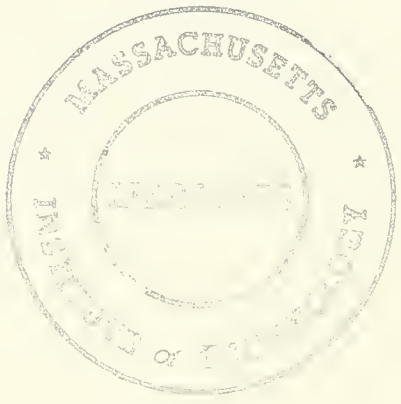

MASSACHUSETTS
INSTITUTE OF TECHNOLOGY

working paper
department
of economics

ON A GENERAL APPROACH TO SEARCH AND INFORMATION GATHERING

Kevin W.S. Roberts
Martin L. Weitzman

Number 263 August, 1980

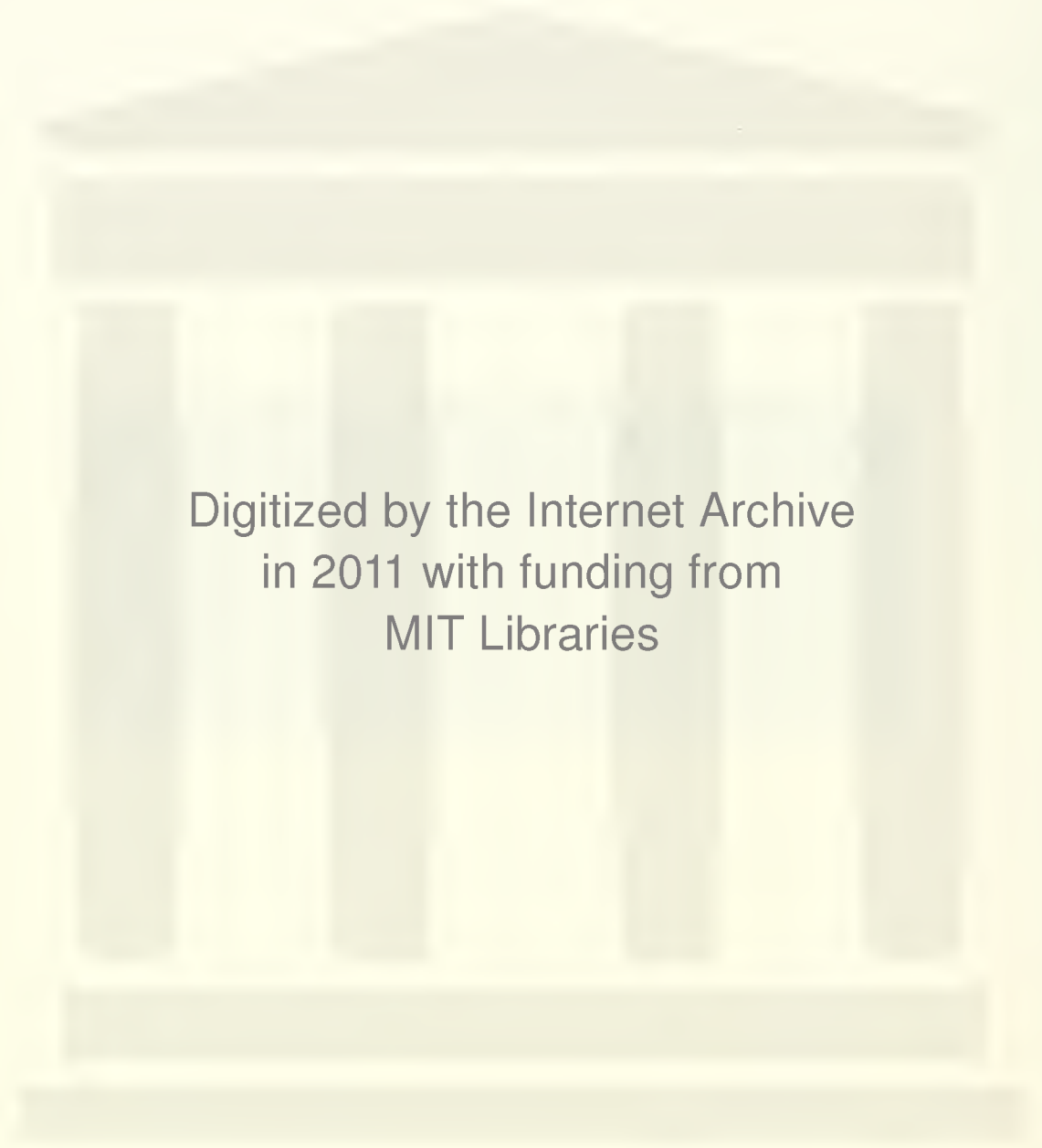
massachusetts
institute of
technology

50 memorial drive
cambridge, mass. 02139

ON A GENERAL APPROACH TO SEARCH AND INFORMATION GATHERING

Kevin W.S. Roberts
Martin L. Weitzman

Number 263 August, 1980



Summary

We show that an important class of multi-stage decision problems, of which conventional search theory is a special case, can be formulated and constructively solved within a unified framework. The optimal strategy is an elementary reservation price rule, allowing an intuitive economic interpretation and permitting problems to be solved in polynomial rather than exponential time. Computationally efficient algorithms are presented which can be used to numerically calculate reservation prices in real situations. We investigate the qualitative properties of an optimal policy, analyze how they depend on various underlying economic features of the problem, and note what they imply about optimal decisions in different contexts.

