MANAGEMENT SCIENCE
Vol. 18, No. 7, March, 1972
Printed in U.S.A.

RISK-SENSITIVE MARKOV DECISION PROCESSES*

RONALD A. HOWARD† AND JAMES E. MATHESON‡§

This paper considers the maximization of certain equivalent reward generated by a Markov decision process with constant risk sensitivity. First, value iteration is used to optimize possibly time-varying processes of finite duration. Then a policy iteration procedure is developed to find the stationary policy with highest certain equivalent gain for the infinite duration case. A simple example demonstrates both procedures.

1. Introduction

An important limitation of previous analyses of Markov reward and decision processes [1] is that there has been no provision for incorporating risk sensitivity. The present paper shows how risk sensitivity may be treated in such processes if the utility function is exponential in form (constant risk aversion).

2. Risk Sensitivity

If a decision maker subscribes to certain arguments regarding risky propositions [3], then his risk preference may be represented by a utility function that assigns a real number to each possible outcome. Furthermore, his preference ranking of these uncertain propositions, called "lotteries," will be in accordance with the expectation of these numbers. We shall call this expectation "the utility of the lottery." Thus if v is the real-valued outcome of a lottery, $u(v)$ is the utility to be assigned to the outcome v. We assume that larger values of v are preferred, and therefore that $u(\cdot)$ is monotonically increasing.

Certain Equivalent

An important concept of our discussion will be that of certain equivalent. The certain equivalent of a lottery is the outcome whose utility is the same as the utility of the lottery. We use the symbol $\bar{v}$ for the certain equivalent of a lottery on an outcome v with utility $u(v)$,

$$u(v) = u(\bar{v}).$$

Exponential Utility

In many situations a decision maker is willing to accept what we call the "delta property": if all prizes in a lottery are increased by the same amount Δ, then he wants his certain equivalent for the lottery to increase by Δ,

$$(\bar{v} + \Delta) = \Delta + \bar{v}.$$

A decision maker who accepts the delta property is saying that his certain equivalent for any proposed new lottery is independent of his current wealth. While few decision makers would accept the delta property in all circumstances, it can be a very useful approximation in practical problems.

* Received February 1971; revised April 1971.

† Stanford University.
‡ Stanford Research Institute.

§ The authors express their appreciation to Arthur F. Veinott, Jr. for the many helpful suggestions he has made regarding the presentation of this paper.

It is easy to show [4] that the utility function of someone accepting the delta property must be either linear or exponential. The linear case implies risk indifference; it was treated in [1]. The exponential case may be described by writing the utility function in the form

$$u(v) = -(\operatorname{sgn}\gamma)e^{-\gamma v} \tag{2.1}$$

with inverse

$$u^{-1}(x) = -\frac{1}{\gamma}\ln(-(\operatorname{sgn}\gamma)x), \tag{2.2}$$

where γ is the risk aversion coefficient, and $\operatorname{sgn}\gamma$ denotes the sign of γ. The exponential utility function implies

$$u(v+\Delta) = -(\operatorname{sgn}\gamma)e^{-\gamma(v+\Delta)} = e^{-\gamma\Delta}u(v); \tag{2.3}$$

adding a constant Δ to all prizes in a lottery causes their utilities to be multiplied by $e^{-\gamma\Delta}$.

A positive risk aversion coefficient implies risk aversion: establishing a certain equivalent for a lottery that is less than its expected value. A negative risk aversion coefficient implies risk preference, the contrary behavior. We shall characterize any risk attitude that is not risk indifferent as "risk sensitive."

3. A Time-Varying Markov Reward Process

Consider an N-state time-varying Markov process that has transition probability matrix $P(n)$ at a time when n transitions (stages) remain. The ijth element of this matrix $p_{ij}(n)$ is the probability that the process will make its next transition to state j if it currently occupies state i and has n transitions remaining. A transition from state i to state j on the nth transition pays a reward $r_{ij}(n)$, positive or negative. The reward structure for any transition n is therefore summarized by a reward matrix $R(n)$ with elements $r_{ij}(n)$. We are interested in analyzing this reward process in the risk sensitive case.

Suppose that the reward process is to be allowed to continue for $n+1$ transitions and that the process is currently in state i. The total reward the process will generate before termination we define to be $v_i(n+1)$. If the decision maker satisfies the delta property, this uncertain reward will have a certain equivalent $\tilde{v}_i(n+1)$ that is independent of his wealth, and thus independent of rewards he has received previously. This quantity represents the amount that he would be willing to take for certain instead of receiving the reward generated by the Markov process.

To compute this quantity, consider what would happen on the next transition. If the process makes its next transition from state i to state j, it will earn a reward $r_{ij}(n+1)$ and place the decision maker in a position where he has n transitions remaining. This position will have a certain equivalent $\tilde{v}_j(n)$ that is independent of $r_{ij}(n+1)$ as well as of previous rewards because the decision maker satisfies the delta property. With probability $p_{ij}(n+1)$ the next transition will be to state j, whereupon the decision maker will use the delta property to assign the certain equivalent $r_{ij}(n+1) + \tilde{v}_j(n)$. Consistency requires that

$$u(\tilde{v}_i(n+1)) = \sum_{j=1}^{N} p_{ij}(n+1)u(r_{ij}(n+1)+\tilde{v}_j(n)), \qquad n = 0, 1, 2, \cdots. \tag{3.1}$$

The utility of accepting the certain equivalent must be the same as the utility of continuing. The quantities $\tilde{v}_j(0)$ can be assigned directly by the decision maker.

