) = \max\{W(\tau''),W(\tau''')\}$ and $V(A,G) = \max\{W(\tau'), W(\tau''), W(\tau''')\}$. All optimal strategies are given as follows: τ' for $\theta_2 \leq 52/73 \doteq 0.7123$, τ''' for $52/73 \leq \theta_2 \leq 4/5$, and τ'' for $\theta_2 \geq 4/5$.

This solution should be compared with Table 1. The interested reader can check that

$$\sup_{\tau\in T_1} W(\tau) \leq \sup_{\tau\in T_2} W(\tau)$$

with strict inequality if and only if $52/73 < \theta_2 < 4/5$.

In this example, not only is Version 2 an improvement over Version 1, but the analysis is simpler. □

5. Mixtures.

As indicated in Section 2, a more general way of introducing random discount sequences is to mix nonrandom sequences. In Version 1, nonobservable discount factors, the problem reduces to one with a nonrandom discount sequence. The reasons given in Section 3 also apply for mixtures. The corresponding nonrandom sequence is simply the mean of the random sequence.

Consider Version 2, observable discount factors. After stage m the mixing distribution is updated via Bayes theorem in a very simple way. Suppose $\alpha_1',\ldots,\alpha_m'$ are known to be the first m discount factors. The total posterior probability of those sequences which disagree with $(\alpha_1',\ldots,\alpha_m')$ in at least one of the first m positions is 0. And the posterior measure of those not ruled out is proportional to the initial measure.

For example, suppose all the sequences in the support of the initial distribution have distinct first factors. Then the true discount sequence will be revealed at stage 1. Learning takes place quickly, but this brings out a difficulty in applying the mixture approach. If one has not been sufficiently careful assigning the initial distribution then every discount sequence may soon be ruled out! And it is difficult to assign a measure rich enough to avoid this problem. In the approach of previous sections, one worries about randomness in a discount sequence on a day-to-day, or stage-to-stage, basis. With mixtures one continually worries about an eternity of randomness.

<u>Example 5.1</u>. Suppose every sequence in the support of the initial measure is of the form (i), finite horizon: $(1,1,\ldots 1,0,\ldots)$, differing only in the length of the horizon. In this rather special

circumstance, observations of the discount factors can be ignored: Version 2 = Version 1. For, the decision maker can always act as though the discount factor "1" was just observed; if it really was a "0" then the remaining actions are of no consequence.

Every nonrandom discount sequence can be expressed as the mean of a mixture of finite horizons. Suppose, for example, the initial probability of $(0,0,\ldots)$ is $1-\alpha$, where α is known and $0 < \alpha < 1$, of $(1,0,\ldots)$ is $(1-\alpha)\alpha$, of $(1,1,0,\ldots)$ is $(1-\alpha)\alpha^2$, etc. Then the mean of this mixture is the geometric sequence, (ii): $(\alpha,\alpha^2,\ldots)$. So in this setting, optimal strategies in Version 1 are also optimal in Version 2. Moreover, they can be found from the nonrandom geometric discounting case. □

6. Conclusion.

When discount factors ΠU_i are random but cannot be observed, the problem is identical with a particular nonrandom problem.

When such discount factors can be observed and the U_i are independent random variables, then again the problem reduces to one that is nonrandom. But this is not the case when the U_i are dependent and learning about the future U_i is possible. The set of available strategies is larger in this version. However, the task of finding an optimal strategy can be easier.

ACKNOWLEDGEMENT. I want to thank Bert Fristedt, John Bather, Alfonso Novales, and David Polansky for helpful discussions.

References

Bather, J. A. (1981). Randomized allocations of treatments in sequential experiments (with discussion). J.R. Statist. Soc. B 43:265-292.

Bellman, R. (1956). A problem in the sequential design of experiments. Sankhya A 16:221-229.

Berry, D. A. (1972). A Bernoulli two-armed bandit. Ann. Math. Statist. 43:871-897.

Berry, D. A., and Fristedt, B. E. (1979). (Called BF79 in text.) Bernoulli one-armed bandits -- Arbitrary discount sequences. Ann. Statist. 7:1086-1105.

Bradt, R. N., Johnson, S. M., and Karlin, S. (1956). On sequential designs for maximizing the sum of n observations. Ann. Math. Statist. 27:1060-1070.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. Ann. Statist. 1:209-230.

Gittins, J. C. (1979). Bandit processes and dynamic allocation indices (with discussion). J. Roy. Statist. Soc. B 41:148-177.

Kelly, F. P. (1981). Multi-armed bandits with discount factor near one: the Bernoulli case. Ann. Statist. 9: 987-1001.