Complexity and Efficiency in the Negotiation Game

Jihong Lee
Birkbeck College, London and Trinity College, Cambridge[*]

Hamid Sabourian
King's College, Cambridge[†]

September 2003

Abstract

This paper considers the "negotiation game" (Busch and Wen [4]) which combines the features of two-person alternating offers bargaining and repeated games. Despite the forces of bargaining, the negotiation game in general admits a large number of equilibria some of which involve delay and inefficiency. In order to isolate equilibria in this game, we investigate the role of complexity of implementing a strategy, introduced in the literature on repeated games played by automata. It turns out that when the players care for less complex strategies (at the margin) only *efficient* equilibria survive. Thus, complexity and bargaining in tandem may offer an explanation for co-operation and efficiency in repeated games.

JEL Classification: C72, C78

Keywords: Negotiation Game, Repeated Game, Bargaining, Complexity, Bounded Rationality, Automaton

1 Introduction

Busch and Wen [4], henceforth referred to as BW, analyze the following game. In each period, two players bargain - in Rubinstein's alternating-offers protocol -

[*] School of Economics, Mathematics and Statistics, Birkbeck College, Malet St, London, WC1E 7HX, United Kingdom (email: J.Lee@econ.bbk.ac.uk)

[†] Faculty of Economics and Politics, Sidgwick Avenue, Cambridge, CB3 9DD, United Kingdom (email: Hamid.Sabourian@econ.cam.ac.uk)

over the distribution of a fixed and commonly known periodic surplus. If an offer is accepted, the game ends and each player gets his share of the surplus according to the agreement at every period thereafter. After any rejection, but before the game moves to the next period, the players engage in a normal form game to determine their payoffs for the period. The Pareto frontier of the disagreement game is contained in the bargaining frontier. We shall refer to this game as the *negotiation game*.

The negotiation game generally admits a large number of subgame-perfect equilibria, as summarized by BW in a result that has a same flavour as the Folk theorem in repeated games. The structure of the disagreement game determines what can be sustained as credible threats in the negotiation game and thus shapes the lowest possible subgame-perfect equilibrium (SPE) payoff for each player. BW then show that, provided the players are sufficiently patient, any payoff vector consistent with these payoffs can be supported as a SPE outcome in the negotiation game. Moreover, one can construct a pair of equilibrium strategies that generate any length of delay in reaching an agreement as well as a sequence of inefficient actions taken after disagreements. The negotiation game has a unique (efficient) equilibrium only in the degenerate case in which any Nash equilibrium payoff of the disagreement game coincides with its minmax point.

The negotiation game and its equilibria can be interpreted from two alternative viewpoints. Naturally, we can think of the game as a standard alternating-offers bargaining game with endogenous disagreement payoffs.[1] In fact, Fernandez and Glazer [7] (and also Haller and Holden [11]) derive much of the insights in a well-known application of the game along this bargaining interpretation. They consider the standoff between a union and a firm. During a contract renewal process, a union and a firm renegotiate over the distribution of a periodic revenue, but a disagreement puts them in a strategic situation. The union can either accept the firm's wage offer or forego the status quo wage for one period and strike before making a counter-offer next period. (The firm is inactive in the disagreement game.) Fernandez and Glazer's characterization of subgame-perfect equilibria in this specific setting contains many of the salient features of the equilibria in the general game, and thus, offers an explanation as to why such socially wasteful activities as strikes may take place even in a situation where the agents are completely rational and fully informed.

The alternative viewpoint focuses on the repeated game aspect of the negotiation game (and this is the interpretation we want to emphasize in the paper). Real world repeated interactions are often accompanied by negotiations which

[1] The issue of endogenous disagreement payoffs in a bargaining situation goes back at least to Nash [14] who considers the problem in a co-operative framework.

can lead to mutual agreement. While equilibria in standard repeated games are usually given the interpretation of implicit, self-enforcing agreements, the situations depicted by the negotiation game are associated with explicit contracts that can bind the players to a particular set of outcomes. For example, we observe firms engaged in a repeated horizontal or vertical relationship negotiating over a long-term contract, or even a merger. Similarly, countries involved in international trade often attempt to settle an agreement that enforces fixed quotas and tariffs.

The Folk theorem gives economic theorists little hope of making any predictions in repeated interactions. However, as the aforementioned examples suggest, it seems that negotiation is often a salient feature of real world repeated interactions, presumably to enforce co-operation and efficient outcomes. Can bargaining be used to isolate equilibria in repeated games? Unfortunately, the contributions of BW and others demonstrate that Folk theorem type results with a large number of equilibria which involve delay and inefficiency may persist even when the players are endowed with an opportunity at the beginning of each period to settle on the efficient outcome once and for all.

In order to enrich this line of enquiry, on the issue of how bargaining can be used to select (efficient) equilibria in repeated games, this paper departs from the standard rationality paradigm and introduces the notion of *complexity* into the negotiation game. Our central message is that the equilibrium strategies supporting inefficient outcomes in this game are unnecessarily too complex to implement. Bargaining combined with the players' preference for less complex strategies (at the margin) select only *efficient* outcomes in the repeated game.

There are many different ways of defining the complexity of a strategy. In the literature on repeated games played by automata the number of states of the machine is often used as a measure of complexity (Rubinstein [20], Abreu and Rubinstein [1], Piccione [17] and Piccione and Rubinstein [18]). This is because the set of states of the machine can be regarded as a partition of possible histories. In particular, Kalai and Stanford [13] show that the counting-states measure of complexity, henceforth referred to as *state complexity*, is equivalent to looking at at the number of *continuation strategies* that the strategy induces at different histories of the game. We extend this notion of strategic complexity to the negotiation game, and facilitate the analysis by considering an equivalent "machine game".

The alternating-offers bargaining imposes an asymmetric structure on the negotiation game which is stationary only every two periods (henceforth we shall refer to every two periods as a "stage"). To account for such structural asymmetry of the game, we shall adopt machine specifications that formally distinguish between the different *roles* played by each player in a given stage. A player can be either proposer or responder. In the main machine specification used in the analysis, there are two “sub-machines”, each playing a role (of a proposer or a responder) with distinct states, output and transition functions. Transition occurs at the end of each period, from a state belonging to one sub-machine to a state belonging to the other sub-machine as roles are reversed.

We first demonstrate that the result of Kalai and Stanford [13] holds for our specification of machines. The total number of states used by each sub-machine under this specification is equivalent to measuring the total number of continuation strategies that the implemented strategy induces at the beginning of each period.

The concept of Nash equilibrium is then refined to incorporate the players' preference for less complex strategies. In our choice of equilibrium notions, complexity enters a player's preferences, together with the payoffs in the underlying game, either lexicographically or as a positive fixed cost c. The larger this cost is, the more is required of complexity. We can thus interpret it as a measure of the players' “bounded rationality”. We will refer to a Nash equilibrium (of the machine game) with fixed complexity cost c by NEMc and adopt the convention of using $c = 0$ (and thus NEM0) to refer to the lexicographic case. We also invoke the notion of subgame-perfection and consider the set of NEMc that are subgame-perfect, referred to as SPEMc.

The selection result is as follows. We first show that, independently of the degree of complexity cost and discount factor, if an agreement occurs in some finite period as the outcome of some NEMc then it must occur within the very first stage of the game, and moreover, the players' equilibrium strategies must be stationary (history-independent). Since any stationary subgame-perfect equilibrium in the negotiation game is efficient (see BW), it then follows that the set of SPEMc inducing an agreement must be efficient.

We then consider the other possible outcome, one in which there is perpetual disagreement. Here the following set of results are shown for a discount factor arbitrarily close to one. We first show that, given any non-negative complexity cost, every SPEMc involving perpetual disagreement is at least *long-run* (almost) efficient; that is, the players must reach a finite period in which the continuation game then on is (almost) efficient. It follows that, in cases where all disagreement game outcomes are inefficient, delay cannot persist indefinitely under any SPEMc. In fact, if we assume a strictly positive complexity cost, then we also derive that perpetual disagreement is not consistent with SPEMc however small that complexity cost is (even when agreement only weakly dominates disagreement). Combined with the previous set of results on agreement, this implies a very strong prediction for the case in which players are sufficiently patient. For any $c > 0$ (or if $c = 0$ and agreement strictly dominates disagreement), every SPEMc of the negotiation game must be efficient such that an agreement is reached in the first stage and the associated strategies are stationary.