Using the property of Equation (2.5), we can write Equation (3.1) as

$$u(\tilde{v}_i(n+1)) = \sum_{j=1}^{N} p_{ij}(n+1)e^{-\gamma r_{ij}(n+1)}u(\tilde{v}_j(n)), \qquad n = 0, 1, 2, \cdots. \tag{3.2}$$

Note that this substitution has allowed us to write an equation relating the utility of the Markov reward process lottery at two successive stages. This happy development is directly traceable to the fact that the risk attitude of the decision maker is independent of his wealth.

We define the utility of the reward process when it occupies state j with n transitions remaining as $u_j(n)$,

$$u_j(n) = u(\tilde{v}_j(n)) = -(\operatorname{sgn}\gamma)e^{-\gamma\tilde{v}_j(n)}, \qquad n = 0, 1, 2, \cdots. \tag{3.3}$$

Equation (3.2) becomes

$$u_i(n+1) = \sum_{j=1}^{N} p_{ij}(n+1)e^{-\gamma r_{ij}(n+1)}u_j(n), \qquad n = 0, 1, 2, \cdots. \tag{3.4}$$

We can directly interpret the terms in Equation (3.4). Thus in the risk-averse case of positive γ, $e^{-\gamma r_{ij}(n)}$, which we shall call $e_{ij}(n)$, is the negative utility or "disutility" of the reward r_{ij} associated with the transition from state i to state j at transition time n. The term "disutility" will hereafter be applied to $e_{ij}(n)$ regardless of the sign of γ. We shall let $q_{ij}(n)$ be the symbol for the product of transition probability and disutility,

$$q_{ij}(n) = p_{ij}(n)e^{-\gamma r_{ij}(n)} = p_{ij}(n)e_{ij}(n).$$

We shall call it the "disutility contribution" of the transition from state i to state j at time n, and define the disutility contribution matrix $Q(n)$ with elements $q_{ij}(n)$. It is clear that all elements of $Q(n)$ are nonnegative.

Now we can write Equation (3.4) in the simple form,

$$u_i(n+1) = \sum_{j=1}^{N} q_{ij}(n+1)u_j(n), \qquad n = 0, 1, 2, \cdots. \tag{3.5}$$

This equation or Equation (3.4) provides a recursive relation for computing the successive utilities of the process. To find the certain equivalents of the process we then use the result implied by Equation (3.3),

$$\tilde{v}_i(n) = -\frac{1}{\gamma}\ln[-(\operatorname{sgn}\gamma)u_i(n)]. \tag{3.6}$$

4. The Stationary Markov Reward Process

Our relationships assume important special form when the transition probabilities and rewards are the same for all transitions:

$$p_{ij}(n) = p_{ij}, \quad r_{ij}(n) = r_{ij}, \qquad n = 0, 1, 2, \cdots.$$

In this case the process is completely specified by the transition probability matrix P and the reward matrix R or, equivalently, by the disutility contribution matrix Q with elements defined by

$$q_{ij} = p_{ij}e^{-\gamma r_{ij}}. \tag{4.1}$$

Example

To illustrate the recursive evaluation of certain equivalents and all other computations in this paper, we shall use the taxicab example of [1]. This example describes the behavior of a taxicab driver who carries out his business in three towns that we identify

with states 1, 2, and 3. His trips between towns are governed by transition probabilities; each trip entails a corresponding reward. The data for the example appear in Table 4.1. In towns 1 and 3, the driver has three alternatives: to cruise, to go to a taxi stand, or to wait for a radio call. In town 2, only the first two alternatives are available. The variable k is used to index the alternatives in each state.

We shall now investigate the policy of going to a stand, the policy composed of alternative 2 in each state. It is shown in the reference that this policy produces the highest possible average reward per transition for the process. The transition probabilities and rewards for this policy are given by:

$$P = \begin{bmatrix} 1/16 & 3/4 & 3/16 \\ 1/16 & 7/8 & 1/16 \\ 1/8 & 3/4 & 1/8 \end{bmatrix}, \qquad R = \begin{bmatrix} 8 & 2 & 4 \\ 8 & 16 & 8 \\ 6 & 4 & 2 \end{bmatrix}. \tag{4.2}$$

We shall use the risk aversion coefficient $\gamma = 1.0$. For this policy and γ, the matrix Q defined by Equation (4.1) becomes:

$$Q = \begin{bmatrix} 2.09664 \times 10^{-5} & 1.01501 \times 10^{-1} & 3.43418 \times 10^{-3} \\ 2.09664 \times 10^{-5} & 9.84683 \times 10^{-8} & 2.09664 \times 10^{-5} \\ 3.09844 \times 10^{-4} & 1.37367 \times 10^{-2} & 1.69169 \times 10^{-2} \end{bmatrix}. \tag{4.3}$$

Table 4.2 shows the results of the computation using Equations (3.5) and (3.6). It provides the certain equivalent $\tilde{v}_i(n)$ for each state i and for a range of number of stages remaining n. It also shows the differences in the certain equivalent of each state on successive stages under the convenient assumption that $\tilde{v}_1(0) = \tilde{v}_2(0) = \tilde{v}_3(0) = 0$.

Notice that this difference approaches the constant 4.07438 for all states as the number of transitions remaining grows. We can interpret this constant difference as the amount that a person with a risk aversion coefficient of 1 should be just willing to pay to increase the number of stages available to him by one when he already has several available. We shall call it the "certain equivalent gain" of the process.