Introduction

A broad class of dynamic allocation models can be roughly described as follows. A decision maker has some number of activities, projects, or opportunities from among which he chooses one for further development. Depending upon the project selected some reward is received and the state of that project is possibly altered, while the other projects remain intact in their previous condition. The basic problem is to choose opportunities at each decision time to maximize expected present discounted value. While complete characterization will have to await the introduction of formal notation, suffice it here to note that there are many useful applications.

Although the basic philosophy of this paper argues that such models are best viewed as multi-stage generalizations of search problems, we adhere to an existing nomenclature in the statistics literature which has classified several examples as so-called "bandit processes" by analogy with the multi-armed bandit problem.[1]

This paper has several objectives. After presenting a discrete-state framework for formulating bandit processes, we show that a rather impressive variety of dynamic allocation problems from seemingly unrelated areas of economics, statistics, and operations research can be cast in that form. Then we prove the existence of a general, constructive method for solving bandit processes.

The solution, which by all reckoning should be complicated to state and very difficult to solve, can in fact be characterized by an elementary rule familiar to economists. Each stage of a project is assigned a reservation price -- a critical number analogous to an internal rate of return, depending only on the project and its stage, independent of all other projects, and possessing a simple, intuitive economic interpretation. The optimal rule is to proceed next with that activity, project, or opportunity having the highest reservation price.

[1] In previous versions of this paper, before the work of Gittens and others in the statistics literature was brought to our attention, we called our model a "multi-stage search process". We still prefer this description, but of course it would now be confusing to persist with our own terminology.

Since knowing reservation prices is tantamount to solving a bandit process, we can understand the basic qualitative features of an optimal policy by analyzing the main factors determining a project's reservation price -- like profitability, uncertainty, information, learning, flexibility, degree of increasing and decreasing returns, etc. In the context of various models, we try to explain why reservation prices differ between projects and how a reservation price is likely to change as a project is developed.

Not only can reservation prices be used to characterize the form of an optimal policy, but they allow any bandit process to be solved in polynomial rather than exponential time. We show that the theory of discrete state reservation prices is essentially constructive. Reservation prices satisfy dynamic programming equations of a form that allow them to be numerically calculated by iterative alogorithms with known and powerful convergence properties. This should make it computationally feasible to actually solve real world problems of a surprisingly large size.

The history of the problem is complicated to trace because a great many examples of bandit process have been effectively treated without real awareness of the underlying connection between them or of the existence of a unified theory.[2] To Gittens and some other statisticians belongs the credit for formulating the first truly

[2] Some of these examples will be presented in a later section.

general mathematical statement of the model, while primary emphasis has been on solving problems in statistical design akin to the multi-armed bandit problem.[3] It seems fair to say that only quite recently have people started to become aware of the full range of problems covered by this kind of theory.

To the economist, bandit processes are important because they form an elegant and operational theory which nicely captures the role of information gathering in dynamic resource allocation. Within economics this role has been played largely by search models, of which bandit processes are a powerful generalization.

The Model

There are N opportunities or projects, indexed n = 1,2,...,N. At any decision time exactly one project of the N must be selected for further development. Let project n be in state i. If project n is chosen, (expected) reward R_i^n is collected and project n makes a transition from state i to state j with probability P_{ij}^n. Every other project remains locked in its previous state. A discount factor β_i^n is applied to future returns.[4]

[3] See Gittens [1979], Whittle [1980], and the references cited therein. The essential results of the present paper were independently stated, proved, and written in an earlier version before we became aware of Gittens' pioneering work. So far as we know, the connection between search theory and bandit processes has not been treated anywhere in the literature, and it is primarily this aspect which is most important in economic applications. The search theory -- reservation price approach to bandit processes is also, in our opinion, the most intuitively appealing way to understand the dynamic programming conditions and to prove the basic theorem.

[4] A superscript on a variable indexes the project and does not mean raising the variable to that exponent.

Thus, if project n is selected, the system as a whole moves from state

$$S = \underset{m=1}{\overset{N}{X}} i(m) \tag{1}$$

to state

$$S - >i(n)< + >j< \tag{2}$$

with probability P_{ij}^{n}, where the notation $S - >i(n)< + >j<$ means a state identical to S except that project n is in state j instead of i(n).

The problem of selecting a project to maximize expected present discounted value can be posed in dynamic programming format. Let $\Psi(S)$ represent the expected present discounted value of following an optimal policy from this time on when the state of the system is S.

For each S, the state valuation functions Ψ must satisfy the fundamental recursive relation

$$\Psi(S) = \max_{1 \le n \le N} \{R_i^n + \beta_i^n \Sigma P_{ij}^n \Psi(S - >i(n)< + >j<)\} \tag{3}$$

In principle the state valuation functions $\{\Psi(S)\}$ might be recursively built up by iterative backwards induction on the stages of each project, using equation (3). In most actual cases the computation would be a brute force task of horrendous proportions because the "curse of dimensionality" is likely to be so strong.