We also explore an alternative machine specification that employs more frequent transitions and hence account for finer partitions of histories and continuation strategies. This machine consists of four sub-machines; while keeping the role distinction, transition occurs twice in each period at the end of bargaining and at the end of the disagreement game. We obtain sharper results in this case. The results on perpetual disagreement do not depend on the discount factor.

Our contribution thus takes the study of complexity in repeated games a step further from the aforementioned literature in which complexity has yielded only a limited selective power. (See also Bloise [3] who shows robust examples of two-player repeated games in which the set of Nash equilibria with complexity costs coincides with the set of individually rational payoffs.) This paper demonstrates that complexity and bargaining in tandem may offer an explanation for co-operation and efficiency in repeated games.

There have been extensive and wide-ranging approaches at restricting the unwieldily large set of equilibria resulting from the Folk theorem. Among these attempts, one literature motivates the notion of bargaining and negotiation by invoking the idea that punishments that are inefficient may be vulnerable to renegotiation and hence not credible. This literature suggests a solution concept based on *renegotiation-proofness*.[2] This line of research takes a "black box" approach to renegotiation. Unlike in the negotiation game, the process of (re)negotiation is not explicitly modelled; rather, the renegotiation arguments are embedded in the additional restrictions imposed on an equilibrium.

We also want to mention several recent papers that have rekindled the issue of complexity in equilibrium selection, and in particular, demonstrated that complexity drives efficient outcomes in some specific games. Chatterjee and Sabourian [5][6] consider the multi-person Rubinstein bargaining game, and Sabourian [21], Gale and Sabourian [9][10] consider market games with matching and bargaining. (These papers are also interested in other issues such as the uniqueness of the equilibrium set and the competitive nature of equilibria in the case of the market games.) In contrast to the present paper, however, these papers build upon a different notion of strategic complexity. They consider the complexity of response rules *within* a period. A simple response rule according to their notion of *response complexity* uses only the information available in the current period and not the history of play up to the period. Introducing this (together with state complexity in Sabourian [21]) delivers the efficiency results in those games.

The paper is organized as follows. In the following section, we describe the negotiation game and BW's main results. We then introduce the notion of com-

[2] There are in fact many competing proposals of the concept with largely different predictions. See Pearce [16] and Chapter 5.4 of Fudenberg and Tirole [8] for a survey.

plexity in terms of strategies and machines. The machine game will be described. Section 4 presents the main analysis and results. We then run the analogous results with an alternative, more elaborate machine specification in Section 5. We finally conclude. The appendices contain some relegated proofs and also explains that the equilibrium concept we use closely parallels that of Abreu and Rubinstein [1].

2 The Negotiation Game

Let us formally describe the negotiation game, as defined by BW. There are two players indexed by $i = 1, 2$. In the alternating-offers protocol, each player in turn proposes a partition of a *periodic* surplus whose value is normalized to one. If the offer is accepted, the game ends and the players share the surplus accordingly at every period thereafter. If the offer is rejected, the players engage in a one-shot game, called the "disagreement game", before moving onto the next period in which the rejecting player makes a counter-offer.

We index the (potentially infinite) time periods by $t = 1, 2, \ldots$ and adopt the convention that player 1 makes offers in odd periods and player 2 makes offers in even periods. Let $\triangle^2 \equiv \{x = (x_1, x_2) \mid \sum_i x_i = 1\}$ be a partition of the unit periodic surplus. A period then refers to a single offer $x \in \triangle^2$ by one player, a response made by the other player - acceptance "Y" or rejection "N" - and the play of the disagreement game if the response is rejection. The common discount factor is $\delta \in (0, 1)$.

The disagreement game is a two-player normal form game, defined as $G = \{A_1, A_2, u_1(\cdot), u_2(\cdot)\}$. A_i is the set of player i's actions and $u_i(\cdot) : A_1 \times A_2 \rightarrow R$ is his payoff function in the disagreement game. We shall denote the set of outcomes in G by $A = A_1 \times A_2$ with its element indexed by a.[3] Define $u(\cdot) : A \rightarrow R^2$ and assume that it is bounded. Each player's minmax payoff is normalized to zero. Also, we assume that for any $a \in A$

$$u_1(a) + u_2(a) \leq 1$$

Agreement weakly dominates disagreement. Thus, the bargaining offers the players an opportunity to settle on the efficient outcome once and for all.

Two types of outcome paths are possible in the negotiation game; one in which an agreement occurs in a finite time, and one in which disagreement continues perpetually. Let T denote the end of the negotiation game and a^t the disagreement

[3] The normal form may involve sequential moves. In this case, A_i will represent player i's set of strategies, rather than actions, in the disagreement game.

game outcome in period $t < T$. If $T = \infty$, we mean an outcome path in which agreement is never reached. Player i's (discounted) *average* payoff in this case is equal to

$$(1 - \delta) \sum_{t=1}^{\infty} \delta^{t-1} u_i(a^t) \ .$$

If $T < \infty$, denote the agreed partition in T by $z = (z_1, z_2) \in \triangle^2$. Player i's payoff from such an outcome path amounts to

$$(1 - \delta) \sum_{t=1}^{T-1} \delta^{t-1} u_i \left(a^t\right) + \delta^{T-1} z_i \ .$$

The negotiation game is stationary only every two periods (beginning with an odd one) or "stage". In specifying the players' strategies (and later machines), we shall formally distinguish between the different *roles* played by each player in each stage game. He can be either the proposer (p) or the responder (r) in a given period. We shall index a player's role by k. The role distinction provides a natural framework to capture the structural asymmetry that the alternating offers bargaining imposes on the repeated (disagreement) game.

In order to define a strategy, we first need to introduce some further notations. We shall use the following notational convention. Whenever superscripts/subscripts i and j both appear in the same exposition, we mean $i, j = 1, 2$ and $i \neq j$. Similarly, whenever we use superscrpts/subscripts k and l together, we mean $k, l = p, r$ and $k \neq l$.

We shall denote a history of outcomes in a period by e, and this belongs to the set $E = \{(x^i, Y), (x^i, N, a)\}_{x^i \in \triangle^2, a \in A, i=1,2}$ where the superscript i represents the identity of the proposer in the period. Let e^t be the outcome of the period t.

We also need notation to represent information available to a player *within* a period when it is his turn to take an action given his role. To this end, we define a "partial history" (information within a period) , d, as an element in the following set

$$D = \{\emptyset, (x^i), (x^i, N)\}_{x^i \in \triangle^2, i=1,2} \ .$$

For example, the null set here refers to the beginning of a period at which the proposer has to make an offer; (x^i, N) represents a partial history of an offer by player i followed by the other player's rejection.

Also, let us define

$$D_{ik} \equiv \{d \in D \mid \text{it is } i\text{'s turn to play in role } k \text{ after } d \text{ in the period}\}$$

Thus, we have

$$D_{ip} = \{\emptyset, (x^i, N)\}_{x^i \in \triangle^2}$$

and

$$D_{ir} = \{(x^j), (x^j, N)\}_{x^i \in \triangle^2} \ .$$

We denote the set of actions available to player i in the negotiation game by

$$C_i \equiv \triangle^2 \cup Y \cup N \cup A_i \ .$$

Let us denote by $C_{ik}(d)$ the set of actions available to player i given his role k and a corresponding partial history $d \in D_{ik}$. Thus, we have

$$C_{ip}(d) = \begin{cases} \Delta^2 & \text{if } d = \emptyset \\ A_i & \text{if } d = (x^i, N) \end{cases}$$

and

$$C_{ir}(d) = \begin{cases} \{Y, N\} & \text{if } d = x^j \\ A_i & \text{if } d = (x^j, N) \end{cases}$$

Let

$$H^t = \underbrace{E \times \cdots \times E}_{t \text{ times}}$$

be the set of all possible histories of outcomes over t periods in the negotiation game, excluding those that have resulted in an agreement. The initial history is empty (trivial) and denoted by $H^1 = \emptyset$. $H^\infty \equiv \cup_{t=1}^{\infty} H^t$ denotes the set of all possible finite period histories.