Notice also that the differences between the certain equivalents of states at the same stage seem to approach a constant as n increases. For example, $\tilde{v}_1(n) - \tilde{v}_3(n)$ approaches 1.55518. We interpret this number as the amount that a person with a risk evaluation coefficient of 1 should be just willing to pay to be in state 1 rather than in

TABLE 4.1

Taxicab Example

State	Alternative	Transition Probabilities p_{ij}^k			Transition Rewards r_{ij}^k		
i	k	$j=1$	$j=2$	$j=3$	$j=1$	$j=2$	$j=3$
1	1 (Cruise)	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	10	4	8
	2 (Stand)	$\frac{1}{16}$	$\frac{3}{4}$	$\frac{3}{16}$	8	2	4
	3 (Call)	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{5}{8}$	4	6	4
2	1 (Cruise)	$\frac{1}{2}$	0	$\frac{1}{2}$	14	0	18
	2 (Stand)	$\frac{1}{16}$	$\frac{7}{8}$	$\frac{1}{16}$	8	16	8
3	1 (Cruise)	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$	10	2	8
	2 (Stand)	$\frac{1}{8}$	$\frac{3}{4}$	$\frac{1}{8}$	6	4	2
	3 (Call)	$\frac{3}{4}$	$\frac{1}{16}$	$\frac{3}{16}$	4	0	8

TABLE 4.2

Recursive Evaluation of Taxicab Problem with Risk Aversion under Policy of Going to a Stand in Every Town

Stage	Certain Equivalent			Change in Certain Equivalent with Stage		
	State			State		
	1	2	3	1	2	3
1	2.25421	10.07710	3.47495			
2	9.08988	12.76828	7.49312	6.83567	2.69118	4.01817
3	13.02518	18.08124	11.56472	3.93530	5.31296	4.07160
4	17.19456	22.12856	15.63873	4.16938	4.04732	4.07401
5	21.26756	26.21985	19.71309	4.07300	4.09129	4.07436
6	25.34264	30.29397	23.78746	4.07508	4.07412	4.07437
7	29.41700	34.36847	27.86183	4.07436	4.07450	4.07437
8	33.49138	38.44284	31.93621	4.07438	4.07437	4.07438
9	37.56575	42.51722	36.01058	4.07437	4.07438	4.07437
10	41.64013	46.59159	40.08495	4.07438	4.07437	4.07437

state 3 with a large number of transitions remaining. For further reference we also note that $\tilde{v}_2(n) - \tilde{v}_3(n)$ approaches 6.50664.

We shall now investigate the nature of the certain equivalent gain and the convergence of certain equivalent differences in more detail.

Certain Equivalent Gain

The nature of the utility transformation implied by Equation (3.5) for this case becomes evident if we define the column vector $\mathbf{u}(n)$ with components $u_j(n)$, recall that $Q(n) = Q$ for the stationary case, and write the equation in the form

$$\mathbf{u}(n+1) = Q\mathbf{u}(n), \qquad n = 0, 1, 2, \cdots,$$

with solution

$$\mathbf{u}(n) = Q^n\mathbf{u}(0), \qquad n = 0, 1, 2, \cdots.$$

In the terminology of Appendix A, if the Markov transition probability is irreducible and acyclic, then the disutility contribution matrix Q is irreducible and primitive. Appendix A shows that for this case

$$\lim_{n\to\infty}\left(\frac{1}{\lambda^n}\right)Q^n\mathbf{u}(0) = \lim_{n\to\infty}\left(\frac{1}{\lambda^n}\right)\mathbf{u}(n) = k\mathbf{u}, \tag{4.4}$$

where λ is the largest eigenvalue of Q and $\mathbf{u}$ is the corresponding eigenvector with k chosen so that $u_N = -\operatorname{sgn}\gamma$. In other words, for large n, the utility of any state will be multiplied by λ at each successive stage.

To find the implications for certain equivalents, we apply to the component form of this equation the transformation indicated in Equation (3.6),

$$-\frac{1}{\gamma}\ln\left[-(\operatorname{sgn}\gamma)\lim_{n\to\infty}\left(\frac{1}{\lambda^n}\right)u_i(n)\right] = -\frac{1}{\gamma}\ln\left[-(\operatorname{sgn}\gamma)ku_i\right],$$

$$\lim_{n\to\infty}\left\{-\frac{1}{\gamma}\ln\left[-(\operatorname{sgn}\gamma)u_i(n)\right] - n\left(-\frac{1}{\gamma}\ln\lambda\right)\right\} = -\frac{1}{\gamma}\ln\left[-(\operatorname{sgn}\gamma)u_i\right] - \frac{1}{\gamma}\ln k,$$

or

$$\lim_{n\to\infty}\left[\tilde{v}_i(n) - n\left(-\frac{1}{\gamma}\ln\lambda\right)\right] = \tilde{v}_i + c, \tag{4.5}$$

where $c = -\ln k/\gamma$, and $\tilde{v}_i$ is defined to be the relative certain equivalent of state i,

$$\tilde{v}_i = -\frac{1}{\gamma} \ln [-(\operatorname{sgn} \gamma)u_i]. \tag{4.6}$$

Note that the normalization of u_N causes $\tilde{v}_N$ to be zero.

We find that the certain equivalent of the process will grow linearly with stage at a rate $-\ln \lambda/\gamma$ that we have called the certain equivalent gain. We designate it by the symbol $\tilde{g}$,

$$\tilde{g} = -\frac{1}{\gamma} \ln \lambda. \tag{4.7}$$

Then the asymptotic form of $\tilde{v}_i(n)$ can be written as $n\tilde{g} + \tilde{v}_i + c$.

