Monitoring Cooperative Agreements in a Repeated Principal-Agent Relationship
Author(s): Roy Radner
Source: *Econometrica*, Vol. 49, No. 5 (Sep., 1981), pp. 1127-1148
Published by: The Econometric Society
Stable URL: http://www.jstor.org/stable/1912747
Accessed: 07/12/2010 13:15

Econometrica, Vol. 49, No. 5 (September, 1981)

MONITORING COOPERATIVE AGREEMENTS IN A REPEATED PRINCIPAL-AGENT RELATIONSHIP[1]

BY ROY RADNER

The situation in which a principal-agent relationship is repeated finitely many times (T) is formulated as a sequential game. For any Pareto-optimal cooperative arrangement in the one-period game that dominates a one-period Nash equilibrium, and any positive number epsilon, there exists for every sufficiently large T a (noncooperative) epsilon equilibrium of the T-period game that yields each player an average expected utility that is at least his expected utility in the one-period cooperative arrangement, less epsilon.

1. INTRODUCTION

THEORIES OF AGENCY and of the design of incentives in organizations typically portray the members of the organization as players in a noncooperative game. The predictive theory that naturally accompanies this point of view is that of Nash equilibria, including Harsanyi's elaboration of that theory to accommodate situations in which the players have incomplete information about the parameters of the game.

On the other hand, much normative theory of organizations uses the framework of cooperative game theory, with its array of alternative "solution" concepts (value, core, von Neumann-Morgenstern solution, Nash bargaining solution, etc.). Furthermore, empirical observations of organizations reveal widespread cooperative behavior, as well as noncooperative behavior, so that cooperative game theory may have descriptive as well as normative value.

What determines whether members of an organization cooperate or not? Conventional wisdom suggests that cooperation is less likely—or less stable—the more players there are, or the greater the difficulty of communication among the players; cooperation is more likely (stable?) if there are mechanisms whereby the players make binding commitments. Thus theories of industrial organization typically *assume* that when the number of firms in an industry is "large" the resulting equilibrium will be of the noncooperative type, whereas when the number of firms is "small" the outcome may be cooperative (collusive).

The theory of repeated games explores in a formal way another piece of conventional wisdom, namely that when members of an organization have long-lasting relationships they can encourage and maintain cooperative behavior (without the device of binding commitments) by signalling intensions to punish defectors from informal agreements. Indeed, the theory of repeated games

[1]This is a revision and extension of "Monitoring Cooperative Agreements between Principals and Agents," Tech. Report No. 3, Center on Decision and Conflict in Complex Organizations, Harvard University, February, 1979. The research on the previous paper was supported by the Office of Naval Research, Contract No. N00014-77-C-0533 and by the National Science Foundation (Grant SOC76-14768 to the University of California, Berkeley). A preliminary version was presented at the CEME-NBER Conference on Decentralization, University of California, San Diego, Feb. 23–25, 1979.

provides conditions under which *noncooperative* equilibria of the entire sequential game can produce *cooperative* outcomes of the component subgames.

Unfortunately, such results typically require an infinite number of repetitions of the subgame; they are not generally valid for a finite number of repetitions, no matter how large that finite number.[2] However, similar results can be obtained for *approximate* noncooperative equilibria in the finite-repetitions case, at least in situations not involving uncertainty. Such an approximate equilibrium is called an *epsilon equilibrium* if each player's sequential strategy is within epsilon (in utility) of being the best response to the other players' strategies. Thus, one gets the result that, for any fixed positive epsilon, if the number of repetitions is large enough then there are noncooperative epsilon equilibria that have cooperative outcomes in each subgame. In a sense, in finite repetitions of a game, the best is the enemy of the good!

In the principal-agent model, the agent observes a (random) environmental variable and then chooses an action; this leads to an outcome that depends on both the action and the environment. The principal observes this outcome (but neither the agent's action nor the environment), and pays the agent according to a previously announced reward function, which depends on the outcome only.

In equilibria of repeated games that sustain cooperative behavior, each player is "punished" by the others for departures from the informal agreement to cooperate.[3] However, in the principal-agent situation, the principal cannot observe the agent's behavior directly, but only the consequences of his behavior, and those consequences are also influenced by the environment. Therefore, if cooperative agreements are to be sustained as equilibria of the repeated game, the principal must have some statistical method of detecting "cheating" by the agent rapidly enough to deter him from doing so; on the other hand, this method should have a very low probability of triggering false alarms. The main theorems of this paper (Sections 5 and 6) show that this is possible.[4]

In Sections 2 and 3, I present the principal-agent model in the form of a one-period game, and state a few of its properties. In Section 4, I introduce the essential concepts in the theory of epsilon equilibria of finitely repeated games. Section 5 contains the first main result on the existence of epsilon equilibria with cooperative outcomes in T-period repetitions of the principal-agent game, when T is large (but finite). The proof is constructive, and exhibits a family of epsilon equilibrium strategy pairs. Using this family of strategy pairs one can approach

[2] This remark is valid for perfect Nash equilibria; see Section 4.

[3] An early important paper on repeated games (supergames) is by Aumann [1]. Characterizations of perfect Nash equilibria in infinite supergames have been provided by Aumann and Shapley (unpublished) and by Rubinstein [12]. For an analysis of altruism in the context of infinite supergames, see Kurz [6]. Examples of epsilon equilibria of finite supergames have been studied by Radner [7, 8].

[4] The main theorems are related to sequential tests of hypotheses that have power one (see Robbins and Siegmund [11], and the references given there). Since the research for the present paper was completed, I had the opportunity to see a paper by A. Rubinstein [13], in which he uses the law of the iterated logarithm to demonstrate the existence of Nash equilibria with close to Pareto optimal average expected utility in an example of an infinite supergame.

arbitrarily close, in terms of average expected utility per period, to any one-period Pareto-optimal cooperative arrangement that dominates a one-period Nash equilibrium, provided that T is large enough.

Unfortunately, the epsilon equilibria described in Section 5 have the property that, with "small" probability, at some period $t > 1$, the expected utility to the agent (conditional on his information up to that date) of defecting from the equilibrium strategy may exceed epsilon. To exclude such equilibria, I propose in Section 6 a condition of "robustness," and demonstrate the existence of robust epsilon equilibria. Section 7 indicates some possible extensions of the theory.

2. A MODEL OF A SEQUENTIAL PRINCIPAL-AGENT RELATIONSHIP

Consider a principal-agent relationship that lasts T periods. In period t, the agent's action is A_t, a number between 0 and M_a (a positive parameter). The outcome of the agent's action is

$$C_t = \gamma(A_t, Z_t),$$

where Z_t is an exogenous random variable (the "state of nature" in period t). We may interpret the variable A_t as a measure of the agent's effort. The principal observes the outcome of the agent's action, and pays the agent W_t. The resulting one-period utility to the agent is $V(W_t, A_t) = P(W_t) - Q(A_t)$, where the functions P and Q are strictly concave and convex, respectively, and increasing. The one-period utility to the principal is assumed to be a linear function of the outcome and the payment to the agent, increasing in the former and decreasing in the latter. By a suitable choice of units one can express the principal's utility as $C_t - W_t$. The agent can observe the state of nature, Z_t, before taking action, but the principal can observe only the resulting outcome, C_t.