At any state S, the optimal project to select, $n^*(S)$, is that alternative which maximizes the right hand side of (3). If two or more policies tie, it makes no difference how the tie is broken. Note that although an optimal strategy is implicitly contained in equation (3),

the form of that strategy is nothing more than a complete enumeration of what to do in all possible situations, with no visible economic or other interpretation.

In the solution of the standard search model, a reservation price equal to its certainty equivalent is assigned to each box. The reservation price of a closed box is that hypothetical cutoff value of a deterministic fallback reward which would make it just equal to the expected net gain of opening the box and having the certain reward to fall back upon.[5] In an optimal policy, that box with highest reservation price is opened next.

The contribution of the present paper is to show that essentially the same idea works for multi-stage search problems. A project in any stage is assigned a reservation price, calculated in an analogous manner to the standard search model. The reservation price of a project-stage determines its ordinal ranking, telling when to fund this project-stage relative to other project stages. Thus, all the advantages of a simple rate of return criterion apply in the context of search with accumulated information.

Examples

We exhibit some examples of bandit processes from three broad application areas.

[5]See Weitzman [1979], Lippman and McCall [1976].

(1) Search

Simple box search is the prototype problem of the present paper. It will become apparent why we have chosen this model as our primary conceptual antecedent when we examine the solution equations for the general bandit process.

Suppose there is a collection of closed boxes, not necessarily identical. Each box contains a potential reward sampled from a probability distribution. At some cost, and after an appropriately discounted waiting interval, a box can be opened and its contents made known.

At each decision node, the decision maker must decide whether or not to open a box. If search is terminated, the maximum reward thus far uncovered is collected. If search is continued, the decision make must select the next box to be opened, pay at that time the fee for opening it, and wait for the outcome. Then will come the next decision node. The object is to find a strategy which maximizes expected present discounted value.[6]

Suppose that box n contains a potential reward Y_i^n with probability q_i^n for i=1,2,... It costs C^n to open box n and learn its contents, which become known after a time lag reflected by the discount factor β^n.

6 See Weitzman [1979], Lippman and McCall [1976], and the references cited therein.

Simple box search can be posed as a bandit process using the following notation. In state o the box is closed and $P^n_{oj} = q^n_j$, $R^n_o = -C^n$, $\beta^n_o = \beta^n$. If the box is opened and in state $i \geq 1$, we say it is in an absorbing state with $P^n_{ii} = 1$, $R^n_i = Y^n_i$, $\beta^n_i = 0$.

Further specialized cases of simple box search, like locating an object, or the gold mining problem, are a fortiori examples of bandit processes.[7]

(2) Scheduling

The so-called "resource pool problem" is the simplest case of a non-trivial deterministic bandit process.[8] It highlights the pure scheduling or fitting aspect of a bandit process in its most direct form, free of search, learning, or other features.

At each instant of time a depletable resource can be drawn from any one of a number of pools. The cost of removing an extra unit from a pool depends on how much has already been taken out of it. What policy supplies a fixed flow of the resource at minimum present discounted cost?

If i units have been drawn from pool n, it costs C^n_i to extract the next unit. Then $P^n_{i,i+1} = 1$, $R^n_i = -C^n_i$, $\beta^n_i = \alpha$ defines a resource pool problem as a bandit process. The non-decreasing cost case $C^n_i \leq C^n_{i+1}$ is the familiar situation where a myopic marginalist rule is optimal.

[7] See Kadane and Simon [1977], Kadane [1969], and the reference cited therein.

[8] See Weitzman [1976].

In the more interesting case, the resource pool problem confronts the issue of evaluating situations with a range of decreasing costs.

Obviously uncertainty can be introduced into a resource pool problem without disturbing its status as a bandit process (although it will typically be somewhat harder to solve). Just as one example, suppose that pool n costs a fixed overhead charge of K^n to open up and has constant marginal cost C^n thereafter, but reserves are uncertain. A similarly structured problem is the task of scheduling a number of jobs to be carried out by a single machine when the jobs differ in value, and completion times are random variables.

Actually, with a judicious interpretation of "resource pool", a bandit process formulation is sufficiently general to include as special cases many of the standard operations research models for such problems as equipment durability selection and replacement, inventory, maintenance, and production scheduling, or capacity expansion. In such situations there are typically several classes of pools, each of which contains an infinite number of identical members. For equipment problems a "resource pool" is a certain piece of equipment and the "amount extracted" is the length of time it has been in service. With scheduling, inventory, and expansion problems, a pool is a certain strategy-schedule (like ordering inventory, or building capacity) which goes up to some expiration or regeneration point (after which costs are interpreted as

being infinite); the "amount extracted" is the length of time the given strategy-schedule has been carried out.

(3) Learning (or Information Gathering)

The following numerical example conveys the flavor of multi-stage learning as a generalization of search.

Suppose the research department of a large organization has been assigned the task of developing some new product. Two independent technologies are being considered, both of which are uncertain. Because they both produce the same product, no more than one technology would actually be used even if both were successfully developed.

For each technology, research goes through two stages. First, a preliminary feasibility study is made. If the outcome of the feasibility study is unfavorable, the technology has no chance of success. If the feasibility study is favorable, the technology may be successful, but this can only be determined after mounting a full scale R&D effort.