It is easy to show that the certain equivalent gain must be bounded by the smallest and largest transition rewards. From Equation (A2) of Appendix A with the matrix A replaced by the matrix Q, we have

$$\min_i \sum_j q_{ij} \leqq \lambda \leqq \max_i \sum_j q_{ij}$$

or

$$\min_i \sum_j p_{ij} e^{-\gamma r_{ij}} \leqq \lambda \leqq \max_i \sum_j p_{ij} e^{-\gamma r_{ij}}.$$

Since each row sum above is a weighted average of the disutilities, we can bound each row sum by the largest and smallest disutilities for that row, i.e.,

$$\min_i \min_j e^{-\gamma r_{ij}} \leqq \min_i \sum_j p_{ij} e^{-\gamma r_{ij}} \leqq \lambda \leqq \max_i \sum_j p_{ij} e^{-\gamma r_{ij}} \leqq \max_i \max_j e^{-\gamma r_{ij}}.$$

This inequality implies

$$\exp(-\max_{i,j} \gamma r_{ij}) \leqq \lambda \leqq \exp(-\min_{i,j} \gamma r_{ij}).$$

By taking the natural logarithm and dividing by minus γ, and using Equation (4.7), we find

$$\min_{i,j} r_{ij} \leqq \tilde{g} \leqq \max_{i,j} r_{ij}$$

regardless of the sign of γ.

We can obtain more insight into the nature of the certain equivalent gain by writing Equation (3.5) for the stationary case,

$$u_i(n+1) = \sum_{j=1}^{N} q_{ij} u_j(n),$$

dividing by λ^n, letting $n \to \infty$, and using Equation (4.4),

$$\lambda u_i = \sum_{j=1}^{N} q_{ij} u_j. \tag{4.8}$$

We recall from Equation (4.6) that we can write u_i as

$$u_i = -(\operatorname{sgn} \gamma) e^{-\gamma \tilde{v}_i}$$

and from Equation (4.7) that λ is related to the certain equivalent gain by $\lambda = e^{-\gamma \tilde{g}}$. Then we can write Equation (4.8) as

$$e^{-\gamma(\tilde{g}+\tilde{v}_i)} = \sum_{j=1}^{N} q_{ij} e^{-\gamma \tilde{v}_j}$$

or, using the explicit form of the disutility contribution,

$$e^{-\gamma(\tilde{g}+\tilde{v}_i)} = \sum_{j=1}^{N} p_{ij} e^{-\gamma(r_{ij}+\tilde{v}_j)}. \tag{4.9}$$

As γ approaches zero and therefore the decision maker approaches risk indifference, these equations imply

$$\tilde{g} + \tilde{v}_i = \sum_{j=1}^{N} p_{ij}(r_{ij} + \tilde{v}_j),$$

which are the usual relative value equations for a Markov reward process. In this case the certain equivalent gain becomes the ordinary gain or expected reward per transition of the process, and the $\tilde{v}_i$'s become the relative values of the states.

Thus we can think of Equation (4.9) as the analog of the usual relative value equations for the case of exponential risk sensitivity. Note that they share with the usual equations the property that adding a constant to all $\tilde{v}_i$'s leaves the equations unchanged; consequently, we may solve for the $\tilde{v}_i$'s only to within a constant. We shall therefore call $\tilde{v}_i$ the relative certain equivalent of state i. When the relative certain equivalent of one state, say state N, is set equal to zero, then we can solve Equation (4.9) for the certain equivalent gain and the relative certain equivalents of the other states for any irreducible acyclic Markov process.

Inspection of Equation (4.9) shows that if we add the same constant Δ to all rewards r_{ij}, the new solution of the equations will have the same relative certain equivalents, but a certain equivalent gain increased by Δ: Increasing the rewards by a constant increases the certain equivalent gain by the same constant.

Let us apply the results of this section to the policy formed by the second alternative in each state in the taxicab example. The transition probabilities and rewards for this policy appear in Equation (4.2); the Q matrix, in Equation (4.3). We find that the largest eigenvalue of Q is $\lambda = 0.0170027$. The certain equivalent gain is then computed from Equation (4.7) to be 4.07438. Note that this is the same value for this quantity indicated by the iterative solution.

We find that the eigenvalue associated with this eigenvector is proportional to (0.17413, 0.00123, 0.82464). If we wish to set the relative certain equivalent value of a state to be zero, we see from Equation (4.6) that the corresponding component of the eigenvector must be set to $-(\operatorname{sgn}\gamma)$. Thus if we wish to set $\tilde{v}_3 = 0$ in this example, we must normalize $\mathbf{u}$ so that $u_3 = -1$. With this normalization, the vector $\mathbf{u}$ becomes $(-0.21115, -0.0014935, -1)$. We then use Equation (4.6) to find:

$$\begin{aligned} \tilde{v}_1 &= -\frac{1}{\gamma}\ln[-(\operatorname{sgn}\gamma)u_1] & \tilde{v}_2 &= -\frac{1}{\gamma}\ln[-(\operatorname{sgn}\gamma)u_2] \\ &= -\ln(0.21115) & &= -\ln(1.49349) \\ &= 1.55517, & &= 6.50664. \end{aligned}$$

We note that these relative certain equivalents are the ones observed earlier in the recursive calculation of certain equivalent values for this policy.

5. A Time-Varying Markov Decision Process

Suppose that at any transition n (measured from the end of the process) different transition probability and reward matrices can be used to govern the process. In any state i at transition n there is a choice among various alternatives k that specify the transition probability $p_{ij}^k(n)$ and rewards $r_{ij}^k(n)$ that will characterize the next transition of the process. The number of alternatives available may be different from transition to transition and from state to state. Our problem is to find which alternative should be used at each stage and state in order to maximize the utility or, equivalently, the certain equivalent of the reward subsequently generated by the process.