Assume that the functions P, Q, and γ are continuously differentiable, that for every Z the function $\gamma(\cdot, Z)$ is concave and increasing in its first argument (the agent's action), and that the partial derivative of γ with respect to the agent's action is bounded away from 0, uniformly in Z, say $\geqq M' > 0$.

Notice that I have assumed that the agent is risk-averse, whereas the principal is risk-neutral. The main results can be extended to cases in which the principal is risk-averse; see Section 7.

Finally, assume that the states of nature, Z_t, are independently and identically distributed, and bounded.

3. THE ONE-PERIOD GAME

In this section I formulate the principal-agent relationship as a one-period noncooperative game.[5] I therefore omit the subscript t on all the variables. The

[5]For material on the principal-agent problem, see Hurwicz and Shapiro [5], Shavell [15], and Holmström [4], and the references cited there. For a more general organizational setting of the problem, see Groves [3]. An early forerunner of the principal-agent literature was Simon [16].

principal's (pure) strategy is a reward function ω that determines the payment to the agent as a function of the outcome of the agent's action:

$$W = \omega(C).$$

Given the reward function ω, the agent chooses a decision function α that determines his action as a function of the state of nature:

$$A = \alpha(Z).$$

The expected utility to the agent is

$$EV(W,A) = EP\{\omega(\gamma[\alpha(Z),Z])\} - EQ[\alpha(Z)],$$

and the expected utility to the principal is

$$EC - EW = E\gamma[\alpha(Z),Z] - E\omega(\gamma[\alpha(Z),Z]).$$

This is in fact a two-move game, in which the principal moves first, choosing the reward function, and the agent moves second, choosing the decision function. The principal's strategy is the same as his move, but the agent's strategy is a mapping from reward functions ω to decision functions α, since the agent learns the reward function before choosing the decision function. The noncooperative solution to the game is taken to be a Nash equilibrium.

A pair (ω,α) of functions is Pareto optimal if there is no other pair that yields each player at least as high an expected utility, and yields at least one of the players strictly more. In this paper, I shall be concerned with situations in which a Nash equilibrium leads to a pair (ω,α) that is not Pareto optimal. Without providing a precise analysis of the circumstances under which this situation would occur, I shall sketch why it would not be atypical.

If the agent is averse to risk, and the principal is neutral towards risk, then in a Pareto optimum the reward function must be constant on the set of realizable outcomes. Thus, if the pair (ω,α) generates the random reward W, the variance of W is strictly positive, and $w \equiv EW$, then the pair (w,α) is strictly better for the agent than (ω,α), and equally good for the principal. Hence, for some w' slightly less than w, both the agent and the principal would prefer (w',α) to (ω,α).

On the other hand, one can exhibit cases consistent with the above assumptions in which, in a Nash equilibrium, the reward function must be strictly increasing. In such cases, a Nash equilibrium cannot be Pareto optimal.

In the typical formulation of the principal-agent relationship, one assumes that the agent has the option of leaving the relationship and achieving some "reservation utility." This would imply a constraint on the set of reward functions that the principal could use and still retain the services of the agent. The addition of such a constraint to the present model would not change the main result of the paper.

4. EPSILON EQUILIBRIA OF REPEATED GAMES

Suppose now that the one-period game is repeated T times (T finite); the resulting sequential game will be called the *T-period game.* Assume that the utility to a player is the average of the T one-period expected utilities. A pure sequential strategy for a player is a sequence of functions, one for each period; the function for period t determines the player's one-period strategy in period t as a function of all of the information available to the player up to that period. A *Nash equilibrium* of the sequential (T-period) game is a pair of sequential strategies such that each player's sequential strategy is a best response to the other player's sequential strategy.

In the present paper, I shall employ the concept of epsilon equilibrium, a weaker condition than that of Nash equilibrium. For any positive number epsilon, an *epsilon equilibrium* is a pair of strategies such that each player's strategy is within epsilon in average expected utility of being a best response to the other player's strategy. To motivate the use of this equilibrium concept in sequential games, I must digress for the moment to discuss (informally) the problem of "noncredible" Nash equilibria. Nash equilibrium pairs of strategies in the sequential game may involve threats of "punishment" by one player if the other player departs from some prescribed sequential strategy. Unfortunately, a literal application of the definition of Nash equilibrium may result in equilibria in which the players use threats that, in a certain sense, are not credible. To rule out such equilibria, Selten [**14**] has introduced the idea of a "perfect" Nash equilibrium[6] (we shall not need a precise definition here). One can show that in every perfect Nash equilibrium of the (finite) T-period game, the outcome in every period is a Nash equilibrium of the one-period game. On the other hand, in repeated games in which each player can observe the other player's one-period strategies, if T is infinite then there are perfect equilibria of the sequential game that result in the use of "cooperative" pairs of strategies in each one-period game, and in particular in the use of Pareto-optimal pairs of strategies.[7] It is this discontinuity at infinity that motivates the definition of epsilon equilibria. In the same situation, one can show that, for any positive epsilon, if T is sufficiently large then there are perfect epsilon equilibria of the T-period game that result in cooperative behavior in all or most of the component one-period games. In other words, for perfect epsilon equilibria, infinite-horizon repeated games may be approximated well by long finite-horizon games.[8]

Cooperative one-period strategies can be sustained in perfect epsilon equilibria of the T-period game by "trigger strategies." Let (s_1^*, s_2^*) be a Nash equilibrium of the one-period game, and let (s_1, s_2) be a Pareto-superior pair of one-period strategies. A trigger strategy for player 1 is defined as follows: player 1 plays

[6]To be precise, I refer here to the concept of *subgame-perfect* equilibrium.

[7]A. Rubinstein [**12**]; R. Aumann and L. Shapley (unpublished).

[8]These results are illustrated in Radner [**7, 8**]. A more general treatment of epsilon equilibria will be presented in a forthcoming paper.

strategy s_1 as long as player 2 plays strategy s_2; thereafter player 1 plays s_1^*. The best response by player 2 to this trigger strategy is to play s_2 until the last period, and then play a best response to s_1. However, the gain in average per-period utility of doing this, over using the corresponding trigger strategy, will be small if T is large.

The efficacy of such simple trigger strategies in sustaining perfect epsilon equilibria of the T-period game depends on each player being able to rapidly detect departures from the cooperative strategies. In the principal-agent situation considered in this paper, the principal cannot observe the agent's actions directly, but only the consequences of his actions, and these consequences also depend on a random state of nature. Therefore, if cooperative arrangements are to be sustained as epsilon equilibria of the T-period game, the principal must have available some method of detecting any "cheating" by the agent, and doing so rapidly enough to reduce the agent's incentive to cheat to negligible levels. That such a method exists is shown in the next two sections.

5. EPSILON EQUILIBRIA OF THE T-PERIOD PRINCIPAL-AGENT GAME

Let ω^* and α^* be the reward and decision functions, respectively, for a Nash equilibrium of the one-period principal-agent game (Section 3), and let $(\hat{w}, \hat{\alpha})$ be a Pareto-optimal and superior pair, where $\hat{w}$ is constant. Thus, if u^* and v^* are the expected one-period utilities of the principal and agent, respectively, corresponding to the use of (ω^*, α), and if $(\hat{u}, \hat{v})$ is the pair of expected one-period utilities corresponding to $(\hat{w}, \hat{\alpha})$, then suppose that $\hat{u} \geqq u^*$, and $\hat{v} \geqq v^*$. In this section I shall exhibit a class of epsilon equilibria of "long" T-period games that use trigger-type strategies and that achieve average expected utilities that are at least close to $(\hat{u}, \hat{v})$. In the next section I shall show how the same method can be refined to construct corresponding epsilon equilibria that satisfy a condition similar to "perfectness."