Table 1 summarizes the relevant information.

TABLE 1

Project	A	B
Probability of Success Estimated Before Feasibility Study	.5	.4
Cost of Feasibility Study	3	4
Probability of Passing Feasibility Study	.8	.5
Cost of Full Sclea R & D	21	24

For which technology -- A or B -- should a feasibility study be ordered? We discuss the answer in the applications section.

General multi-stage problems of this sort can be written as bandit processes. The corresponding bandit process is typically a unidirectional branching tree with terminal absorbing states; if $P_{ij} > 0$, then either $j > i$ or $j = i$ and $P_{ii} = 1$. In the previous example, the terminal absorbing states offer as reward either some large unspecified constant (success), or zero (failure). Rewards in a transition state are the negative of development costs at that stage.

In previous work, we have modeled a research, development or exploration project where the potential reward, which can only be collected after all development work has been completed, is viewed as a sum of independent random variables across component development stages.[9] As additional research money is paid to develop another stage, the "contribution" of that stage to the final reward becomes known. Hence, the distribution of final rewards is continually shifting as the contribution of each stage turns out better or worse than expected; and the distribution narrows with development because less uncertainty remains to be resolved. Research costs are paid both to move the project towards completion and to find out more information about potential rewards. If the decision maker has paid all development costs leading up to the final stage, development uncertainty has been eliminated and the

[9] See Roberts and Weitzman [1979]. The model is solved and analyzed by continuous time methods.

reward can be collected. At any stage the project can be abandoned and, viewed ex post, the previously sunk development costs have nothing to show. A collection of such projects is an example of a bandit process.

Perhaps the most classic model of learning is the multi-armed bandit problem which served Gittens and his co-workers as a prototype example.[10] At any stage the decision maker has an estimate of the distribution of success probabilities for each arm, which in our language is the state of the arm. When an arm is played, some reward is expected and depending on what is actually received, the estimate (or state) is updated in Bayesian fashion. The multi-armed bandit problem embodies a classical trade-off between taking high expected rewards now and acquiring information which may be valuable later.

It should be obvious that many economic aspects of "learning by doing" can be modeled as a bandit process.

The Basic Theorem

Consider the following functional equation:

$$V_i^n(Z) = \max \{ Z, R_i^n + \beta_i^n \Sigma P_{ij}^n V_j^n(Z) \} \qquad \forall i \qquad (4)$$

Under very weak conditions, (4) will have a unique solution for each Z.

10 See Gittens [1979], Whittle [1980], and the references cited therein. For applications to economics, see Rothschild [1974] and the references he cites.

$V_i^n(Z)$ possesses an important economic interpretation. Consider the artificial bandit process where one of the projects is n in state i and the other is a fallback lump sum reward Z which can be collected at any time, whereupon the entire process must be discontinued. $V_i^n(Z)$ is the expected present discounted value of being in state i of project n and following an optimal stopping rule when the consolation prize is Z. The difference $V_i^n(Z) - Z$ might be called the "option value" of being in i(n).

A fixed point of $V_i^n(Z)$ which will play an indispensable role is

$$Z_i^n,$$

defined to satisfy

$$Z_i^n = V_i^n(Z_i^n) = R_i^n + \beta_i^n \Sigma P_{ij}^n V_j^n(Z_i^n). \tag{5}$$

It is not hard to prove existence and uniqueness of Z_i^n. It is also not difficult to show that

$$V_i^n(Z) = Z \qquad \text{for } Z \geq Z_i^n$$

$$V_i^n(Z) > Z \qquad \text{for } Z < Z_i^n$$

Now Z_i^n has two important interpretations and one truly extraordinary property.

An interesting interpretation is that Z_i^n represents that value of the fallback position which would make a decision maker just indifferent between continuing project n at stage i and abandoning it in favor of taking the fallback reward immediately. In economist's terminology, Z_i^n is a reservation price.

An equally valuable way of comprehending Z_i^n is to note that it represents the expected present discounted value of an optimal policy for a bandit process consisting of an infinite number of projects, all of type n in state i. This interpretation follows from the fact that Z_i^n so defined must satisfy (5).

The extraordinary property is that Z_i^n contains <u>all</u> relevant information about project n in state i for <u>any</u> bandit process of which it is a member. The optimal rule is to proceed next with that project-state having the highest reservation price. This unusual feature can lead to striking results and powerful characterizations of an optimal policy. Note that Z_i^n does not at all represent the value of project n in any traditional economic sense; there is a crucial distinction between the marginal value of a project and the order in which it should be undertaken.

Adopting the notation

$$V_{ji}^n = V_j^n(Z_i^n), \tag{6}$$

a standard contraction mapping argument shows that the system of equations

$$V_{ii}^n = R_i^n + \beta_i^n \Sigma P_{ij}^n V_{ji}^n \tag{7}$$

$$V_{ji}^n = \max \{V_{ii}^n, R_j^n + \beta_j^n \Sigma P_{jk}^n V_{ki}^n\} \tag{8}$$

has a unique solution if $0 \leq \beta_j^n < 1$.

The following theorem is the basic result of this paper.