Let $d_i(n)$ be the number of the best alternative to use in state i when n transitions remain. When we have specified the column vector $\mathbf{d}(n)$ for all n, we have solved the problem. We call $\mathbf{d}(n)$ the optimum policy at time n. We can find the optimum policy by applying the principle of optimality to Equation (3.4). We define $u_i(n)$ to be the highest utility achievable from the process when it occupies state i, and n transitions remain. Then

$$(5.1) \quad u_i(n+1) = \max_k \sum_{j=1}^{N} p_{ij}^k(n+1)e^{-\gamma r_{ij}^k(n)}u_j(n), \qquad n = 0, 1, 2, \cdots,$$

where $d_i(n+1)$ is the maximizing value of k. This equation allows us to compute the optimum policy for all stages as well as the utility of the process under this policy.

If we define $\tilde{v}_i(n)$ as the certain equivalent of the lottery implied by being in state i with n stages remaining under the optimum policy, then we can find $\tilde{v}_i(n)$ from $u_i(n)$ by using Equation (3.6).

Example

Table 5.1 shows the results of applying this procedure to the taxicab decision problem whose data appear in Table 4.1. The risk aversion coefficient used is $\gamma = 1.0$; rewards are assumed to be zero after all available transitions have been made. We see that when only one transition remains, the best policy is to use the first alternative in both states 1 and 2, and the second alternative in state 3. For all greater numbers of transitions remaining, the best policy is to use the first alternative in each state.

The table also shows the change in the certain equivalent of each state as the number of transitions increases. Note that it approaches approximately 8.48 for all states. As we shall see, this is the certain equivalent gain of the policy composed of the first alternative in each state.

Table 5.2 shows the same procedure with the sole change that the risk aversion coefficient has been changed to $\gamma = -1.0$ to illustrate a case of risk preference. When one transition remains, the best policy is to use the first alternative in each state. However, when more than one transition remains, the best policy is to use the third alternative in state 1 and the second alternative in states 2 and 3. Note that the differences in certain equivalent for any state appear to approach approximately 15.87 when more than a few transitions remain. We shall see that this is the certain equivalent gain of

TABLE 5.1

Value Iteration in Taxicab Problem with Risk Aversion Coefficient $\gamma = 1.0$

Stage	Policy			Certain Equivalent			Change in Certain Equivalent with Stage		
	State			State			State		
	1	2	3	1	2	3	1	2	3
1	1	1	2	5.36329	14.67500	3.47495			
2	1	1	1	12.82038	19.94218	12.15518	7.45709	5.26718	8.68023
3	1	1	1	21.39147	27.47853	20.73632	8.57109	7.53635	8.58114
4	1	1	1	29.93613	36.04996	29.18795	8.53466	8.57143	8.45163
5	1	1	1	38.40430	44.59130	37.66343	8.46817	8.54134	8.47548
6	1	1	1	46.88206	53.05974	46.14960	8.47776	8.46844	8.48617
7	1	1	1	55.36650	61.53781	54.63268	8.48444	8.47807	8.48308
8	1	1	1	63.84950	70.02220	63.11497	8.48300	8.48439	8.48229
9	1	1	1	72.33197	78.50518	71.59763	8.48247	8.48298	8.48266
10	1	1	1	80.81462	86.98765	80.08033	8.48265	8.48247	8.48270

TABLE 5.2

Value Iteration in Taxicab Problem with Risk Aversion Coefficient $\gamma = -1.0$

Stage	Policy State			Certain Equivalent State			Change in Certain Equivalent with Stage State		
	1	2	3	1	2	3	1	2	3
1	1	1	1	9.37349	17.32500	8.85351			
2	3	2	2	21.24580	33.19147	21.03776	11.87231	15.86647	12.18425
3	3	2	2	37.11204	49.05794	36.90380	15.86624	15.86647	15.86604
4	3	2	2	52.97850	64.92441	52.77027	15.86646	15.86647	15.86647
5	3	2	2	68.84497	80.79088	68.63673	15.86647	15.86647	15.86646
6	3	2	2	84.71144	96.65735	84.50320	15.86647	15.86647	15.86647
7	3	2	2	100.57791	112.52381	100.36967	15.86647	15.86646	15.86647
8	3	2	2	116.44438	128.39028	116.23614	15.86647	15.86647	15.86647
9	3	2	2	132.31085	144.25675	132.10261	15.86647	15.86647	15.86647
10	3	2	2	148.17732	160.12322	147.96908	15.86647	15.86647	15.86647

the policy of using the third alternative in state 1, the second alternative in states 2 and 3 for every transition.

6. The Stationary Markov Decision Process

Suppose again the individual has various alternatives available for operating the system. However, unlike the earlier case, whatever alternative is selected in a state must be used for all transitions. The alternative k in state i therefore specifies transition probabilities p_{ij}^k and transition rewards r_{ij}^k that will govern the system whenever state i is entered. We describe the policy for the system by a vector $\mathbf{d}$ whose ith element is the decision in state i, the number of the alternative to be used in state i. We seek the policy that will maximize the certain equivalent gain of the system.

We can find the optimum policy by a procedure analogous to policy iteration; it appears in Figure 6.1.[1] First we select an arbitrary policy and solve the relative certain equivalent Equations (4.9) to find the certain equivalent gain and relative certain equivalents corresponding to it. Then we perform a policy improvement by selecting as the new decision in each state the alternative k that maximizes

$$\tilde{V}_i^k = -\frac{1}{\gamma} \ln \left[\sum_{j=1}^N p_{ij}^k e^{-\gamma(r_{ij}^k + \tilde{v}_j)}\right]$$

using the relative certain equivalents of the previous policy. When this has been done for all states, we have a new policy which we evaluate and attempt to improve in the same manner. When no change is possible in the alternative selected in any state, we have found the optimum policy. The proof of the optimality of this procedure appears in Appendix B.