The definition of an effective trigger strategy for the agent presents no problem; the agent simply uses the decision function $\hat{\alpha}$ until the first period that the principal does not use the (constant) reward $\hat{w}$; at that period and in each period thereafter the agent optimizes against the reward function announced for the period. Call this sequential strategy σ_A.

It is important to emphasize at this point that in each period the principal's action is an announcement of a reward function, and he is required to use that reward function for that period. The agent then observes the current state, Z_t, and takes an action, A_t.

Defining a suitable trigger strategy for the principal is more difficult. In each period t, based on the history of outcomes through period $(t-1)$, the principal must decide whether to make the payment $\hat{w}$ or to switch to the one-period Nash-equilibrium reward function ω^*. If his switching rule is too lax, then the agent may be able to accumulate a large enough extra expected utility by cheating before getting caught so as to make cheating attractive. On the other hand, if the switching rule is too strict (too "trigger happy"!), then there will be a

substantial probability that the principal will switch to ω^* before the agent ever starts cheating.

Define

$$C_t \equiv \gamma[\alpha_t(Z_t), Z_t];$$

thus C_t is the realized consequence in period t if the agent uses decision function α_t in that period. The consequences C_t are bounded, say by B. Define S_n to be the sum of the realized consequences in periods 1 through n, that is,

$$S_n = C_1 + \cdots + C_n.$$

In particular, let $\hat{C}_t$ denote the realized consequence in period t if the agent uses the decision function $\hat{\alpha}$, and let $\hat{S}_n$ be the corresponding cumulative sum of consequences by the end of period n. The random variables $\hat{C}_t$ are independent and identically distributed, with expected value, say, $\hat{c}$.

Let (b_n) be a strictly increasing sequence of positive numbers ($n \geqq 1$), and define the random variables $\tilde{N}$ and N by:

$$\begin{aligned} \tilde{N} &= \min\{n \geqq 1 : S_n - n\hat{c} \leqq -b_n\}, \\ N &= \min\{\tilde{N}, T\}. \end{aligned} \tag{5.1}$$

Consider the following trigger strategy for the principal: pay the agent $\hat{w}$ in each period 1 through N, and thereafter use the reward function ω^*. I shall denote this strategy by $\sigma_P((b_n))$.

Recall that $\hat{S}_n = \hat{C}_1 + \cdots + \hat{C}_n$. If the agent uses some sequential strategy other than σ_A, then the principal's loss (positive or negative) during periods 1 through n is

$$L_n \equiv \hat{S}_n - S_n.$$

LEMMA 5.1: *If the principal uses the trigger strategy $\sigma_P((b_n))$, then a bound on his expected loss*[9] *during periods* 1 *through N is given by*

$$EL_N \leqq Eb_N + B \leqq b_T + B.$$

PROOF: Recall that the C_t are bounded by B. By the definition of N, $S_N \geqq N\hat{c} - b_N - B$, so that

$$L_N \leqq \hat{S}_N - N\hat{c} + b_N + B. \tag{5.2}$$

Note that $(\hat{S}_n - n\hat{c})$ is a martingale, and N is a bounded stopping time. Hence, by the "systems theorem" for martingales (see e.g., Chung [**2**, equation (3) on

[9]Although it is convenient to interpret L_N as the principal's loss, this is not essential to the argument that follows. What is essential is that this is the cumulated difference in outcome in the direction of the *agent's gain*. This will become clear in Lemma 5.2.

p. 319, and the Corollary on p. 325]),

$$E(\hat{S}_N - N\hat{c}) = 0. \tag{5.3}$$

If one takes the expected value of both sides of (5.2), and uses (5.3) and the fact that the sequence (b_t) is increasing, one immediately obtains the conclusion of the lemma.

Lemma 5.1 establishes a limit to the cumulated expected loss that the principal can suffer up through period N. The next lemma establishes a corresponding limit on the agent's gain. Let A_t be the agent's actual action in period t, and let $\hat{A}_t$ denote what his action would be if he used the decision function $\hat{\alpha}$, i.e., $\hat{A}_t = \hat{\alpha}(Z_t)$. The corresponding difference in the agent's utility is

$$D_t = V(\hat{w}, A_t) - V(\hat{w}, \hat{A}_t),$$

if the agent receives the payment $\hat{w}$. The agent's total gain in utility during periods 1 through n is

$$G_n = D_1 + \cdots + D_n.$$

LEMMA 5.2: *If the principal uses the trigger strategy $\sigma_P((b_n))$, then a bound on the agent's possible expected gain in utility up through period N is given by*

$$EG_N \leqq K(Eb_N + B) \leqq K(b_T + B),$$

where K is some suitably chosen positive number.

PROOF: The regularity properties of γ and V, and the fact that $(\hat{w}, \hat{\alpha})$ is Pareto optimal, imply that there is a (finite) positive number K such that, for any period t and any decision function α_t,

$$EV(\hat{w}, A_t) - \hat{v} \leqq -K(EC_t - \hat{c}), \tag{5.4}$$

where A_t and C_t are determined by α_t. Since the random variables (Z_t) are independent, it follows that for any sequential strategy of the agent, and any partial history $H_{t-1} = (Z_1, \ldots, Z_{t-1})$,

$$E(D_t \mid H_{t-1}) \leqq -KE(C_t - \hat{C}_t \mid H_{t-1}). \tag{5.5}$$

For the purposes of this proof, write $X_0 = 0$, and

$$X_t = D_t + K(C_t - \hat{C}_t) \qquad (t \geqq 1); \tag{5.6}$$

then $(X_1 + \cdots + X_t)$ is a supermartingale, and hence, again by the systems theorem, $E(X_1 + \cdots + X_N) \leqq 0$. The conclusion of the lemma now follows from this last inequality and Lemma 5.1.

Lemma 5.2 shows how to make the principal's trigger strategy strict enough to keep the agent's incentive to cheat small when T is large. It suffices to use a

sequence (b_n) such that (b_n/n) approaches zero as n increases without limit. But how can the principal, with the same trigger strategy, keep small the probability that the agent would be "unjustly" punished if he never cheats? The law of the iterated logarithm provides an answer.[10]

Recall that, by the law of the iterated logarithm (see, e.g., Chung [2]),

$$\liminf_{n\to\infty} \frac{\hat{S}_n - n\hat{c}}{\sqrt{n \ln \ln n}} = -\sqrt{2 \operatorname{var} \hat{C}_t} \equiv -\lambda_0, \tag{5.7}$$

where $\operatorname{var} \hat{C}_t$ denotes the variance of $\hat{C}_t$. Define

$$X \equiv \inf_{n>2} \frac{\hat{S}_n - n\hat{c}}{\sqrt{n \ln \ln n}}.$$

Then $X > -\infty$ almost surely. Hence

$$\lim_{\lambda\to\infty} \operatorname{prob}\{X > -\lambda\} = 1.$$

Hence the following has been proved:

LEMMA 5.3: *For every $\delta > 0$ there exists a $\lambda > 0$ such that*

$$\operatorname{prob}\left\{\hat{S}_n > n\hat{c} - \lambda\sqrt{n \ln \ln n}\ , \textit{for all } n > 2\right\} \geqq 1 - \delta.$$

Define the sequence (b_n^0) by

$$\begin{aligned} b_1^0 &\equiv -\operatorname{ess\,inf}(\hat{C}_1 - \hat{c}), \\ b_2^0 &\equiv -\operatorname{ess\,inf}(\hat{C}_1 + \hat{C}_2 - 2\hat{c}), \\ b_n^0 &\equiv \lambda_0\sqrt{n \ln \ln n}\ , \quad n \geqq 3. \end{aligned} \tag{5.8}$$

Note that (b_n^0/n) approaches zero as n increases without limit.