For the analysis, we shall divide H^∞ into two smaller subsets according to the different roles that the players play in each stage. Let H^t_{ik} be the set of all possible histories of outcomes over t periods after which player i's role is k. Notice that $H^t_{ik} = H^t_{jl}$. Also, let $H^\infty_{ik} = \cup_{t=1}^{\infty} H^t_{ik}$. Thus, the set of all possible periodic histories of the negotiation game can be written as $H^\infty = H^\infty_{ip} \cup H^\infty_{ir}$ $(i = 1, 2)$.

A strategy for player i is then a function

$$f_i : (H^\infty_{ip} \times D_{ip}) \cup (H^\infty_{ir} \times D_{ir}) \to C_i$$

such that for any $(h, d) \in H^\infty_{ik} \times D_{ik}$ we have $f_i(h, d) \in C_{ik}(d)$. The set of all strategies for player i is denoted by F_i.

We can define a *stationary* (or history-independent) strategy in the following way.

Definition 1 *A strategy f_i is stationary if and only if $f_i(h, d) = f_i(h', d)$ $\forall h, h' \in H^\infty_{ik}$ and $\forall d \in D_{ik}$ for $k = p, r$. A strategy profile $f = (f_i, f_{-i})$ is stationary if f_i is stationary for all i.*

The behavior induced by such a strategy may depend on the partial history within the current period but not on the history of the game up to the period. Notice also that a stationary strategy profile always induces the same outcome in each stage of the game.

In the spirit of the Folk theorem, BW characterize the set of subgame-perfect equilibrium (SPE) payoffs of the above game. BW, to this end, compute the lower bound of each player's SPE payoff in the negotiation game with discount factor δ.

Define

$$w_j = \max_{a \in A} \left\{ u_j(a) - \left[\max_{a_i' \in A_i} u_i(a_i', a_j) - u_i(a) \right] \right\}$$

which BW assume to be well-defined. Note also that $w_i \leq 1$ given the assumption that $u(a) \leq 1 \; \forall a \in A$, and $w_i \geq 0$ if G has at least one Nash equilibrium (given the minmax point). Then, the infimum of player i's SPE payoffs in the negotiation game beginning with his offer (given δ) is not less than

$$\underline{v}_i(\delta) = \frac{1 - w_j}{1 + \delta}$$

while the infimum of the other player's SPE payoffs in the same game is not less than

$$\underline{v}_j(\delta) = \frac{\delta(1 - w_i)}{1 + \delta} \; .$$

BW show that, provided the players are sufficiently patient, these exists a SPE of the negotiation game (beginning with i's offer) in which the players obtain these lower bounds.

Define the limit of these infima as δ goes to unity such that

$$\underline{v}_i = \frac{1 - w_j}{2} \quad \text{and} \quad \underline{v}_j = \frac{1 - w_i}{2} \; .$$

We are now ready to formally recite the key results of BW below.

BW Result 1 *For any payoff vector (v_1, v_2) of the negotiation game such that $v_1 \geq \underline{v}_1$ and $v_2 \geq \underline{v}_2$, $\exists \; \bar{\delta} \in (0,1)$ such that $\forall \delta \in (\bar{\delta}, 1)$ (v_1, v_2) is a SPE payoff vector of the negotiation game with discount factor δ.*

This is BW's main Theorem. Several comments are due. First, many outcome paths are possible to support a feasible payoff vector in equilibrium, some of which will involve delays, and moreover, inefficient disagreement game outcomes before agreement. Perpetual disagreement is also possible.

Second, notice that what determines the nature of equilibria in the negotiation game is the structure of the disagreement game, and not the discount factor or the bargaining surplus available. In particular, the negotiation game will admit a unique subgame-perfect equilibrium only if $w_1 = w_2 = 0$ which implies that any Nash equilibrium payoff vector of the disagreement game has to coincide with its minmax point. Thus, in general, the negotiation game will have a continuum of equilibria much in the way the Folk theorem characterizes the repeated game (even when the disagreement game payoffs are always uniformly small relative to agreement). Nonetheless the forces of bargaining still restrict the set of feasible equilibrium payoffs in the negotiation game substantially compared to the set of individually rational payoffs in the disagreement (repeated) game.

Another relevant result of BW concerns stationary strategies. For a pair of stationary strategies to constitute a SPE of the negotiation game, only a Nash equilibrium of the disagreement game can be played after a rejection; otherwise, there will be a profitable deviation for some player. We can thus analyze the negotiation game as if there is a fixed sequence of disagreement game plays, and consequently, the Rubinstein [19] bargaining result carries over. When we henceforth refer to an equilibrium as being *efficient*, we mean that its outcome is such that either an agreement takes place immediately in the first period or otherwise the disagreement payoffs sum up to one in every period up to agreement. The following puts together Proposition 1 and Corollary 1 of BW.

BW Result 2 *If G has a Nash equilibrium, denoted by $a^* \in A$, then the negotiation game has a subgame-perfect equilibrium in which the strategies are stationary and player 1's offer $z = (z_1, z_2) \in \triangle^2$ such that*

$$z_1 = \frac{1 + \delta u_1(a^*) - u_2(a^*)}{1 + \delta}$$

is accepted immediately. Any other stationary SPE is efficient.

Thus, any stationary SPE of the negotiation game, if exists (which is guaranteed if the disagreement game has at least one Nash equilibrium), must be efficient. Delay is possible (either over one period or indefinitely), but then the Nash equilibrium payoffs must be efficient, i.e. $\sum_i u_i(a^*) = 1$, such that the players are always indifferent between agreement and delay of one period. Note also that if the disagreement game has multiple Nash equilibria there can be many different (but all efficient) payoff distributions that will support the above equilibrium outcome of the negotiation game.

3 Complexity, Machines, and Equilibrium

There are many alternative ways to think of the "complexity" of a strategy in dynamic games. One natural and intuitive way to measure strategic complexity, which we shall adopt in the paper, is to consider the total number of distinct *continuation strategies* that the strategy induces at different histories (Kalai and Stanford [13]).

In a repeated game, it is natural to take the measure over all its possible subgames. In the negotiation game, each stage game is sequential and this means that we can have several different measures of complexity this way. For instance, we can take all possible subgames at the beginning of each period of the negotiation game to correspond with our definition of periodic histories H^t. Let $f_i|h$ be a continuation strategy at history $h \in H^\infty$ induced by $f_i \in F_i$. Thus,

$$f_i|h(h', d) = f_i(h, h', d) \text{ for any } (h, h', d) \in H^\infty_{ik} \times D_{ik} \text{ for any } k\ .$$

Also, let us define the set of all such continuation strategies by $F_i(f_i) = \{f_i|h : h \in H^\infty\}$. Then the cardinality of this set provides a measure of strategic complexity. Let us call it $comp(f_i)$.

The set of continuation strategies can also be divided into smaller sets according to the role specification. Define $F_{ik}(f_i) = \{f_i|h : h \in H^\infty_{ik}\}$. We have $F_i(f_i) = \cup_k F_{ik}(f_i)$. Complexity can then be equivalently measured by $comp(f_i) = \sum_k |F_{ik}(f_i)|$. We can also measure complexity over finer partitions of histories and corresponding continuation strategies. As we shall see, the precise definition of complexity is going to play some role in shaping the results.