Theorem: The optimal policy in state $S = \sum_{n=1}^{N} i(n)$ is to select the project n^* for which

$$Z^{n^*} = \max_n Z^n, \tag{9}$$

where

$$Z^n \equiv V^n_{i(n),i(n)}.$$

Proof of the Basic Theorem

Henceforth we suppress the cumbersome notation i(n) where its use is superfluous. Unless otherwise noted, project n is in state i(n).

Throughout the proof it will be convenient to work with an equivalent undiscounted problem where $1-\beta_i^n$ is now interpreted as the probability that, if project n in state i is chosen, everything stops and the entire process ends (with zero reward). Transition probabilities are then

$$Q^n_{ij} = \beta_i^j \, p^n_{ij}.$$

The two formulations are mathematically identical, but the interpretation of the equivalent undiscounted problem is easier.

Let $G_i^n(Z)$ be the probability that Z^n will eventually fall to or below Z if project n is continued forever starting from state i. More formally,

$$G_i^n(Z) = 1 \qquad \text{if } Z_i^n \le Z$$

$$G_i^n(Z) = \Sigma Q^n_{ij} G_j^n(Z) \qquad \text{if } Z_i^n > Z$$

An interesting relation which we will not use directly is

$$\frac{dV_i^n(Z)}{dZ} = G_i^n(Z) \qquad \text{a.e.}$$

At any stage let A(Z) be a "continuation set" of projects whose reservation prices are greater than Z. More formally,

$$Z^n > Z \quad \leftrightarrow \quad n \in A.$$

Define the function

$$W(Z)$$

to be the maximum expected present discounted value of a bandit-like process played under the following conditions:

stopping rule: retire the process and collect Z if and only if A is empty.

selection restriction: the next project to be selected must belong to A.

It is not difficult to prove that W(Z) is a continuous function.

Analogously, define

$$W^n(Z)$$

to be the value of an optimal policy under the above rules with the overriding contraint, only initially operative, that project n must be selected first. Then

$$W(Z) = Z \qquad \text{if A is empty}$$

$$W(Z) = \max_{n \in A} W^n(Z) \qquad \text{otherwise.}$$

Note that when $Z = -\infty$, in effect there are no restrictions and $W(-\infty)$ is the value of an optimal policy in the original problem. The theorem will be proved if we can show that

$$W(-\infty) = W^1(-\infty),$$

where without loss of generality the projects are so ordered that

$$Z^1 = \max_n Z^n.$$

Lemma 1:

$$\frac{dW}{dZ} = \prod_{n=1}^{N} G^n(Z) \qquad \text{a.e.}$$

Proof: In whatever order they are used, projects are abandoned if and only if their reservation prices fall to Z or below. The abandoning of each project is an independent stochastic event. The probability that the entire process is retired and Z is collected is therefore

$$\prod_{n=1}^{N} G^n(Z)$$

In the problem as we have formulated it, there are essentially a finite number of reservation prices. Suppose $Z \neq Z_i^n$ for all i and n. Define

$$\underline{Z} = \max_{Z_i^n < Z} Z_i^n$$

$$\bar{Z} = \min_{Z_i^n > Z} Z_i^n.$$

Let X and Y be any two points in the interval $(\underline{Z},\bar{Z})$. Then $A(X) = A(Y)$ and $G^n(X) = G^n(Y)$ under all possible states of the system.

Consider a policy identical to the optimal policy for X, except that Y is collected instead of X. This policy is feasible for Y, hence

$$W(Y) \geq W(X) + (Y-X) \prod_{n=1}^{N} G^n(X).$$

By a symmetric argument,

$$W(X) \geq W(Y) + (X-Y) \prod_{n=1}^{N} G^n(Y).$$

Thus, the function W is linear in the interval $(\underline{Z},\bar{Z})$ and piecewise linear in $(-\infty, \infty)$. Except for (a finite number of) policy switch points, the slope of W(Z) is given by $\prod_{n=1}^{N} G^n(Z)$.

Lemma 2:

$$\frac{dW}{dZ} = \frac{dW^1}{dZ}$$

for almost all $Z \leq Z^1$.

Proof: That

$$\frac{dW^1}{dZ} = \prod_{n=1}^{N} G^n(Z) \qquad \text{a.e.}$$

for $Z \leq Z^1$ follows by the same logic as Lemma 1.

Lemma 3:

$$W(Z^1) = W^1(Z^1).$$

Proof: Follows from the relevant definitions.

Theorem:

$$W(-\infty) = W^1(-\infty).$$

Proof: Follows from Lemmas 2 and 3.

Computation

For project n, we need to calculate $\{V^n_{ji}\}$ only for that state i(n) currently occupied. Once V^n_{ii} is determined, further calculations for project n are unnecessary until and unless project n is actually selected.

The equations (7), (8) are decomposable by project n and by originating state i, so that the calculations of $\{V^n_{ji}\}$ with j varying, i and n fixed, can be done independently of other i and n. Without loss of generality, therefore, we suppress indices i, n and show how to calculate the solution to the system

$$V_i = R_i + \beta_i \Sigma P_{ij} V_j \tag{10}$$

$$V_j = \max\{V_i, R_j + \beta_j \Sigma P_{jk} V_k\} \qquad \forall j \neq i \tag{11}$$

where

$$V_i = V^n_{ii}$$

$$V_j = V^n_{ji}$$

The system of equations (10), (11) is of a classical form familiar from input-output theory with variable techniques, or Markov chain theory with alternative policies.[11] Two basic solution methods are commonly available.