Define $\boldsymbol{B}$ to be the class of positive sequences (b_n) that satisfy:

$$b_n \text{ are strictly increasing, and } \lim_{n\to\infty} \frac{b_n}{n} = 0; \tag{5.9}$$

$$\text{there exists } \lambda > 1 \text{ such that } b_n \geqq \lambda b_n^0, \quad n \geqq 1. \tag{5.10}$$

In particular, $\boldsymbol{B}$ contains all the sequences (λb_n^0) with $\lambda > 1$.

Recall that $\hat{u}$ and $\hat{v}$ denote the expected one-period utilities of the principal and agent, respectively, under the pair $(\hat{w}, \hat{\alpha})$.

[10] In order to demonstrate the existence of a sequence (b_t) with the desired properties it is sufficient to use an argument based on the strong law of large numbers. The argument used here, which is based on the more powerful law of the iterated logarithm, has the advantage of being constructive. It also provides a family of sequences that, in a sense, grow as slowly as possible.

THEOREM: *For any* $\epsilon > 0$ *there exists a sequence* (b_n) *in* $\boldsymbol{B}$ *and a* T_ϵ *such that for all* $T \geqq T_\epsilon$ *the pair* $[\sigma_P((b_n)), \sigma_A]$ *is an* ϵ *equilibrium, and yields the principal and agent average expected utilities at least* $(\hat{u} - \epsilon)$ *and* $(\hat{v} - \epsilon)$, *respectively.*

PROOF: Recall that u^* and v^* denote the expected one-period utilities of the principal and agent, respectively, under the (Nash equilibrium) pair (ω^*, α^*). Consider a pair $[\sigma_P((b_n)), \sigma_A]$ of sequential trigger strategies, with (b_n) in $\boldsymbol{B}$. The corresponding average expected utility to the principal is

$$\left(\frac{1}{T}\right)\left[(EN)\hat{u} + (T - EN)u^*\right]. \tag{5.11}$$

(Use the martingale systems theorem again.) Recall that $\hat{u} \geqq u^*$. Define

$$\delta \equiv \text{prob}(N < T).$$

Then (5.11) is at least as large as

$$(1 - \delta)\hat{u} + \delta u^*. \tag{5.12}$$

This is as large as $\hat{u} - \epsilon$ if

$$\delta \leqq \frac{\epsilon}{\hat{u} - u^*} \tag{5.13}$$

(with the obvious interpretation if $\hat{u} = u^*$).

If the principal were to switch in any period n to a reward function other than the constant $\hat{w}$, then in that period and thereafter the agent would optimize against the announced reward functions; hence in periods n through T it would be optimal for the principal to use the reward function ω^*. Hence the principal's optimal response to the agent's strategy σ_A is to use the constant reward $\hat{w}$ in *all* periods. The resulting average expected utility to the principal is $\hat{u}$. Therefore, if (5.13) is satisfied, the strategy $\sigma_P((b_n))$ is within ϵ of being optimal against σ_A.

If the agent follows strategy σ_A against $\sigma_P((b_n))$, then his average expected utility is

$$\left(\frac{1}{T}\right)\left[(EN)\hat{v} + (T - EN)v^*\right]. \tag{5.14}$$

Since $\hat{v} \geqq v^*$, it follows that (5.14) is not less than

$$(1 - \delta)\hat{v} + \delta v^*, \tag{5.15}$$

which is at least $(\hat{v} - \epsilon)$ if

$$\delta \leqq \frac{\epsilon}{\hat{v} - v^*}. \tag{5.16}$$

If the agent uses some sequential strategy σ instead of σ_A against the principal's

strategy $\sigma_P((b_n))$, then his average expected utility is

$$(5.17) \qquad \left(\frac{1}{T}\right)\left[E\sum_{t=1}^{N} V(\hat{w}, A_t) + (T - EN)v^*\right],$$

where N is the stopping time under the agent's strategy σ. If the agent uses σ_A, his average expected utility is, by (5.15), at least

$$(5.18) \qquad \hat{v} + \delta(v^* - \hat{v}).$$

The *increment* in average expected utility to the agent from using σ instead of σ_A is therefore not more than the difference between (5.17) and (5.18), which can be written as

$$(5.19) \qquad \left(\frac{1}{T}\right)\left[EG_N + (T - EN)(v^* - \hat{v})\right] + \delta(\hat{v} - v^*).$$

Using Lemma 5.2, and recalling that $\hat{v} \geqq v^*$, we see that (5.19) is not greater than ϵ if, for example,

$$(5.22) \qquad \begin{aligned} &\delta \leqq \frac{\epsilon}{2(\hat{v} - v^*)}, \\ &K\left(\frac{b_T + M}{T}\right) \leqq \frac{\epsilon}{2}. \end{aligned}$$

Therefore, the proof of the theorem is completed by taking (i) δ to satisfy both (5.13) and (5.21), (ii) λ corresponding to δ as in Lemma 5.3, (iii) a sequence (b_t) in $\boldsymbol{B}$ corresponding to λ, and (iv) T_ϵ to satisfy (5.22); the last is of course possible because (b_T/T) approaches zero as T increases without limit.

6. ROBUST EPSILON EQUILIBRIA

The epsilon equilibria described in Section 5 have the property that, with "small" probability, at some period $t > 1$ the agent may be able to gain more than ϵ in average expected utility (conditional on his information up to that date) by departing from the equilibrium strategy. I therefore propose a more stringent equilibrium condition, which I shall now describe.

For every t, at the end of period t the agent knows the history H_t, whereas the principal knows only the random variables $C_1, \ldots, C_t$. Call this latter history H_{Pt}. Suppose that some strategy pair (σ_P, σ_A) has been used in periods 1 through t, with a history H_t. To the remaining $(T - t)$ periods corresponds a game in which the principal and agent have the initial information H_{Pt} and H_t, respectively, and in which the payoff to a player is $(1/T)$ times the total expected utility in all T periods. Call this the *continuation game, given t and* H_t. Among the strategies available to the principal in the continuation game is the *continuation of* σ_P, defined in the obvious way; a corresponding remark applies to the agent. A

player's strategy is a *robust epsilon-optimal response to the other player's strategy* if, for every t and almost every H_t, the continuation of the player's strategy is within epsilon (in payoff) of being an optimal response to the continuation of the other player's strategy, in the continuation game given t and H_t. A pair of strategies is a *robust epsilon equilibrium* if each player's strategy is a robust epsilon-optimal response to the other player's strategy. A robust epsilon equilibrium is, of course, an epsilon equilibrium.

Notice that, for every t, I have required the continuation of the principal's strategy to be epsilon-optimal for almost every history H_t, not just for every H_{Pt}, even though the principal will not know H_t completely.[11] Nevertheless, as I shall show in this section, there exist robust epsilon equilibria that are approximately as efficient as the (nonrobust) epsilon equilibria described in Section 5.