In dynamic games any strategy can be *implemented* by an automaton or a "machine" (we shall clarify this statement below). Moreover, Kalai and Stanford [13] show that in repeated games the above notion of complexity of a strategy (the number of continuation strategies) is equivalent to counting the number of states of the (smallest) automaton that implements the strategy. Thus, one could equivalently describe any result either in terms of underlying strategies and their complexity ($comp(\cdot)$) or in terms of machines and their number of states.

We shall establish below that this equivalence between the two representations of strategic complexity also holds in the negotiation game. Our approach to complexity will then be facilitated in machine terms as this will provide a more economical platform to present the analysis of complexity. Each player's strategy space in the negotiation game will be taken as the set of all machines and the players simultaneously and independently choose a single machine at the beginning of the negotiation game. This is the "machine game", a term which we shall interchangeably use with the negotiation game.

Since each stage game of the negotiation game has a sequential structure, many different machine specifications are possible to equivalently represent a strategy. (The same is also the case in other sequential dynamic games; see Piccione and Rubinstein [18], Chatterjee and Sabourian [5][6] and Sabourian [21]). The fact that the stage game is also asymmetric across its two periods - a player switches his role in the bargaining process - adds to this issue of multiple possible machine specifications.

In this paper, we present two particular machine specifications. We choose to run the analysis first with the simpler of the two. The results are in fact sharper under the other specification, but our chosen order of analysis will serve to strengthen the expositional flow. As we shall see later, counting the number of states for these machines corresponds precisely to the manner in which we divide the histories and accordingly define the notion of complexity in terms of (continuation) strategies.

The following defines a machine that employs two "sub-machines".

Definition 2 (Two sub-machine (2SM) specification) *A machine (automaton), $M_i = \{M_{ip}, M_{ir}\}$, consists of two sub-machines $M_{ip} = (Q_{ip}, q^1_{ip}, \lambda_{ip}, \mu_{ip})$ and $M_{ir} = (Q_{ir}, q^1_{ir}, \lambda_{il}, \mu_{il})$ where for any $k, l = p, r$*

$$\begin{aligned} &Q_{ik} \textit{ is the set of states};\\ &q^1_{ik} \textit{ is the initial state belonging to } Q_{ik};\\ &\lambda_{ik} : Q_{ik} \times D_{ik} \to C_i \textit{ is the output function such that}\\ &\qquad \lambda_{ik}(q_{ik}, d) \in C_{ik}(d),\ \forall q_{ik} \in Q_{ik} \textit{ and } \forall d \in D_{ik};\ \textit{ and}\\ &\mu_{ik} : Q_{ik} \times E \to Q_{il} \textit{ is the transition function.} \end{aligned}$$

Each sub-machine in the above definition of a machine consists of a set of *distinct* states, an initial state and an output function enabling a player to play a given role. Transitions take place at the end of each period from a state in one sub-machine to a state in the other sub-machine as roles are reversed each period. We shall sometimes refer to a machine in the above definition simply as a 2SM.

We shall assume that each sub-machine has to have at least one state.[4] But notice that we do not assume finiteness of a machine; each sub-machine may have any arbitrary (possibly infinite) number of states. This is in contrast to Abreu and Rubinstein [1] and others who consider finite automata. Assuming that machines can only have a finite number of states is itself a restriction on the players' choice of strategies.

[4] We could also define a distinct terminal state for each sub-machine. This is immaterial. We are assuming that if an offer is accepted by the responder, M_i enters the terminal state of the relevant sub-machine and shuts off.

Notice also that the initial state of the sub-machine that operates in the second period is in fact redundant because the first state used by this sub-machine depends on the transition function taking place between the first two periods of the game (in terms of strategies, the continuation strategy from the second period on can be contingent on what happens in the first period). Nevertheless, we endow both sub-machines with an initial state for expositional ease.

We can now formally state what we mean by a machine implementing a strategy in the negotiation game. Consider a machine $M_i = \{M_{ip}, M_{ir}\}$ where, for $k = p, r$, $M_{ik} = (Q_{ik}, q^1_{ik}, \lambda_{ik}, \mu_{ik})$. For every $k = p, r$ and for any $h \in H^\infty_{ik}$, denote the state at history h by $q_i(h) \in Q_{ik}$. Formally if $h = (e^1, \ldots, e^{t-1})$ then $q_i(h) = q^t_i$ where for any $0 \leq \tau \leq t$, q^τ_i is defined inductively by

$$q^1_i = \begin{cases} q^1_{ik} & \text{if } i \text{ is in role } k \text{ initially at } t = 1 \\ q^1_{il} & \text{if } i \text{ is in role } l \text{ initially at } t = 1 \end{cases}$$

and for $\tau > 0$

$$q^\tau_i \equiv \begin{cases} \mu_{il}(q^{\tau-1}_i, e^{\tau-1}) & \text{if } i \text{ is in role } k \text{ at } \tau \\ \mu_{ik}(q^{\tau-1}_i, e^{\tau-1}) & \text{if } i \text{ is in role } l \text{ at } \tau \end{cases}$$

Definition 3 *M_i implements f_i if $\forall k$, $\forall h \in H^\infty_{ik}$ and $d \in D_{ik}$*

$$\lambda_{ik}(q(h), d) = f_i(h, d)$$

where $q(h)$ is defined inductively as above.

The following defines a *minimal* machine.

Definition 4 *A machine is minimal if and only if each of its sub-machines has exactly one state.*

A minimal 2SM implements the same actions in every period regardless of the history of the preceding periods, provided that the partial history within the current period (given a role) is the same. Hence, it corresponds to a stationary strategy as in Definition 1.

We have thus far established that machines and strategies are equivalent in the negotiation game. Now let us formally show that *comp*(f_i) is equivalent to counting the total number of states of the machine that implements the strategy f_i. It must be stressed here that the exact specification of a machine is important in qualifying this statement. Since in defining *comp*(f_i) we take continuation strategies at the beginning of each period, we need transitions to take place between *periods* in accordance with the continuation points chosen. It is also important that each sub-machine uses its own distinct set of states.

Let $\|M_i\| = \sum_k |Q_{ik}|$ be the total number of states (or size) of machine M_i in the 2SM specification. The cardinality of the set of continuation strategies that a strategy induces at the beginning of each period of the negotiation game corresponds to the size of the smallest 2SM implementing the strategy.

Proposition 1 *For every $f_i \in F_i$ let $\Phi(f_i)$ be the set of 2SMs that implement f_i. Also, let $\bar{M}_i = \{\bar{M}_{ip}, \bar{M}_{ir}\}$ be such that*

$$\|\bar{M}_i\| \in \{M_i \in \Phi(f_i) \mid \|M_i\| \leq \|M'_i \mid \forall M'_i \in \Phi(f_i)\} \ .$$

Then, we have $|F_{ik}(f_i)| = \|\bar{M}_{ik}\|$ for any $k = p, r$ and thus $\|\bar{M}_i\| = comp(f_i)$.

Proof. The proof is a direct application of the proof of Theorem 1 in Kalai and Stanford [13]. For ease of exposition, it is relegated to Appendix A. $\|$

Given this result, we now formally define the notion of complexity in terms of machines, as adopted in the literature on repeated games played by automata *à la* Rubinstein [20] and Abreu and Rubinstein [1].[5]

Definition 5 (State complexity) *A machine M'_i is more complex than another machine M_i, or $M'_i \succ M_i$, if $\|M'_i\| > \|M_i\|$. Also, we say that M'_i is at least complex as M_i, or $M'_i \succeq M_i$, if $\|M'_i\| \geq \|M_i\|$.*

To wrap up the description of the machine game, let us fix some more notational conventions. Let $M = (M_1, M_2)$ be a machine profile. There are several variables that will depend on the particular machine profile chosen. Given the machine profile M, $T(M)$ is the end of the negotiation game; $z(M) \in \triangle^2$ is the agreement offer if $T(M) < \infty$; $a^t(M)$ is the disagreement game outcome in period $t < T(M)$; and $q_i^t(M)$ is the state of player i's (sub-)machine appearing in period $t \leq T(M)$ induced by M.