Perhaps the easiest algorithm is successive approximations. The

[11] See, e.g., Weitzman [1967] or Howard, [1960].

iterations

$$V_i(t+1) = R_i + \beta_i \Sigma P_{ij} V_j(t) \tag{12}$$

$$V_j(t+1) = \max\{V_i(t),\ R_j + \beta_j \Sigma P_{jk} V_k(t)\} \qquad j \neq i \tag{13}$$

converge to the solution of (10), (11) from any initial $V_i(0)$, $\{V_j(0)\}$.

Problems of immense dimensions could be calculated because the principal effective constraint on (12), (13) is not computation, but storage size.

Furthermore, after project n has been selected and has moved from state i to state k, the previously calculated $\{V^n_{ji}\}$ are natural starting values for $\{V^n_{jk}(0)\}$, the initial approximations for $\{V^n_{jk}\}$ under the new state k.

An alternative approach is policy iteration. Starting with a prescribed policy in each state $j \neq i$ of choosing either to continue or to stop, we calculate the solution to the (now linear) set of value equations to which (10), (11) reduces. These are used to select a new policy on the basis of value maximization, which defines another step of the iteration. The advantage of this approach is finite, speedy convergence. The disadvantage, considerable in large problems, is that a matrix must be inverted at each iteration.

More exotic approaches, like fixed point algorithms, could in principle be employed.

Whatever methods are used to solve (10), (11), we note the computational superiority of the reservation price approach over traditional dynamic programming. If there are N projects, each with I states, there are at most NI^2 numbers to be calculated by the present approach. By the traditional backwards recursion dynamic programming approach of equation (3), there are at most I^N numbers to be determined. If $I = N = 10$, this is a difference between computing one thousand numbers and ten billion numbers!

Applications

(1) Search

Studying simple box search is a very useful way to understand the basic features of bandit process solutions because the equation defining the reservation price of a simple box

$$Z = -C + \beta \Sigma q_j \max\{Z, Y_j\} \tag{14}$$

is a miniature version of the bandit process reservation price equations (10), (11).

From (14), the reservation price of a box is completely insensitive to the probability distribution of rewards at the lower end of the tail. Any rearrangement of the probability mass located below Z leaves Z unaltered. It is important to understand this feature. Considering that a box could be opened at any time, the only rationale for opening it now is the possibility of drawing a relatively high reward. That is why the lower end of its reward distribution is irrelevant to the order in which box i

should be sampled even though it may well influence the value of an optimal policy.

On the other hand, as rewards become more dispersed at the upper end of the distribution, the reservation price increases. Other things being equal, it is optimal to sample first from distributions which are more spread out or riskier in hopes of striking it rich early. This is a major conclusion. Low-probability high-payoff situations should be prime candidates for early investigation even though they may have a smaller chance of ending up as the source ultimately yielding the maximum reward when search ends.

The standard comparative statics exercises performed on (14) yield anticipated results. Reservation price decreases with greater search cost, increased search time, or a higher interest rate. Moving the probability mass of rewards to the right makes Z larger. Thus although there is no necessary connection between the mean reward and the reservation price, there is a well-defined sense in which higher rewards increase the reservation price. Similarly, performing a mean preserving spread on the distribution function makes Z bigger. In this sense a riskier distribution of rewards implies a higher reservation price.

(2) Scheduling

The reservation prices of the deterministic resource pool problem are[12]

$$Z_i^n = -\frac{1}{1-\alpha} \inf_{T \geq i} \frac{\sum_{t=i}^{T} C_t^n \alpha^{t-i}}{\sum_{t=i}^{T} \alpha^{t-i}} \qquad (15)$$

[12]See Weitzman [1976].
In the present context α^{t-i} means α multiplied by itself t-i times.

The reservation price of a pool is (a negative multiple of) the minimum equivalent stationary cost per barrel of oil from that source, which we might call the implicit cost of the pool.

Converting arbitrary cost streams to stationary equivalents for the purpose of finding the cheapest alternative is an old economist's trick. The optimality of the max Z rule can in a sense be interpreted as justifying this heuristic procedure under certain conditions.

Note that the implicit cost of a pool reduces to its marginal cost for the special case of nondecreasing costs.

At the opposite extreme, with decreasing costs over the entire range, the infimum in (15) is obtained for $T = \infty$. Once an infinite capacity non-increasing cost source is opened up, in an optimal policy it should operate forever.

To illustrate the typical form of an optimal policy for exploiting depletable natural resources, consider the following simple example. Suppose there are but two resource pools. Each pool has an initial range of decreasing costs, followed by a final section of increasing costs. Let the pools be ordered so that the first has lower implicit cost than the second. The optimal strategy will be to initially exploit the pool with the lowest implicit cost, pool number one. This will be done until the marginal cost of extracting one more unit (in the increasing cost range) becomes greater than the implicit cost of the second pool. At that time pool two will start being exploited, and it will be the exclusive source until its marginal cost in the increasing cost range becomes greater than the marginal cost of pool one. Then pool one or two will alternately be exploited, depending on which is currently the cheaper source at the margin.