I should emphasize, too, that in each continuation game each player calculates his average expected utility per period over *all* periods (1 through T). Equivalently, his payoff in the continuation game may be taken to be $(1/T)$ times the total expected utility in the remaining $(T-t)$ periods, since the past history is given. From the behavioral point of view, it might be more attractive to take this payoff to be the average expected utility in the remaining periods, i.e., to divide the total expected utility by $(T-t)$. In fact, the results of the present section can be extended to cover the corresponding alternative definition, but at the cost of more complex calculations. Therefore, to simplify the exposition (which is in any case somewhat complex), I shall use the first definition in the present section, and briefly discuss the implications of the second definition in the next section.

One reason that the epsilon equilibria may not be robust is that when the cumulative sum of consequences, S_t, is sufficiently far above the "boundary" $t\hat{c} - b_t$, the agent has an opportunity to "loaf" and still keep the probability of reaching the boundary acceptably low. Therefore, in a robust equilibrium, the principal must modify his strategy so as to discourage the agent from taking advantage of this situation. One simple way to do this is for the principal to impose an *upper* boundary, as well as a lower one, to the region in which he maintains the constant reward $\hat{w}$. Thus I shall demonstrate the existence of robust epsilon equilibria in which the principal switches to the reward function ω^* after the first t such that S_t crosses either the lower or the upper boundary.

A second cause of nonrobustness is that, the closer S_t is to the boundary of the region in which the reward $\hat{w}$ is maintained, the greater is the agent's incentive to take some action to avoid the boundary. For a lower boundary, this means increasing his effort; for an (additional) upper boundary, this means decreasing (!) his effort. In the robust epsilon equilibria described in this section, the agent will use the decision rule $\hat{\alpha}$ until he reaches the boundary of a region that is smaller than the principal's region for $\hat{w}$, and thereafter he will optimize sequentially against the principal's strategy. In particular, the agent will switch to the

[11] The definition of robust epsilon equilibrium bears a superficial resemblance to that of subgame-perfect equilibrium (see Selten [**14**]), but the concepts are significantly different. In fact, the repeated principal-agent game discussed here has no subgames other than the entire game (as the term "subgame" is used). Hence every Nash equilibrium is subgame-perfect.

one-period Nash equilibrium decision rule α^* as soon as the principal switches to the reward function ω^*.

It may appear inefficient for the principal to switch to ω^* when S_t reaches his *upper* boundary, since this "punishes" both players for the agent's good luck or increased effort. Indeed, Pareto-superior strategy-pairs can be derived; this will be discussed in Section 7.

Before giving a precise statement of the main result of this section, I must introduce some new notation. With the sequence (b_t^0) defined by (5.8), let $\boldsymbol{B}$ denote the set of sequences of the form (λb_t^0) such that $\lambda > 1$. (Warning: this set $\boldsymbol{B}$ is smaller than the corresponding set denoted by $\boldsymbol{B}$ in Section 5.) By an "extended integer" I shall mean a positive integer or $+\infty$. For any extended integers r and s, and any set R of extended integers, define

$$\wedge R \equiv \min\{x : x \text{ in } R\},$$

$$r \wedge s \equiv \wedge\{r,s\},$$

$$r \wedge R \equiv r \wedge (\wedge R).$$

Let $(b_t) = (\lambda b_t^0)$ be a sequence in $\boldsymbol{B}$, and, for this sequence and a given T, define

(6.1a) $$N \equiv T \wedge \{t : |S_t - t\hat{c}| \geqq b_t\},$$

(6.1b) $$M \equiv T \wedge \{t|S_t - t\hat{c}| \geqq b_t/2\},$$

(6.1c) $$D_P \equiv \wedge\{t : \omega_t \neq \hat{w}\},$$

(6.1d) $$D_A \equiv (N+1) \wedge D_P.$$

I shall say that the agent *optimizes myopically* in period t if he uses a one-period optimal decision rule in response to the reward function ω_t.

Given the sequence $(b_t) = (\lambda b_t^0)$ in $\boldsymbol{B}$, the strategy $\sigma_P(\lambda)$ for the principal is defined by

(6.2) $$\omega_t = \begin{cases} \hat{\omega}, & t = 1, \ldots, N, \\ \omega^*, & t > N. \end{cases}$$

The strategy $\sigma_A(\lambda)$ for the agent is defined by:

(6.3a) $\alpha_t = \hat{\alpha}, \quad t = 1, \ldots, M \wedge (D_A - 1).$

(6.3b) The agent optimizes myopically in all periods $t \geqq D_A$.

(6.3c) If $M < D_A - 2$, then the agent uses a sequential strategy in periods $(M+1)$ through $(D_A - 1)$ that is (sequentially) optimal against $\sigma_P(\lambda)$, given (6.3b).

Notice that $\sigma_A(\lambda)$ is sequentially optimal from period $(M+1)$ on, given $\sigma_P(\lambda)$.

The main result of this section is the following theorem. The situation and notation are as described in Section 5, except as noted above.

THEOREM: *For any $\epsilon > 0$ there is a sequence (λb_t^0) in $\mathbf{B}$ and a T_ϵ such that for all $T \geqq T_\epsilon$ the pair $[\sigma_P(\lambda), \sigma_A(\lambda)]$ is a robust ϵ equilibrium, and yields the principal and the agent expected average utilities per period at least $(\hat{u} - \epsilon)$ and $(\hat{v} - \epsilon)$, respectively.*

The proof of the theorem uses arguments similar to those used in Section 5, so I shall omit some of the details. Let H_t denote history of the random variables $Z_1, \ldots, Z_t$, and for any random variable X let $E_t X$ denote the conditional expectation $E\{X \mid H_t\}$. A *random time* is a random variable X such that, for every extended integer t the event $\{X = t\}$ is measurable[12] with respect to H_t. The proof of the theorem makes use of a "comparison path," in which $(\hat{w}, \hat{\alpha})$ is used until some random time, $\hat{D}$, and then (ω^*, α^*) is used thereafter. (This is, in fact, the kind of path generated by the equilibria of Section 5.) The first lemma characterizes the difference in total expected utility to the principal, after a given period t, in two situations: (i) the principal pays the reward $\hat{w}$, and the agent uses some arbitrary strategy, through some random time D, and they then use the pair (ω^*, α^*) thereafter; (ii) the comparison path.

Define:

$$
\begin{aligned}
&A_t^* = \alpha^*(Z_t), \quad \hat{A}_t = \hat{\alpha}(Z_t), \\
&C_t^* = \gamma(A_t^*, Z_t), \quad \hat{C}_t = \gamma(\hat{A}_t, Z_t), \\
&W_t^* = (C_t^*), \quad \hat{W}_t = \hat{w}, \\
&u^* = E(C_t^* - W_t^*), \quad \hat{u} = E(\hat{C}_t - \hat{w}) = \hat{c} - \hat{w}.
\end{aligned} \tag{6.4}
$$

LEMMA 6.1: *If D and $\hat{D}$ are random times bounded above by T, then for every pair of strategies, and almost every history H_t such that $D > t$ and $\hat{D} > t$,*

$$
\begin{aligned}
&E_t\left\{\sum_{n=t+1}^{D}(C_n - \hat{w}) + \sum_{n=D+1}^{T}(C_n^* - W_n^*)\right\} \\
&\quad - E_t\left\{\sum_{n=t+1}^{\hat{D}}(\hat{C}_n - \hat{w}) + \sum_{n-\hat{D}+1}^{T}(C_n^* - W_n^*)\right\} \\
&\quad = E_t\left\{\sum_{n=t+1}^{D}(C_n - \hat{C}_n) + (\hat{u} - u^*)(D - \hat{D})\right\}.
\end{aligned}
$$

[12] Strictly speaking, measurable with respect to the sigma-field of the underlying probability space that is induced by H_t.