[5] We also draw attention to the work of Binmore, Piccione, and Samuelson [2] who propose another notion of complexity similar to state complexity considered in this paper and others. According to their "collapsing state condition", an automaton M^1 is less complex than another automaton M^2 if the same implementation can be obtained by consolidating a collection of states belonging to M^2 into a single state in M^1. It will not be difficult to see that our results will also hold under this scheme.

Similarly, we denote by $\pi_i^t(M)$ player i's (discounted) average *continuation payoff* at period t when the machine profile M is chosen, and this amounts to

$$\pi_i^t(M) = \begin{cases} (1-\delta)\sum_{\tau=t}^{\infty} \delta^{\tau-t} u_i(a^\tau(M)) & \text{if } T(M) = \infty \\ (1-\delta)\sum_{\tau=t}^{T-1} \delta^{\tau-t} u_i\left(a^\tau(M)\right) + \delta^{T-t} z_i(M) & \text{if } t < T(M) < \infty \\ z_i(M) & \text{if } t = T(M) \end{cases}$$

We shall use the abbreviation $\pi_i^1(M) = \pi_i(M)$.

For ease of exposition, the argument in M will sometimes be dropped when we refer to one of these variables that depends on the particular machine profile, e.g. $\pi_i^t \equiv \pi_i^t(M)$. Unless otherwise stated, the variable will refer to the profile in the *claim*.

We now introduce an equilibrium notion that captures the players' preference for less complex strategies. There are several ways of refining Nash equilibrium with complexity. We choose an equilibrium notion in which complexity enters a player's preferences *after* the payoffs and with a (non-negative) fixed cost c.[6]

To facilitate this concept, we first define the notion of ϵ-best response. (The following definition can equivalently refer to underlying strategies.)

Definition 6 *For any $\epsilon \geq 0$, a machine M_i is a ϵ-best response to M_{-i} if, $\forall M_i'$,*

$$\pi_i(M_i, M_{-i}) + \epsilon \geq \pi_i(M_i', M_{-i}) \ .$$

If a machine is a 0-best response, then it is a best response in the conventional sense.

Using this, we define a NEMc.

Definition 7 *A machine profile $M^* = (M_1^*, M_2^*)$ constitutes a Nash equilibrium of the machine game with complexity cost $c \geq 0$ (NEMc) if, $\forall i$,*

(i) M_i^* is a best response to M_{-i}^*; and
(ii) There exists no M_i' such that M_i' is a c-best response to M_{-i}^* and $M_i^* \succ M_i'$.

By definition, the set of NEMc is a subset of the set of Nash equilibria in the negotiation game. The case of zero complexity cost $c = 0$ is closest to the standard equilibrium and corresponds to the case of *lexicographic preferences*. Any NEMc with a positive complexity cost $c > 0$ must also be a NEMc with $c = 0$. The

[6] Sabourian [21] employs this equilibrium notion.

magnitude of c therefore can be interpreted as a measure of how much the players care for less complex strategies, or indeed the players' *bounded rationality.*

Abreu and Rubinstein [1], henceforth referred to as AR, propose a general way of describing a player's preference ordering over machine profiles that is increasing in his payoff of the game and decreasing in the complexity of his machine. A Nash equilibrium can then be written in terms of machines that are most preferred against each other. In contrast, our equilibrium concept directly finds a subset of Nash equilibria of the underlying game that fits our complexity cost criterion (at the margin). There is, however, an analytical parallel between our choice of solution concept and that of AR because the latter must also be a Nash equilibrium of the negotiation game (see Appendix B). Our complexity cost criterion can be thought of as an alternative way to embed the trade-off between payoff and complexity that underlies AR's preference ordering.

NEMc strategy profiles are not necessarily credible however. We could introduce credibility, as in Chatterjee and Sabourian [5][6], by introducing trembles into the model and considering the limit of extensive form trembling hand equilibrium (Nash equilibrium with independent trembles at each information set) with complexity cost as the trembles become small. The noise will ensure that strategies are optimal (allowing for complexity) after all histories that occur with a positive probability.

A more direct, and simpler, way of introducing credibility would be to consider NEMc strategy profiles that are subgame-perfect equilibria of the negotiation game without complexity cost.

Definition 8 *A machine profile $M^* = (M_1^*, M_2^*)$ constitutes a subgame-perfect equilibrium of the machine game with complexity cost $c \geq 0$ (SPEMc) if M^* is both a NEMc and a subgame-perfect equilibrium (SPE) of the negotiation game.*

We shall denote by $\Omega^\delta(c)$ the set of SPEMc profiles in the negotiation game with common discount factor δ when the complexity cost is c.

Given Proposition 1, we can equivalently define these notions of equilibrium (NEMc and SPEMc) in terms of underlying strategies and the corresponding measure of complexity $comp(\cdot)$. As mentioned earlier, we prefer the machine game analysis for its expositional economy.

4 Analysis: Complexity and Efficiency

4.1 Some Preliminary Results

In this sub-section, we lay out some Lemmas that will pave way for the main results below. These results are derived independently of the magnitude of complexity cost.

We first state an obvious, yet very important, implication of the complexity requirement. Every state belonging to the equilibrium machines has to appear on the equilibrium path. If there is a state that does not appear on the equilibrium path, it can be "dropped" to reduce complexity cost without affecting the outcome and payoff.

Lemma 1 *Assume that $M^* = (M_1^*, M_2^*)$ is a NEMc with $c \geq 0$. Let $M_i^* = \{M_{ip}^*, M_{ir}^*\}$ where, for $k = p, r$, $M_{ik}^* = (Q_{ik}^*, q_{ik}^{1*}, \lambda_{ik}^*, \mu_{ik}^*)$. Then, $\forall q_i \in Q_{ik}^*$, $\forall i$ and $\forall k$, there exists a period t such that $q_i^t(M^*) = q_i$.*

Proof: Suppose not. So suppose that there exists some $\bar{q}_i \in Q_{ik}^*$ that does not appear in any period t on the equilibrium path.

But then, consider player i using another machine $M_i' = \{M_{ip}', M_{ir}'\}$ which is identical to M_i^* except only that $\bar{q}_i$ is dropped (so the set of states of M_{ik}' is just $Q_{ik}^* \backslash \bar{q}_i$).

Clearly $\pi_i(M_i', M_j^*) = \pi_i(M_i^*, M_j^*)$, and moreover, we have $M_i^* \succ M_i'$. Hence, we have contradiction against the assumption that M^* is a NEMc profile. ||

A NEMc machine may have an infinite number of states. But, It follows from Lemma 1 that:

Corollary 1 *If $M^* = (M_1^*, M_2^*)$ is a NEMc with $c \geq 0$, then M_i^* $(i = 1, 2)$ must have a countable number of states.*

Next note that since any strategy can be implemented by a machine it follows from its definition that any NEMc profile $M^* = (M_1^*, M_2^*)$ corresponds to a Nash equilibrium of the underlying negotiation game; thus $(\forall c \geq 0)$

$$\pi_i(M_i^*, M_{-i}^*) = \max_{f_i \in F_i} \pi_i(f_i, M_{-i}^*) \ \forall i \tag{1}$$

More generally, the equilibrium machines must be best response (in terms of payoffs) *along* the equilibrium path of the negotiation game. The following must be the case:

Lemma 2 *Assume that $M^* = (M_1^*, M_2^*)$ is a NEMc with $c \geq 0$. Then, $\forall i, j$ and $\forall \tau \leq T(M^*)$, we have*

$$\pi_i^\tau(M^*) = \max_{f_i \in F_i} \pi_i(f_i, M_j^*(q_j^\tau)) \ .$$

where $q_j^\tau \equiv q_j^\tau(M^*)$, and $M_j^*(q_j^\tau)$ is the machine that is identical to M_j^* except that it starts with the sub-machine which operates in period τ with the initial state q_j^τ.