The method can also be used when strictly periodic and hence deterministic processes are possible. In this case $p_{ij} = 1$ for only one j, $j \neq i$, and otherwise $p_{ij} = 0$. The procedure of Figure 6.1 then reduces to the usual policy iteration of [1] and [2] involving the solution of linear simultaneous equations.

The procedure of Figure 6.1 may also be viewed in the utility formulation as shown

[1] We thank Arthur F. Veinott, Jr. for pointing out that the existence of a solution to an equation with the form of Equation (5.1) for the stationary infinite-horizon case is proved by Richard Bellman in the book *Dynamic Programming*, Princeton University Press, Princeton, N.J., 1957, p. 329.

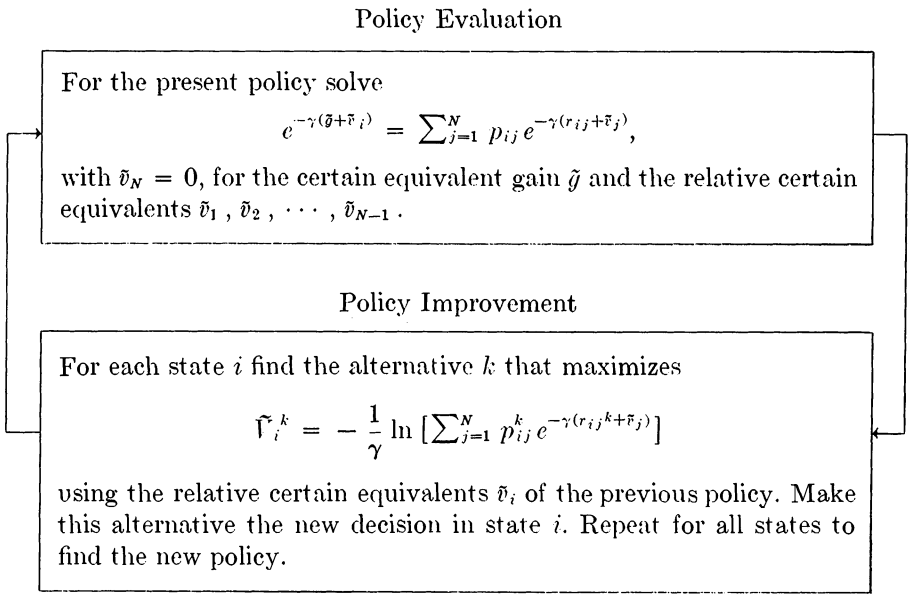

FIGURE 6.1. The Policy Iteration Cycle with Risk Sensitivity—Certain Equivalent Form

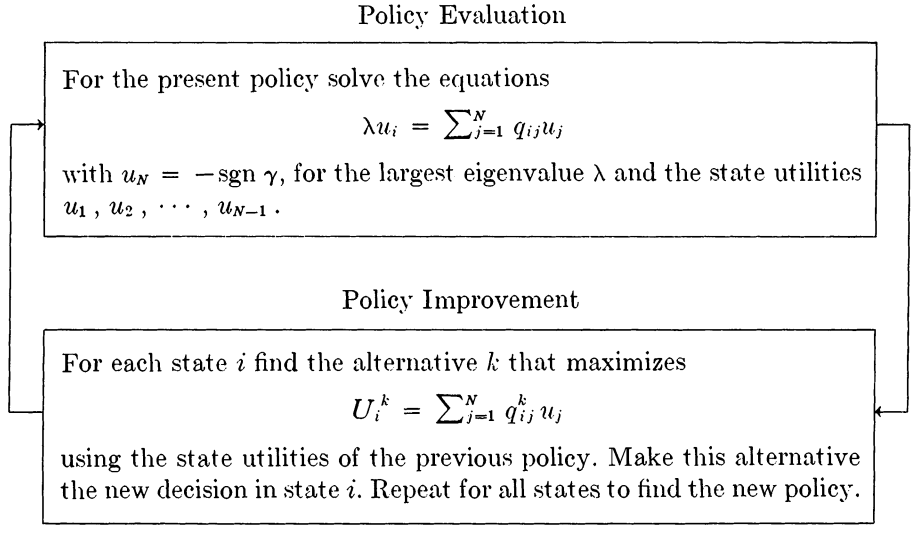

FIGURE 6.2. The Policy Iteration Cycle with Risk Sensitivity—Utility Form.

in Figure 6.2. The quantity q_{ij}^k is defined to be $p_{ij}^k e^{-\gamma r_{ij}^k}$ in agreement with the definition of q_{ij}. The nature of the iteration is clear from the figure.

Example

Table 6.1 brings the results of applying this policy iteration procedure to the taxicab example of Table 4.1. For a range of risk aversion coefficients γ from -1.0 to 1.0, the table shows for each γ the optimum stationary policy, the certain equivalent gain, and the relative certain equivalents of states 1 and 2 relative to state 3. In each case the

TABLE 6.1

Results of Policy Iteration in Taxicab Problem for a Range of Risk Aversion Coefficients