PROOF: By the systems theorem for martingales,

$$E_t \sum_{n=t+1}^{D} (\hat{C}_n - \hat{w}) = \hat{u} E_t(D - t), \tag{6.5}$$

$$E_t \sum_{n=t+1}^{\hat{D}} (\hat{C}_n - \hat{w}) = \hat{u} E_t(\hat{D} - t); \tag{6.6}$$

and by the strong Markov property,

$$E_t \sum_{n=D+1}^{T} (C_n^* - W_n^*) = u^* E_t(T - D), \tag{6.7}$$

$$E_t \sum_{n=\hat{D}+1}^{T} (C_n^* - W_n^*) = u^* E_t(T - \hat{D}). \tag{6.8}$$

The conclusion of the lemma now follows with a straightforward calculation.

The next lemma puts an upper bound on the principal's expected loss after any period t, if the principal uses the strategy $\sigma_P(\lambda)$, and the agent uses—instead of the strategy $\sigma_A(\lambda)$—*any* sequential strategy such that he optimizes myopically after some random time $D \leqq N$.

LEMMA 6.2: *Suppose that the principal uses the strategy* $\sigma_P(\lambda)$. *For any sequential strategy of the agent define* N *by* (6.1a), *let* D *be any random time* $\leqq N$, *and define*

$$L(t, D) = \sum_{n=t+1}^{D} (\hat{C}_n - C_n), \tag{6.9}$$

provided $t < D$; *then for almost every history* H_t *such that* $D > t$,

$$|E_t\{L(t, D) - (S_t - t\hat{c})\}| \leqq E_t b_N + B. \tag{6.10}$$

The proof of this lemma is similar to that of Lemma 5.1, and is omitted.

LEMMA 6.3: *For any* $\lambda > \lambda_0$ *and any* $\epsilon > 0$ *there is a* $T_P(\epsilon, \lambda)$ *such that, for all* $T \geqq T_P(\epsilon, \lambda)$, $\sigma_P(\lambda)$ *is a robust* ϵ*-optimal response to* $\sigma_A(\lambda)$.

PROOF: First note that if the principal uses a reward function different from $\hat{w}$ at t, then the agent will optimize myopically from period t on, so it will be optimal for the principal to use ω^* from t on. Let σ_P' be a sequential strategy with this property, i.e. the principal uses $\hat{w}$ through some random time D, and then uses ω^*. Since the agent will optimize myopically after N, an optimal strategy for the principal must have $D \leqq N$, so assume this property for σ_P'.

For $t \leqq D$, the total conditional expected utility to the principal in periods $(t + 1), \ldots, T$, if he uses σ_P' from period $(t + 1)$ on, conditional on H_t (*not* H_{P_t}),

is

$$(6.11)\qquad E_t\left\{\sum_{n=t+1}^{D}(C_n-\hat{w})+\sum_{n=D+1}^{T}(C_n^*-W_n^*)\right\}.$$

On the other hand, if he uses the strategy $\sigma_P(\lambda)$ from $(t+1)$ on, the corresponding expected utility is

$$(6.12)\qquad E_t\left\{\sum_{n=t+1}^{N}(C_n-\hat{w})+\sum_{n=N+1}^{T}(C_n^*-W_n^*)\right\}.$$

Finally, the corresponding expected utility for the "comparison path," with $\hat{D}=N$, is

$$(6.13)\qquad E_t\left\{\sum_{n=t+1}^{N}(\hat{C}_n-\hat{w})+\sum_{n=N+1}^{T}(C_n^*-W_n^*)\right\}.$$

By Lemma 6.1, (6.11)–(6.13) equals

$$(6.14)\qquad E_t\left\{\sum_{n=t+1}^{D}(C_n-\hat{C}_n)+(\hat{u}-u^*)(D-N)\right\}.$$

Also, (6.12)–(6.13) equals

$$(6.15)\qquad E_t\left\{\sum_{n=t+1}^{N}(C_n-\hat{C}_n)\right\}.$$

Applying Lemma 6.2 twice, one immediately concludes that (6.11)–(6.12) is not greater than

$$(6.16)\qquad 2\left[E_t b_N+B\right]+(\hat{u}-u^*)E_t(D-N).$$

Recall that $N\leqq T$, (b_n) is increasing in n, $\hat{u}\geqq u^*$, and $D\leqq N$. Hence (6.16) is not greater than

$$(6.17)\qquad 2(b_T+B).$$

To complete the proof of the lemma it suffices to take T large enough so that

$$(6.18)\qquad \frac{2(b_T+B)}{T}\leqq\epsilon.$$

(Note that, in (6.18), $b_T=\lambda b_T^0$, so that the critical value of T depends both on λ and on ϵ.)

LEMMA 6.4: *For any $\epsilon>0$ there exists $\lambda_\epsilon>0$ and $T_A(\epsilon)$ such that, for all $T\geqq T_A(\epsilon)$, $\sigma_A(\lambda_\epsilon)$ is a robust ϵ-optimal response to $\sigma_P(\lambda_\epsilon)$.*

PROOF: For all $t > M$, the continuation of strategy $\sigma_A(\lambda)$ is (strictly) optimal against the continuation of $\sigma_P(\lambda)$, conditional on H_t. Therefore it remains to consider the case in which $t \leqq M$. Let $\lambda > \lambda_0$ be given, and $(b_t) = (\lambda b_t^0)$; suppose that the principal uses $\sigma_P(\lambda)$ in all periods, and that the agent has used $\sigma_A(\lambda)$ through period t.

Recall the definition of the "comparison path" $(\hat{C}_t)$ as in (6.4), and define

$$\begin{aligned} &\hat{S}_t = \hat{C}_1 + \cdots + \hat{C}_t, \\ &\hat{N} \equiv T \wedge \{t: |\hat{S}_t - t\hat{c}) \geqq b_t\}. \end{aligned} \tag{6.19a}$$

Consider the policy $\hat{\sigma}_A(\lambda)$ for the agent defined by

$$\alpha_t = \begin{cases} \hat{\alpha}, & \text{for } t \leqq \hat{N}, \\ \alpha^* & \text{for } t > \hat{N}. \end{cases} \tag{6.19b}$$

The strategy $\hat{\sigma}_A(\lambda)$ generates the comparison path (given $\sigma_P(\lambda)$), whereas the strategy $\sigma_A(\lambda)$ generates the comparison path for $t \leqq M$, and then optimizes sequentially thereafter. Hence, for all t, $\sigma_A(\lambda)$ is at least as good as $\hat{\sigma}_A(\lambda)$, conditional on H_t, and so it suffices to show that, for $t \leqq M$, the strategy $\hat{\sigma}_A(\lambda)$ is within ϵ of being an optimal response to $\sigma_P(\lambda)$, conditional on H_t (for suitably chosen λ and T).