***Proof*.** Suppose not. Then, for some i and $\tau \leq T(M^*)$, there exists another machine $\bar{M}_i = \{\bar{M}_{ip}, \bar{M}_{ir}\}$ such that

$$\pi_i^\tau(M^*) < \pi_i(\bar{M}_i, M_j^*(q_j^\tau)) \ .$$

Now, consider player i using at the outset another machine $M_i' = \{M_{ip}', M_{ir}'\}$ where, for $k = p, r$, $M_{ik}' = (Q_{ik}', q_{ik}^{1\prime}, \lambda_{ik}', \mu_{ik}')$. This machine is constructed in the following way. Let $q_i^t \in Q_{ik}^*$ denote the state of M_i^* appearing in period t (where i is in role k). Also let e^t be the outcome in period t induced by M^*. For every $t < \tau$, there exists a *distinct* state $q_i'(t) \in Q_{ik}'$ such that

$$\lambda_{ik}'(q_i'(t), d) = \lambda_{ik}^*(q_i^t, d) \text{ for all } d \in D_{ik} \ .$$

The transition function of the new machine is such that $\forall t < \tau - 1$

$$\mu_{ik}'(q_i'(t), e^t) = q_i'(t+1)$$

and for $t = \tau - 1$

$$\mu_{ik}'(q_i'(t), e^t) = \bar{q}$$

where $\bar{q} \in \bar{Q}_{ik}$ is another distinct state such that $M_i'(\bar{q}) = \bar{M}_i$.

Thus, M_i' played against M_j^* replicates the outcome path up to τ such that

$$\sum_{t=1}^{\tau-1} \delta^{t-1} u_i\left(a^t(M_i', M_j^*)\right) = \sum_{t=1}^{\tau-1} \delta^{t-1} u_i\left(a^t(M_i^*, M_j^*)\right)$$

followed by activation of $\bar{M}_i$ at τ. It follows that

$$\pi_i(M_i', M_j^*) > \pi_i(M_i^*, M_j^*) \ .$$

But this contradicts (1) above. $\|$

Now it follows that if a state belonging to a player's equilibrium machine appears twice on the outcome path then the continuation payoff of the other player must be identical at both periods.

Lemma 3 *Assume that $M^* = (M_1^*, M_2^*)$ is a NEMc with $c \geq 0$. Then, $\forall i, j$ and $\forall t, t' \leq T(M^*)$, we have the following:*

$$\text{if } \ q_j^t(M^*) = q_j^{t'}(M^*), \text{ then } \pi_i^t(M^*) = \pi_i^{t'}(M^*) \ .$$

Proof. This follows from Lemma 2. $\|$

Using this information, we can show that if a state belonging to a player's equilibrium machine appears on the outcome path for the first time, then the state of the other player's machine in that period must also be appearing for the first time. This Lemma will provide a critical tool behind the derivation of some of the main results below.

Lemma 4 *Assume that $M^* = (M_1^*, M_2^*)$ is a NEMc with $c \geq 0$. Then, for any i and any $\tau \leq T(M^*)$, we have the following:*

$$\textit{if } \; q_i^\tau(M^*) \neq q_i^t(M^*) \;\forall t < \tau, \;\textit{ then } q_j^\tau(M^*) \neq q_j^t(M^*) \;\forall t < \tau \;.$$

Proof. Suppose not. So, there exists some i and some $\tau \leq T$ such that $q_i^\tau \neq q_i^t$ $\forall t < \tau$ and $q_j^{\tau'} = q_j^\tau$ for some $\tau' < \tau$. Then, by Lemma 3, $\pi_i^\tau = \pi_i^{\tau'}$.

Consider player i using another machine $M_i' = \{M_{ip}', M_{ir}'\}$ where, for $k = p, r$, $M_{ik}' = (Q_{ik}', q_{ik}^{1\prime}, \lambda_{ik}', \mu_{ik}')$. This machine is identical to M_i^* except that:

- q_i^τ is dropped; and
- the transition function is such that $\mu_{ik}'(q_i^{\tau-1}, e^{\tau-1}) = q_i^{\tau'}$ $(k \in \{p, r\})$.

To be precise, M_i' is such that (assume that i is in role k in period τ)

- $Q_{ik}' = Q_{ik}^* \backslash q_i^\tau$ and $Q_{il}' = Q_{il}^*$;
- $q_{ik}^{1\prime} = q_{ik}^{1*}$ and $q_{il}^{1\prime} = q_{il}^{1*}$;
- for every $k' = p, r$, every $q_i \in Q_{ik'}'$, and every $d \in D_{ik}$

$$\lambda_{ik'}'(q_i, d) = \lambda_{ik'}^*(q_i, d)$$

- for every k, every $q_i \in Q_{ik}'$ and every $e \in E$

$$\mu_{ik}'(q_i, e) = \mu_{ik}^*(q_i, e)$$

  and for every l, every $q_i \in Q_{il}'$ and every $e \in E$

$$\mu_{il}'(q_i, e) = \begin{cases} q_i^{\tau'} & \text{if } q_i = q_i^{\tau-1} \\ \mu_{il}^*(q_i, e) & \text{otherwise} \end{cases}$$

  where $e^{\tau-1} \in E$ is the outcome that M^* generates in period $\tau - 1$.

Since q_i^τ appears for the first time in period τ on the original equilibrium path, we cannot have $q_i^{\tau-1}$ and $e^{\tau-1}$ appearing together before $\tau - 1$. Otherwise, the transition function of the equilibrium machine would induce q_i^τ before τ which contradicts our assumption of τ.

Thus, playing M_i' against M_j^* does not alter the outcome path up to τ. But from τ onwards, the outcome path between τ' and $\tau - 1$ will repeat itself *ad infinitum.*

This does not change i's payoff from the machine game (given M_j^*). We know that

$$\begin{aligned}
\pi_i^{\tau'}(M_i^*, M_j^*) &= \sum_{t=\tau'}^{\tau-\tau'} \delta^{t-\tau'} u_i(a^t) + \delta^{\tau-\tau'} \pi_i^{\tau}(M_i^*, M_j^*) \\
&= \sum_{t=\tau'}^{\tau-\tau'} \delta^{t-\tau'} u_i(a^t) + \delta^{\tau-\tau'} \pi_i^{\tau'}(M_i^*, M_j^*) \\
&= \frac{1}{1-\delta^{\tau-\tau'}} \sum_{t=\tau'}^{\tau-1} \delta^{t-\tau'} u_i(a^t) \qquad (2)
\end{aligned}$$

where the second equality follows from Lemma 3. The new machine also yields the same payoff because

$$\begin{aligned}
\pi_i^{\tau'}(M_i', M_j^*) &= \sum_{t=\tau'}^{\tau-1} \delta^{t-\tau'} u_i(a^t) + \delta^{\tau-\tau'} \sum_{t=\tau'}^{\tau-1} \delta^{t-\tau'} u_i(a^t) + \ldots \\
&= \sum_{t=\tau'}^{\tau-1} \delta^{t-\tau'} u_i(a^t)(1 + \delta^{\tau-\tau'} + \delta^{2(\tau-\tau')} + \ldots) \\
&= \frac{1}{1-\delta^{\tau-\tau'}} \sum_{t=\tau'}^{\tau-1} \delta^{t-\tau'} u_i(a^t)\ . \qquad (3)
\end{aligned}$$

Since (M_i', M_j^*) and (M_i^*, M_j^*) induce the same outcome before τ', it follows that $\pi_i(M_i', M_j^*) = \pi_i(M_i^*, M_j^*)$. But then, since q_i^τ is dropped, $M_i^* \succ M_i'$. Thus, we have contradiction against NEMc.[7] ||

[7] Notice that this result turns on the assumption that each sub-machine uses a distinct set of states. If the sub-machines shared the states, we could not simply "drop" q_i^τ since it could be used for the other sub-machine (playing a different role) before τ'.

4.2 Agreement

In this sub-section, we shall show that, independently of c, if an agreement occurs at some finite period as a NEMc outcome, then it must occur within the very first stage (two periods) of the negotiation game.

We can immediately state that if an agreement occurs within the first stage as a NEMc outcome, then the associated equilibrium machines must be minimal, and thus, the implemented strategies must be stationary.