Reservation prices for stochastic resource pool problems are often easy to calculate and typically are interpretable as a probabilistic version of (15). For example, if it costs a fixed overhead charge of K^n to open pool n and a variable cost C^n per unit extracted thereafter, but reserves are a random variable τ, then

$$Z_o^n = -\frac{1}{1-\alpha} \quad \frac{K^n + C^n E \sum_{t=o}^{\tau} \alpha^t}{E \sum_{t=o}^{\tau} \alpha^t}$$

before the pool is opened (state i=o). For $i \geq 1$,

$$Z_i^n = \frac{-C^n}{1-\alpha}$$

after the pool is opened. With the above specification, once a pool is tapped in an optimal policy it is run until dry.

Note that for a situation where all pools are the same and there is an unlimited collection of them, the optimal policy will be cyclic or recursive. The same conclusion holds if there are several classes of pools, each class containing an infinite number of identical pools (because in an optimal strategy only pools from one class will be tapped). This is why so many of the standard operations research models with stationary probability distributions (for example, equipment durability selection and replacement, inventory, maintenance, and production scheduling, or capacity expansion) end up having a repetitive solution which may be universally characterized as follows. At each decision node, choose the strategy element with lowest expected equivalent stationary cost.

(3) Learning

In simple box search we observed that if rewards are more spread out, or in other words the probability distribution collapses more completely when drawing a sample from it, the reservation price is increased. This principle generalizes.

The possibility of learning increases the reservation price of a project by a premium reflecting how rapidly the probability spread of rewards narrows as more steps of the project are undertaken. This is a crucial feature of information gathering processes.

The learning effect is quite pivotal in the numerical R&D example. Before the feasibility study, project A with probability .8 has a .625 chance of success and with probability .2 has a 0 chance of success; likewise project B with probability .5 has a .8 chance of success and with probability .5 has a 0 chance of success. Viewed as a binomial event, the pre-feasibility variance of A is (.5) (.5) = .25, and of B is (.4) (.6) = .24. The expected post-feasibility variance of A is (.8) (.625) (.375) = .19, and of B is (.5) (.8) (.2) = .08. On average, the feasibility study reduces the variance of A by only .06 compared with .16 for B. The learning or information effect is strong enough to favor starting with B, which is an inferior project to A in all other respects. If in the first stage (feasibility study) the cost is C_1, and the probability of success is P_1, and in the second stage (full scale R&D) the cost is C_2 and the probability of success conditional upon passing stage 1 is P_2, then $Z = -(C_1+P_1C_2)/P_1P_2$. The reader can verify $Z^B > Z^A$.

With the multi-armed bandit problem, there is a comparable learning effect favoring arms with more diffuse priors. Mean preserving spreads of bandit arm priors increase the reservation price. On average, the reservation price of an arm is expected to decline over time, as the probability distribution for the arm contracts after sampling. There will of course be instances when the decision maker should sample a high-variance, low-mean arm even though he knows he is likely to abandon it in favor of a low-variance, high-mean arm.

From the analysis of theoretical models, new insights are possible into the properties of the R&D search process as a whole. Examination of how reservation prices tend to change with the development of a line of research allow one to describe the way in which research can be expected to proceed.

To take an example, it is possible to infer from the Roberts/Weitzman analysis of a single research project that the reservation price can be expected to fall over time when the uncertainty about ultimate rewards is high in comparison to the costs required to complete the research project. If a planner is facing a situation with several projects of this type, he can expect the optimal selection rule to involve considerable switching and reswitching between projects. The real world implication is that in such situations it may be optimal to pursue a parallel research effort with more than one project being developed at the same time. By contrast, when the uncertainty about ultimate rewards is low relative to R&D costs, the reservation price of a project can be expected to rise

because as more stages are developed, total remaining costs to completion are lowered, without much gain in information. In a situation with several projects of this type, the optimal selection rule will tend to go with one "best" project from beginning to end.

(4) A Composite Example

Most bandit processes exhibit features common to more than one "pure case". Consider, for example, the following stylized mining problem which illustrates nicely the interacting of search, scheduling, and learning.

A company seeks to extract a natural resource at a fixed predetermined flow rate from any one at a time of a number of different potential mine locations.

For notational convenience we drop the superscript referring to mine location.

If a given location contains ore, it contains enough to last for τ years at the fixed extraction rate, where τ is a random variable with a known distribution. The initial overhead cost of opening the mine would be K, and the operating cost would be C per year.

By paying a (relatively inexpensive) testing cost of C_1, the company can perform a geological survey which, with probability $1-p_1$, will rule the site out as altogether implausible. If the survey is favorable, by paying a (relatively expensive) cost of C_2, the company can do a test drilling which will strike ore with probability p_2.

The interest rate is r. What is the next site to explore?

Applying the formula for Z_i^n, we can derive

$$Z_o = - \frac{C_1 + p_1 C_2 + p_1 p_2 (K + (C/r)(1 - E e^{-r\tau}))}{p_1 p_2 (1 - E e^{-r\tau})}$$

With $C_1 = C_2 = 0$, this is a pure scheduling problem; with $C_1 = 0$, $C_2 > 0$, a search aspect is tacked on, with $C_1 > 0$, $C_2 > 0$, a learning stage is added.