Let $\tilde{\sigma}_A$ be any continuation strategy for the agent, and for $n > t$ define the corresponding random variables $\tilde{A}_n$, $\tilde{C}_n$, and $\tilde{W}_n$, as in (6.4). Define

$$\begin{aligned} &\tilde{S}_n = \hat{S}_t + \hat{C}_{t+1} + \cdots \tilde{C}_n, \\ &\tilde{N} = T \wedge \{n : |\tilde{S}_n - n\hat{c}| \geqq b_n\}. \end{aligned} \tag{6.20}$$

Suppose that $\tilde{\sigma}_A$ has the property that the agent optimizes myopically (and hence sequentially) against $\sigma_P(\lambda)$ for $t > N$, so that $\alpha_t = \alpha^*$ for $t > N$. For $t \leqq M$ and any history H_t, the conditional expected total utility to the agent in periods $(t+1)$ through T, given H_t, from the strategy $\tilde{\sigma}_A$ is

$$E_t\left\{\sum_{n=t+1}^{\tilde{N}} V(\hat{w}, \tilde{A}_n) + \sum_{n=\tilde{N}+1}^{T} V(W_n^*, A_n^*)\right\}. \tag{6.21}$$

The corresponding conditional expected utility from the strategy $\hat{\sigma}_A(\lambda)$ is

$$E_t\left\{\sum_{n=t+1}^{\hat{N}} V(\hat{w}, \hat{A}_n) + \sum_{n=\hat{N}+1}^{T} V(W_n^*, A_n^*)\right\}. \tag{6.22}$$

By an argument similar to that of Lemma 6.1, the difference between (6.21) and (6.22) is

$$E_t\left\{\sum_{n=t+1}^{\tilde{N}} \left[V(\hat{w}, \tilde{A}_n) - V(\hat{w}, \tilde{A}_n)\right] + (\hat{v} - v^*)(\tilde{N} - \hat{N})\right\}. \tag{6.23}$$

Consider first the first part of (6.23), namely,

$$G(t,\tilde{N}) = E_t \sum_{n=t+1}^{\tilde{N}} \left[V(\hat{w},\tilde{A}_n) - V(\hat{w},\hat{A}_n) \right]. \tag{6.24}$$

By the argument used in Lemma 5.2, there is a number $K > 0$ such that

$$G(t,\tilde{N}) \leqq KE_t \sum_{n=t+1}^{\tilde{N}} (\hat{C}_n - \tilde{C}_n). \tag{6.25}$$

By an argument like that used in Lemma 5.1,

$$\left| E_t \sum_{n=t+1}^{\tilde{N}} (\hat{C}_n - \tilde{C}_n) - (S_t - t\hat{c}) \right| \leqq E_t b_{\tilde{N}} + B \tag{6.26}$$

$$\leqq b_T + B.$$

Remember that $S_t = \hat{S}_t$, since $t \leqq M$ and $\sigma_A(\lambda)$ and $\hat{\sigma}_A(\lambda)$ agree through period M. By the definition of M (see (6.1b)),

$$|S_t - t\hat{c}| < b_t/2,$$

so that, from (6.25) and (6.26),

$$E_t G(t,\tilde{N}) \leqq K\left(\frac{3b_T}{2} + B \right). \tag{6.27}$$

Consider now the second part of (6.23), namely

$$(\hat{v} - v^*)E_t(\tilde{N} - \hat{N}). \tag{6.28}$$

Observe that

$$E_t(\tilde{N} - \hat{N}) \leqq E_t(T - \hat{N}) = T - E_t\hat{N},$$

and

$$E_t\hat{N} \geqq 0 + T\,\text{prob}\{\hat{N} = T \mid H_t\},$$

so that

$$E_t(\tilde{N} - \hat{N}) \leqq T\,\text{prob}\{\hat{N} < T \mid H_t\}. \tag{6.29}$$

One can verify that for every m and every $t \leqq m$,

$$b_{m-t} + b_t \leqq 2b_{m/2}.$$

Hence

$$\left(\frac{1}{2}\right)b_{m-t} \leqq b_{m/2} - \left(\frac{1}{2}\right)b_t$$

$$\leqq b_m - \left(\frac{1}{2}\right)b_t.$$

But, given $\hat{N} > t$, $\hat{N} < T$ if and only if there is an m, with $t < m < T$, such that

$$|\hat{S}_m - m\hat{c}| \geqq b_m,$$

which, since $t \leqq M$, implies that

$$|(\hat{S}_m - \hat{S}_t) - (m-t)\hat{c}| \geqq b_m - \left(\frac{1}{2}\right)b_t \geqq \left(\frac{1}{2}\right)b_{m-t}. \tag{6.30}$$

By an argument like that leading up to Lemma 5.3, one can show that, for every $\delta > 0$ there is a $\lambda(\delta) > \lambda_0$ such that

$$\text{prob}\left\{|(\hat{S}_m - \hat{S}_t) - (m-t)\hat{c}| < \left(\frac{1}{2}\right)b_{m-t}, \text{ all } m > t \,|\, H_t\right\} > 1 - \delta, \tag{6.31}$$

uniformly in all t for which $M \geqq t$, and where $b_{m-t} = \lambda(\delta)b^0_{m-t}$. Hence

$$\text{prob}\{\hat{N} < T \,|\, H_t\} \leqq \delta,$$

so that

$$E_t(\tilde{N} - \hat{N}) \leqq T\delta. \tag{6.32}$$

In summary, (6.24), (6.27), and (6.32) imply that, for any $\delta > 0$, $\lambda > \lambda(\delta)$, and $t \leqq M$, and almost every H_t, the expression (6.23) is not greater than

$$K\left[\left(\frac{3}{2}\right)b_T + B\right] + (\hat{v} - v^*)T\delta. \tag{6.33}$$

Hence, to complete the proof of the lemma, it suffices to take

$$\delta < \frac{\epsilon}{2(\hat{v} - v^*)}, \quad \lambda > \lambda(\delta); \tag{6.34}$$

$$T_A(\epsilon) \text{ large enough so that } \frac{K}{T}\left[\left(\frac{3}{2}\right)\lambda(\delta)b^0_T + B\right] < \frac{\epsilon}{2} \text{ for } T \geqq T_A(\epsilon). \tag{6.35}$$

To complete the proof of the theorem, it remains to establish lower bounds on the expected average utilities of the two players, with the strategy pair $[\sigma_P(\lambda, \sigma_A(\lambda)]$.

The expected average utility to the principal is

$$\bar{u} \equiv \frac{1}{T}\left[\hat{u}E\{M\} + E\left\{\sum_{n=M+1}^{N}(C_n - \hat{w})\right\} + u^*E(T - N)\right]. \tag{6.36}$$

Recall that the variables $|C_t|$ are bounded by B, and define

$$B' = \max\{B + \hat{c}, u - u^*\}. \tag{6.37}$$

From (6.36) and (6.37),

$$\begin{aligned}\bar{u} - \hat{u} &\geqq -\left(\frac{B'}{T}\right)E(T - M) \\ &\geqq -B'\,\text{prob}\{M < T\}.\end{aligned} \tag{6.38}$$

By taking $t = 0$ in (6.31), one sees that

(6.39a) $\quad \text{prob}\{M < \infty\} < \frac{\epsilon}{B'}$

if

(6.39b) $\quad \lambda > \lambda(\epsilon/B')$.

Hence, if (6.39b) is satisfied, then $\bar{u} \geqq \hat{u} - \epsilon$, *for all* T.