Lemma 5 *If $M^* = (M_1^*, M_2^*)$ is a NEMc with $c \geq 0$ and $T(M^*) \leq 2$, then $|Q_{ik}| = 1$ $\forall i$ and $\forall k$.*

Proof: Suppose not. So, suppose that $|Q_{ik}| > 1$ for some i and for some k. But then, for this player i, dropping every state in his machine other than the two states appearing the first and second periods leaves his payoff unchanged and yet reduces complexity cost. Hence, we have contradiction against NEMc. $\|$

Next we show that if a NEMc induces an agreement in a finite period beyond the first stage, it must be that the pair of states appearing in the final period are distinct.

Lemma 6 *Assume that $M^* = (M_1^*, M_2^*)$ is a NEMc with $c \geq 0$ and $T(M^*) < \infty$. Then, $q_i^t(M^*) \neq q_i^T(M^*)$ $\forall t < T(M^*)$ and $\forall i$.*

Proof. Suppose not. So, suppose that $q_i^t = q_i^T$ for some i and some $t < T$. Let $z = (z_1, z_2) \in \triangle^2$ be the agreement at T. There are two possible cases to consider.

Case A: Player i is the proposer at T.

Define $\tau = \min\{t | q_i^t = q_i^T\}$. By Lemma 3, $\pi_j^\tau = \pi_j^T$. Since there is an agreement on z at T, we have $\pi_j^\tau = z_j$.

Now consider player j using another machine $M_j' = \{M_{jp}', M_{jr}'\}$ where, for $k = p, r$, $M_{jk}' = (Q_{jk}', q_{jk}^{1\prime}, \lambda_{jk}', \mu_{jk}')$. This machine is identical to M_j^* except that:

- q_j^τ is dropped (i.e. $Q_{jr}' = Q_{jr}^* \backslash q_j^\tau$); and
- the transition function is such that $\mu_{jp}'(q_j^{\tau-1}, e^{\tau-1}) = q_j^T$.

Since, by Lemma 4, q_j^τ (as does q_i^τ by definition) appears for the first time at τ on the original equilibrium path, this new machine (given M_i^*) generates an identical outcome path as the original machine M_j^* up to τ and then induces the agreement z at τ. We know $\pi_j^\tau = z_j$, and thus, it follows that $\pi_j(M_i^*, M_j') = \pi_j(M_i^*, M_j^*)$.

But since q_j^τ is dropped, $M_j^* \succ M_j'$. This contradicts NEMc.

Case B: Player i is the responder at T.

We can show contradiction similarly to Case A above. $\|$

We are now ready to present our first major result. For any value of complexity cost, any NEMc outcome that reaches an agreement must do so in the very first stage (period 1 or 2) of the negotiation game and hence the associated strategies must be *stationary.* The intuition is as follows. The state of each player's machine occurring in the last period must be distinct. This implies that, if the last period occurs beyond the first stage of the game, one of the players must be able to drop it and instead use another state in his (sub-)machine to condition his behavior in that period without affecting the outcome of the game. This reduces complexity cost.

Proposition 2 *Assume that* $M^* = (M_1^*, M_2^*)$ *is a NEMc with* $c \geq 0$ *and* $T(M^*) < \infty$*. Then (i)* $T(M^*) \leq 2$*; and (ii)* M_1^* *and* M_2^* *are minimal.*

Proof. If part (i) of the claim is true, part (ii) must be true because of Lemma 5. Let us consider part (i).

Suppose not. So, suppose that an agreement $z \in \triangle^2$ occurs at some $T \in (2, \infty)$. We know from Lemma 6 that q_1^T and q_2^T are both distinct. Now suppose that player i is the proposer at T and consider two possible cases.

Case A: $x^\tau = z$ at some $\tau < T$ where i proposes.

Consider another machine $M_i' = \{M_{ip}', M_{ir}'\}$ where, for $k = p, r$, $M_{ik}' = (Q_{ik}', q_{ik}^{1\prime}, \lambda_{ik}', \mu_{ik}')$ which is identical to M_i^* except that:

- q_i^T is dropped (i.e. $Q_{ip}' = Q_{ip}^* \backslash q_i^T$); and
- the transition function is such that $\mu_{ir}'(q_i^{T-1}, e^{T-1}) = q_i^\tau$.

Since $\lambda_{ip}'(q_i^\tau, \emptyset) = z$ and q_i^T appears for the first time at T on the original outcome path, this new machine (given M_j^*) generates an identical outcome path and payoff as the original machine M_i^*. But then, q_i^T is dropped and therefore we have $M_i^* \succ M_i'$. This contradicts NEMc.

Case B: $x^\tau \neq z$ $\forall \tau < T$ where i proposes.

Consider another machine $M_j' = \{M_{jp}', M_{jr}'\}$ where, for $k = p, r$, $M_{jk}' = (Q_{jk}', q_{jk}^{1\prime}, \lambda_{jk}', \mu_{jk}')$ which is identical to M_j^* except that:

- q_j^T is dropped (i.e. $Q_{jr}' = Q_{jr}^* \backslash q_j^T$);

- the transition function is such that $\mu'_{jp}(q_j^{T-1}, e^{T-1}) = q_j$ for some arbitrary but fixed $q_j \in Q'_{jr}$; and
- the output function is such that $\lambda'_{jr}(q_j, z) = Y$.

Since the offer z does not appear anywhere before T on the original outcome path (when i proposes), the modified output function does not affect the outcome and payoff. But then, q_j^T is dropped and therefore we have $M_j^* \succ M_j'$. This contradicts NEMc. $\|$

Together with subgame-perfectness requirement (see BW Result 2 above), Proposition 2 tells us that if there is an agreement in the negotiation game the outcome must be efficient. Also, non-emptiness of the set of SPEMc ($\Omega^\delta(c)$) is guaranteed (for any δ and any c) if the disagreement game has at least one Nash equilibrium.

Corollary 2 *For any $c \geq 0$ and any $\delta \in (0,1)$, if any $M^* \in \Omega^\delta(c)$ is such that $T(M^*) < \infty$, then M^* must be efficient and minimal (implements stationary strategies). If G has at least one Nash equilibrium, then such SPEMc exists in the negotiation game.*

4.3 Perpetual Disagreement

We now consider SPEMc outcomes in which agreement never occurs. The results here are sensitive to whether the complexity cost is zero $c = 0$ (lexicographic preferences), or positive $c > 0$.

First, we show that, given any complexity cost and a discount factor arbitrarily close to one, any SPEMc outcome with perpetual disagreement must be at least *long-run* (almost) efficient; that is, the players must eventually reach a finite period at which the sum of their continuation payoffs is approximately equal to one.

The argument behind this statement turns critically on the fact that every state of each player's equilibrium machine must appear on the outcome path (Lemma 1). This implies the following. Suppose that a player deviates from a SPEMc of the negotiation game by making a different offer in some period. What can the other player obtain if he rejects this offer? Since the state of each player's (sub-)machine is fixed for each period (not each decision node), the ensuing disagreement game of the period may see an outcome that never happens on the original equilibrium path; but then, Lemma 1 implies that the subsequent transition must take the players to some point along the original path for next period. Thus, any punishment for a player who deviates from the proposed equilibrium must itself occur on the equilibrium path (except for the play of the disagreement game immediately after the deviating offer), and as a consequence, the set of equilibrium outcomes is severely restricted.

In loose terms, we consider the period in which a player gets his maximum continuation payoff in the proposer role. Bargaining can then be used by the other player in the *preceding* period to break up the on-going disagreement if there is any (continuation) inefficiency from then on. In such cases, there exists a Pareto-improving deviation offer because the responder in that period, who will be proposing next, cannot obtain more from punishing the deviant than what he is already getting from the original outcome as of next period. We need the discount factor to be sufficiently large so as to eliminate the importance of the current period in which the deviation is followed immediately by an off-the-equilibrium play of the disagreement game.