Without a first learning stage, the value of Z would be

$$- \frac{C_2 + p_1 p_2 (K + (C/r)(1 - E e^{-r\tau}))}{p_1 p_2 (1 - E e^{-r\tau})}$$

Thus, provided $C_1 < (1-p_1)C_2$, adding the possibility of a geological survey makes it more likely a decision maker will want to investigate a site, even though it increases overall cost if the site contains ore.

The strength of this effect increases as p_1 is smaller. Holding other things constant, including the overall probability that the site contains ore $p_1 p_2$, a decrease in p_1 means an increase in p_2. The less likely is the geological survey to be successful, the more discriminating power does it have, and the more desirable is it to investigate the site now because the relatively expensive test drilling is less likely to be in vain.

(5) Irreversible Investment and The Option Value of Flexibility

The following stylized example shows the value of flexibility in determining reservation prices.[13]

[13]For background, see Henry [1974] and the references cited there.

Suppose that an irreplaceable asset (a forest, say) can be put to some number of alternative uses. When used for a given purpose, the annual (imputed) income of the forest follows a random walk from current value s, with zero drift and annual variance σ^2. (For notational convenience we drop the superscript referencing use option). An irreversible usage could be represented by $\sigma^2 = 0$. The policy question is which usage to favor at the current time.

In our terminology, s is the state of a usage and $U(s;Z)$ is the expected present discounted value of an optimal policy if the only alternative to the proposed usage is irreversibly cutting down the forest for a present discounted reward of Z. Let $Z(s)$ be the reservation price of the usage under consideration. If r is the instantaneous interest rate, for $Z < Z(s)$ there is a sufficiently small δt such that

$$U(s;Z) = s\,\delta t + (1 - r\delta t)EU(s + X\sigma\sqrt{\delta t};Z)$$

where X is a random variable taking on the values +1 and -1 each with probability $\frac{1}{2}$.

Employing Taylor series approximations and passing to the limit yields the differential equation.[14]

$$\frac{\sigma^2}{2} U_{ss} - rU + s = 0$$

for $Z > Z(s)$.

[14] The general technique being used is explained in greater detail in Roberts and Weitzman [1979]. Here we are more concerned with presenting and interpreting a new result than deriving it rigorously.

At $Z = Z(s)$ we have the condition

$$U(s;Z(s)) = Z(s).$$

Also, for sufficiently small δt,

$$Z(s) = s\,\delta\,t + (1 - r\delta t)\,[\tfrac{1}{2}U\,(s + \sigma\sqrt{\delta}t;Z(s)) + \tfrac{1}{2}\,Z(s)],$$

which yields, passing to the limit after a Taylor expansion,

$$U_s(s;Z(s)) = 0.$$

The differential equation, along with the boundary conditions, can be explicitly solved to yield the formula

$$Z(s) = \frac{1}{r}\,(s + \frac{\sigma}{\sqrt{2r}})$$

The reservation price $Z(s)$ is the sum of the "certainty equivalent"

$$\frac{s}{r}$$

plus the "option value"

$$\frac{\sigma}{r\sqrt{2r}}$$

The option value of a given flexible usage measures its incremental worth over the hypothetical irreversible alternative of receiving its current certainty equivalent income forever. The option value is directly proportional to σ, which parameterizes the uncertainty in the difference between the flexible and irreversible options.

To give some idea of the orders of magnitude involved, suppose $s = \$1$ million, $\sigma = \$100$ thousand, $r = 5\%$. Then the certainty equivalent is \$20 million, whereas the reservation price is \$26.4 million. The option value, here 32% of the certainty equivalent, can easily be a non-negligable component of cost-benefit analysis.

References

Gittens, J.C.[1979]: "Bandit Processes and Dynamic Allocation Indices", J.R. Statist. Soc. B, 41, No. 2, 148-177.

Henry, C. [1974]: "Option Values in the Economics of Irreplaceable Assets", R.E.Stud, 89-104.

Howard, R.A. [1960]: Dynamic Programming and Markov Processes, M.I.T. Press.

Kadane, J.B., [1969]: "Quiz Show Problems", Journal Math. Anal. and Applic., 27, 609-623.

Kadane, J.B., and H.A. Simon [1972]: "Optimal Strategies for a Class of Constrained Sequential Problems", Annals Stat. 5, 237-255.

Lippman, S.A., and J.J. McCall [1976]: "The Economics of Job Search: A Survey", Economic Inquiry, 4, 155-189.

Roberts, K.W.S., and M.L. Weitzman [1979]: "Funding Criteria for Research, Development, and Exploration Projects", submitted to Econometrica.

Rothschild, M. [1974]: "A Two-Armed Bandit Theory of Market Pricing", J. Econ. Th., 9, No. 2, 185-202.

Weitzman, M.L. [1967]: "On Choosing an Optimal Technology", Man. Sci., No. 5, 413-428.

Weitzman, M.L. [1976]: "The Optimal Development of Resource Pools", J. Econ. Th. 12, No. 3, 351-364.

Weitzman, M.L. [1979]: "Optimal Search for the Best Alternative", Econometrica, 47, No. 3, 641-654.

Whittle, P. [1980]: "Multi-armed Bandits and the Gittens Index", J.R. Statist. Soc. B, 42, No. 2.