The expected average utility, $\bar{v}$, to the agent is at least as large as the expected average utility for the comparison path, so that

$$\begin{aligned}(6.40) \quad \bar{v} &= \left(\frac{1}{T}\right)\left[\hat{v}E(\hat{N}) + v^*E(T - \hat{N})\right] \\ &= \hat{v} - (\hat{v} - v^*)E\left(-\frac{N}{T}\right) \\ &\geqq \hat{v} - (\hat{v} - v^*)\text{prob}\{N < T\}.\end{aligned}$$

By a variation on (6.39ab), one sees that

(6.41a) $\quad \text{prob}\{\hat{N} < \infty\} < \frac{\epsilon}{\hat{v} - v^*}$

if

(6.41b) $\quad \lambda > \lambda(\epsilon/(\hat{v} - v^*))$.

Hence, if (6.41b) is satisfied, then $\bar{v} \geqq \hat{v}$ for all T.

I am now in a position to complete the proof of the theorem by summarizing all of the sufficient conditions on λ and T. For any positive number δ let $\lambda(\delta)$ be a number $\lambda > 1$ such that

(6.42) $\quad \text{prob}\{|\hat{S}_t - t\hat{c}| < \lambda b_t^0, \text{ for all } t\} > 1 - \delta$;

further choose the function λ to be decreasing.

From (6.34), (6.37), (6.39b), and (6.41b) one sees that it is sufficient to take

$$\begin{aligned}&\lambda_\epsilon > \lambda\left(\frac{\epsilon}{B''}\right), \quad \text{where} \\ (6.43) \quad &B'' \equiv \max\{B + \hat{c}, \hat{u} - u^*, 2(\hat{v} - v^*)\};\end{aligned}$$

and from (6.18) and (6.35) one sees that it is sufficient to take T_ϵ to be the smallest T such that

$$\frac{\lambda_\epsilon b_T^0 + B}{T} \leqq \frac{\epsilon}{2},$$

$$\left(\frac{K}{T}\right)\left[\left(\frac{3}{2}\right)\lambda_\epsilon b_T^0 + B\right] \leqq \frac{\epsilon}{2}.$$

This completes the proof of the theorem.

7. EXTENSIONS

Recall that in the definition of the continuation of the game after a date t (Section 6), the payoff to a player was taken to be $(1/T)$ times the total expected utility in the remaining $(T-t)$ periods. This is equivalent to taking his payoff to be the average expected utility per period over all T periods, since in the continuation of the game after date t, the actions and consequences through date t have already been determined. As was pointed out in Section 6, from a behavioral point of view it might be more attractive to take the payoff in the continuation game to be the average expected utility per period in the remaining $(T-t)$ periods, i.e., to divide the total remaining utility by $(T-t)$. Using such a definition, one can show that the robust equilibria described in Section 6 need to be modified so that cooperation breaks down near the end of the game. For example, for each epsilon one can find a number k such that each player will switch to the one-period Nash equilibrium strategy after date $(T-k)$; the number k may be taken to be independent of T. For many, such behavior would be intuitively more plausible than the equilibria of Sections 5 and 6. (For similar results in the case of certainty, see Radner [**8**].)

In Section 6 attention was called to the apparent "inefficiency" of the equilibrium strategies described there. The source of the inefficiency was the property of the principal's strategy in which he switched to the one-period Nash reward function as soon as the agent's cumulated performance reached an upper boundary. To improve the efficiency of the equilibrium, one can use a number of devices. For example, the principal can translate the two boundaries upward by some prescribed amount whenever the upper boundary is reached. In addition, the switch by the two players to the one-period Nash equilibrium after the lower boundary is reached need not last until the end of the game, but only long enough to deter the agent from cheating. I should emphasize, however, that strategies modified in these two ways, *as well as the strategies described in the theorem of Section 6*, differ from the equilibrium strategies of Section 5 only on histories of "small probability."

I note here that the results of Sections 5 and 6 can be extended, in an appropriate form, to cases with more general classes of utility functions for the principal and agent than the classes considered here. In particular, the principal need not be neutral towards risk; however in this latter case one would not expect a Pareto-optimal arrangement to be characterized by a constant reward.

One can use similar techniques to demonstrate the existence of exact Nash equilibria in the infinite supergame that exactly achieve Pareto-optimal long-run average expected utility pairs. Also, one can extend these methods to situations with more than two players (see [**9**]).

Finally, I note that, if the principal and agent discount future utilities, an equilibrium of the infinite supergame typically cannot be efficient; however, such equilibria can be close to efficient if the discount rates are small (see [**10**]).

Bell Laboratories, Murray Hill, N.J.

Manuscript received September, 1979; revision received May, 1980.

REFERENCES

[1] AUMANN, R. J.: "Acceptable Points in General Cooperative N-Person Games," in *Contributions to the Theory of Games*, Vol. IV, ed. by A. W. Tucker and R. D. Luce. Princeton: Princeton University Press, 1959, pp. 287–324.

[2] CHUNG, K. L.: *A Course in Probability Theory*, Second Edition. New York: Academic Press, 1974.

[3] GROVES, T.: "Incentives in Teams," *Econometrica*, 41(1973), 617–631.

[4] HOLMSTRÖM, B.: "Moral Hazard and Observability," *Bell Journal of Economics*, 10(1979), 74–91.

[5] HURWICZ, L., AND L. SHAPIRO: "Incentive Structures Maximizing Residual Gain Under Incomplete Information," *Bell Journal of Economics*, 9(1978), 180–191.

[6] KURZ, M.: "Altruism as an Outcome of Social Interaction," *American Economic Review*, 68(1978), 216–222.

[7] RADNER, R.: "Can Bounded Rationality Resolve the Prisoner's Dilemma?" Bell Telephone Laboratories Discussion Paper, Murray Hill, New Jersey, March, 1979.

[8] ———: "Collusive Behavior in Noncooperative Epsilon-Equilibria of Oligopolies with Long but Finite Lives," *Journal of Economic Theory*, 22(1980), 136–154.

[9] ———: "Optimal Equilibria in a Class of Repeated Games with Imperfect Monitoring," Bell Telephone Laboratories Discussion Paper, Murray Hill, New Jersey, May, 1981.

[10] ———: "A Repeated Principal-Agent Game with Discounting," Bell Telephone Laboratories Discussion Paper, Murray Hill, New Jersey, May, 1981.

[11] ROBBINS, H., AND D. SIEGMUND: "The Expected Sample Size of Some Tests of Power One," *Annals of Statistics*, 2(1974), 415–436.

[12] RUBINSTEIN, A.: "Equilibrium in Supergames," Center Res. Math. Econ. Game Theory, Research Memorandum No. 25, Hebrew University, May, 1977.

[13] ———: "Offenses that May Have Been Committed by Accident—An Optimal Policy of Retribution," in *Applied Game Theory*, ed. by S. J. Brams, A. Schotter, and G. Schrödiauer. Würzburg: Physica-Verlag, 1979.

[14] SELTEN, R.: "Re-examination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory*, 4(1975), 25–55.

[15] SHAVELL, S.: "Risk-Sharing and Incentives in the Principal and Agent Relationship," *Bell Journal of Economics*, 10(1979), 55–73.

[16] SIMON, H.: "A Formal Theory of the Employment Relation," *Econometrica*, 19(1951), 293–305.