Risk Aversion Coefficient	Policy			Certain Equivalent Gain	Relative Certain Equivalents		No. of Iterations
γ	d			$\tilde{g}$	$\tilde{v}_1$	$\tilde{v}_2$	
−1.0	3	2	2	15.86647	0.20824	12.15414	2
−0.7	3	2	2	15.80924	−0.55936	12.22008	3
−0.50	3	2	2	15.73295	−1.57543	12.30717	3
−0.45	3	2	2	15.70329	−1.96342	12.34054	3
−0.44	2	2	2	15.69655	−1.99156	12.34803	3
−0.3	2	2	2	15.55569	−1.96311	12.49427	3
−0.2	2	2	2	15.34197	−1.89234	12.66973	3
−0.1	2	2	2	14.82655	−1.68692	12.88752	3
−0.01	2	2	2	13.56641	−1.24330	12.94136	3
−0.001	2	2	2	13.36751	−1.18326	12.66499	3
−0.0001	2	2	2	13.34678	−1.17715	12.65638	3
0	2	2	2	13.34454	−1.17647	12.65546	3*
0.0001	2	2	2	13.34216	−1.17579	12.65445	3
0.001	2	2	2	13.32137	−1.16966	12.64571	3
0.01	2	2	2	13.10536	−1.10755	12.54814	3
0.09	2	2	2	10.88344	−0.57796	11.00438	3
0.1	1	2	2	10.62679	−0.47318	10.76876	2
0.16	1	2	2	9.56203	1.03740	9.59114	3
0.17	1	2	1	9.42216	1.20493	9.39993	2
0.2	1	2	1	9.21697	1.21201	8.47294	2
0.24	1	2	1	8.98541	1.22502	7.50767	2
0.25	1	1	1	8.95762	1.22286	7.41749	1
0.3	1	1	1	8.91227	1.18909	7.39437	1
0.5	1	1	1	8.75025	1.04280	7.26815	1
0.7	1	1	1	8.62182	0.90394	7.11748	1
1.0	1	1	1	8.48267	0.73431	6.90733	3

* Run by usual procedure presented in [1].

procedure was started using the policy derived by setting $\tilde{v}_i = 0$ for all i and then entering the policy improvement box. The number of iterations required to attain the optimum policy appear in the last column of the table.

The table includes all values of γ at which policy changes occur. Note that only 5 of the 18 possible policies are optimum for any value of γ. As γ increases from -1, the policy $\mathbf{d} = [3\ 2\ 2]$ is optimum for γ through -0.45. Then $\mathbf{d} = [2\ 2\ 2]$ becomes optimum until γ reaches 0.1. At this point $\mathbf{d} = [1\ 2\ 2]$ is best and remains so until $\mathbf{d} = [1\ 2\ 1]$ becomes optimum at $\gamma = 0.17$. When γ reaches 0.25, the optimum policy changes permanently to $\mathbf{d} = [1\ 1\ 1]$.

As γ ranges from -1 to 1, the certain equivalent gain decreases from 15.86647 to 8.48267. The number of iterations required for convergence is never more than 3.

Figure 6.3 is a plot of certain equivalent gain of the optimum policy versus risk aversion coefficient for the range covered by the table. The optimum policy regions are indicated at the top of the figure. We see that most of the dependence of gain on risk aversion coefficient occurs for γ in the range $-0.4 \leqq \gamma \leqq 0.4$. The figure shows quite clearly just how sensitive the optimum policy and certain equivalent gain are to changes in risk attitude.

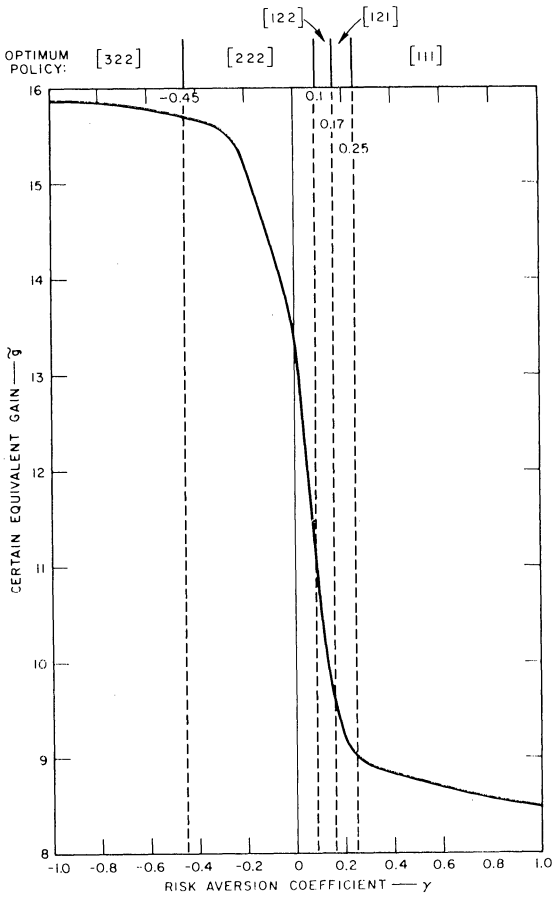

FIGURE 6.3. Certain Equivalent Gain of the Optimum Policy as a Function of Risk Aversion Coefficient.

7. Conclusion

The ability to extend Markov decision process analysis to the case of risk sensitivity should have important application in many areas. We have found the model exceptionally useful in considering optimum buying and selling strategies for a commodity market. Employing this approach in such traditional applications as the optimization of replacement and investment systems will provide interesting new insight into the robustness of maximum-expected-value-per-transition policies.

Appendix A—Matrix Results Utilized in This Paper

The results in this paper are derived from the theory of matrices with nonnegative elements. An excellent discussion of this subject can be found in F. R. Gantmacher, *The Theory of Matrices*, Vol. 2, Chelsea, 1960, Chapter XIII. We shall now summarize the important properties of irreducible nonnegative matrices used in the paper.

A matrix A is called reducible if there is a permutation that can place it in the form

$$\begin{pmatrix} B & 0 \\ C & D \end{pmatrix},$$

where B and D are square matrices; otherwise, A is called irreducible. (An irreducible Markov transition probability matrix represents a process in which all states communicate.)

An irreducible nonnegative matrix A always has a positive eigenvalue λ that is a simple root of the characteristic equation. The moduli of all the other eigenvalues do not exceed λ. To this "maximal" eigenvalue there corresponds an eigenvector with positive components having a unique direction. The matrix A is called "primitive" if some power of A has all elements positive. (A primitive Markov transition probability matrix is called "acyclic.") If A is primitive, then the moduli of all other eigenvalues are strictly less than λ. This means that asymptotically

$$\lim_{n\to\infty} \left(\frac{1}{\lambda^n}\right) A^n \mathbf{x} = \mathbf{v}, \tag{A1}$$

where $\mathbf{x}$ is any vector with nonnegative elements, and $\mathbf{v}$ is an appropriately normalized eigenvector (of unique direction) corresponding to λ.

If A is an irreducible matrix with nonnegative elements a_{ij} and maximal eigenvalue λ, and $\mathbf{x}$ is a vector with positive components x_i, then the following inequalities hold:

$$\min_i \sum_j a_{ij} \leqq \lambda \leqq \max_i \sum_j a_{ij}, \tag{A2}$$

$$\min_i \frac{\sum_j a_{ij} x_j}{x_i} \leqq \lambda \leqq \max_i \frac{\sum_j a_{ij} x_j}{x_i}. \tag{A3}$$

Equation (A3) holds with the equality signs if and only if $\mathbf{x}$ is an eigenvector corresponding to λ.

Appendix B—Proof of Convergence of the Policy Iteration Procedure to an Optimal Policy

The proof will be carried out only for the case of risk aversion, i.e., positive risk aversion coefficient and negative utilities. The same conclusions can be reached for risk preference with appropriate inequality changes.

Assume that the policy iteration has converged on policy B and that policy A is any other policy. Then from the policy improvement step we know that

$$\sum_j q_{ij}^A u_j^B \leqq \sum_j q_{ij}^B u_j^B, \tag{B1}$$

where $\mathbf{u}^B$ is an eigenvector of Q^B corresponding to the maximal eigenvalue λ^B,

$$\sum_j q_{ij}^B u_j^B = \lambda^B u_i^B. \tag{B2}$$

Thus, we have

$$\sum_j q_{ij}^A u_j^B \leqq \lambda^B u_i^B. \tag{B3}$$

Since all components u_j^B are negative, we can write

$$\frac{\sum_j q_{ij}^A \mid u_j^B \mid}{\mid u_i^B \mid} \geqq \lambda^B, \qquad i = 1, 2, \cdots, N. \tag{B4}$$

The left inequality of Equation (A3) provides the condition

$$\min_i \frac{\sum_j q_{ij}^A \mid u_j^B \mid}{\mid u_i^B \mid} \leqq \lambda^A,$$

where λ^A is the maximal eigenvalue of Q^A. Therefore we obtain

$$\lambda^A \geqq \min_i \frac{\sum_j q_{ij}^A \mid u_j^B \mid}{\mid u_i^B \mid} \geqq \lambda^B. \tag{B5}$$

To find the implication for certain equivalent gains of both policies, we note that (B5) implies

$$-\frac{1}{\gamma} \ln \lambda^A \leqq -\frac{1}{\gamma} \ln \lambda^B$$

and $\tilde{g}^A \leqq \tilde{g}^B$. Thus policy iteration can only converge on an optimum policy.

It now remains to be shown that the policy iteration procedure will converge on an optimum policy. Assume that the procedure has improved policy A to arrive at policy B. Then from the policy improvement step we know that

$$\sum_j q_{ij}^B u_j^A \geqq \sum_j q_{ij}^A u_j^A \tag{B6}$$

with inequality for some i.

Recognizing $\mathbf{u}^A$ as an eigenvector of Q^A corresponding to λ^A, we can write

$$\sum_j q_{ij}^B u_j^A \geqq \lambda^A u_i^A. \tag{B7}$$

If $\mathbf{u}^A$ happens also to be an eigenvector of Q^B then the above equation implies directly that $\lambda^B u_i^A > \lambda^A u_i^A$ for some i, which because the u_i^A are negative leads to $\lambda^B < \lambda^A$.

If $\mathbf{u}^A$ is not an eigenvector of Q^B then we rewrite (B7) as

$$\frac{\sum_j q_{ij}^B \mid u_j^A \mid}{\mid u_i^A \mid} \leqq \lambda^A. \tag{B8}$$

In view of Equation (A3), we have

$$\lambda^B < \max_i \frac{\sum_j q_{ij}^B \mid u_j^A \mid}{\mid u_i^A \mid} \leqq \lambda^A. \tag{B9}$$

Note the strict inequality on the left occurs because we have assumed that $\mathbf{u}^A$ is not an eigenvector of Q^B.

In terms of certain equivalent gain these results imply

$$\tilde{g}^A < \tilde{g}^B. \tag{B10}$$

Thus at each iteration the certain equivalent gain must increase.

Since there is only a finite number of possible policies, the procedure will converge in a finite number of iterations.

References

1. Howard, Ronald A., *Dynamic Programming and Markov Processes*, The M.I.T. Press, Cambridge, 1960.
2. ——, *Dynamic Probabilistic Systems* (two volumes), John F. Wiley & Sons, Inc., New York, 1971.
3. Luce, Robert D. and Raiffa, Howard, *Games and Decisions*, John F. Wiley & Sons, Inc., New York, 1957.
4. Raiffa, Howard, *Decision Analysis: Introductory Lectures on Choices under Uncertainty*, Addison Wesley, Menlo Park, 1968.