For the results below,

Proposition 3 *For any $\epsilon \in (0,1)$, $\exists\, \bar{\delta} < 1$ such that, for any $\delta \in (\bar{\delta}, 1)$, any $c \geq 0$ and any $M^* \in \Omega^\delta(c)$ with $T(M^*) = \infty$, $\exists\, \tau < \infty$ such that $\sum_i \pi_i^\tau(M^*) > 1 - \epsilon$.*

Proof: Fix any $\epsilon \in (0,1)$. Define

$$\beta = \max\left\{1, \sup_{a,a' \in A, i} [u_i(a) - u_i(a')]\right\} \tag{4}$$

which is bounded since $u(\cdot)$ is. Define also

$$\bar{\delta} = 1 - \frac{\epsilon}{\beta} \ .$$

Given these, consider any $\delta \in (\bar{\delta}, 1)$ (thus $\epsilon > \beta(1-\delta)$), any $c \geq 0$ and any $M^* = (M_1^*, M_2^*) \in \Omega^\delta(c)$. As before, let $M_i^* = \{M_{ip}^*, M_{ir}^*\}$ where, for $k = p, r$, $M_{ik}^* = (Q_{ik}^*, q_{ik}^{1*}, \lambda_{ik}^*, \mu_{ik}^*)$.

Define η, t_{ik} and τ_η such that

$$0 < \eta < \epsilon - \beta(1-\delta), \tag{5}$$

$$t_{ik} = \{t |\ i \text{ plays role } k\},$$

and

$$\tau_\eta = \min\{t |\ \pi_i^t + \eta > \pi_i^{t'}\ \forall t, t' \in t_{ip}\}$$

where, as before, π_i^t is player i's continuation payoff at period t if M^* is chosen. Clearly, $\tau_\eta < \infty$.

Now, take any machine profile (M_i, M_j^*) and consider i's continuation payoff after rejecting any offer in any period belonging to t_{ir}. Notice that since

- every state of M_j^* appears on the equilibrium path of M^* (Lemma 1)

- $\pi_i^t = \max_{f_i}\left(f_i, M_j^*(q_j^t)\right)\ \forall t$ (Lemma 2)

player i's continuation payoff *at the next period* if he rejects any offer (given M_j^*) is at most $\sup_{t\in t_{ip}} \pi_i^t$. We know that $\pi_i^{\tau_\eta} + \eta > \pi_i^t\ \forall t \in t_{ip}$.

This implies that under profile M^* if i receives an offer $(\pi_{ir}^{\max}, 1-\pi_{ir}^{\max}) \in \triangle^2$ where

$$\pi_{ir}^{\max} = (1-\delta)\sup_{a\in A} u_i(a) + \delta\left(\pi_i^{\tau_\eta} + \eta\right), \tag{6}$$

he must always accept because of the subgame-perfectness of M^*.

Now, consider player j using another machine $M_j' = \{M_{jp}', M_{jr}'\}$ where, for $k = p, r$, $M_{jk}' = (Q_{jk}', q_{jk}^{1\prime}, \lambda_{jk}', \mu_{jk}')$. This machine is identical to M_j^* except for the output function which is such that $\lambda_{jp}'(q_j^{\tau_\eta - 1}, \emptyset) = (\pi_{ir}^{\max}, 1-\pi_{ir}^{\max})$.

Define

$$\tau = \min_t\{t|\ q_j^t = q_j^{\tau_\eta-1}\}\ . \tag{7}$$

Since i always accepts the offer $\pi_{ir}^{\max}$ given M_j^* and M_j' differs from M_j^* only in offers, it follows that (M_i^*, M_j') results in an agreement $(\pi_{ir}^{\max}, 1-\pi_{ir}^{\max})$ in period τ.

We also know by Lemma 3 that $\pi_i^\tau = \pi_i^{\tau_\eta - 1}$. Thus, we have

$$\pi_i^\tau = (1-\delta)u_i(a^{\tau_\eta-1}) + \delta\pi_i^{\tau_\eta}\ .$$

Now, since $\sup_{a\in A} u_i(a) - u_i(a^{\tau_\eta - 1}) \le \beta$ (where β is given by (4)), we have, by the definition of $\pi_{ir}^{\max}$,

$$\pi_{ir}^{\max} - \pi_i^\tau \le (1-\delta)\beta + \delta\eta\ .$$

Using this, we can write

$$1 - \pi_{ir}^{\max} \ge 1 - \left(\pi_i^\tau + (1-\delta)\beta + \delta\eta\right)\ . \tag{8}$$

Since M^* is a SPEMc it must be that $\pi_j^\tau \ge 1 - \pi_{ir}^{\max}$; otherwise the deviation is profitable. This implies that (given $\delta < 1$)

$$\pi_i^\tau + \pi_j^\tau > 1 - \left((1-\delta)\beta + \eta\right)\ .$$

But, since by (5) we have $\epsilon > (1-\delta)\beta + \eta$, it follows that at period $\tau < \infty$, $\sum_i \pi_i^\tau > 1 - \epsilon$ as in the claim. ||

Proposition 3 does not however rule out the possibility that we observe inefficiency (in terms of continuation payoffs) early on in the negotiation game.[8] Given any $\epsilon > 0$ and δ sufficiently close to one, we can write the total equilibrium payoff from the negotiation game as

$$\sum_i \pi_i(M^*) > (1-\delta)\sum_{t=1}^{\tau-1}\delta^{t-1}u^t + \delta^{\tau-1}(1-\epsilon) \tag{9}$$

where M^* indicates the equilibrium machine profile, $u^t = \sum_i u_i\left(a^t\left(M^*\right)\right)$, and τ is the period in which continuation (first) becomes (almost) efficient. The limit of the right-hand side as $\epsilon \to 0$ and $\delta \to 1$ is not necessarily the efficient level. The reason is that as we increase δ we are changing the equilibrium strategy profile itself, and consequently, τ may also increase, that is, it may take longer and longer to reach the efficient long-run.[9]

But, it immediately follows from Proposition 3 that if the structure of the disagreement game is such that there exists no action profile delivering the efficient surplus, the players cannot disagree forever. Then, the results in the previous section imply that any SPEMc must induce an agreement in the very first stage of the game and thus be efficient (and stationary). We summarize this below.

Corollary 3 *If $\sum_i u_i(a) < 1$ $\forall a \in A$, then $\exists\ \bar{\delta} \in (0,1)$ such that, for any $\delta \in (\bar{\delta}, 1)$ and any $c \geq 0$, every $M^* \in \Omega^\delta(c)$ is efficient (and stationary) with $T(M^*) \leq 2$.*

Agreement will strictly dominate any disagreement if playing the disagreement game involves some cost to the players (that bargaining does not). They may, for instance, discount the time between bargaining and disagreement game within a period.

In fact, we derive a qualitatively same result from a complexity argument. If complexity cost is strictly positive, i.e. $c > 0$, disagreement cannot persist indefinitely however small that complexity cost is, and thus, any SPEMc of the negotiation game ends in the first stage and is efficient.

Proposition 4 *For any $c \in (0,1)$, $\exists\ \bar{\delta} < 1$ such that, for any $\delta \in (\bar{\delta}, 1)$, every $M^* \in \Omega^\delta(c)$ is efficient (and stationary) with $T(M^*) \leq 2$.*

[8] To be precise, neither does it rule out the possibility that there will be inefficient disagreement game outcomes even after τ. It is just that the continuation game from then on is almost efficient.

[9] If we restrict each player's machine to use only a finite number of states, then any machine profile must generate *cycles*. But this is not enough to guarantee that Proposition 3 implies ex ante efficiency in the limit. For this, we need for instance to additionally assume that the size of a machine is uniformly bounded so that the first cycle cannot last beyond a fixed period.

Proof. We shall prove the claim by way of contradiction.

Fix any $c \in (0,1)$.[10] Define $\bar{\delta} = 1 - \frac{c}{\beta}$ where β is given by (4) above. Given these, consider any $\delta \in (\bar{\delta}, 1)$, and any $M^* = (M_1^*, M_2^*) \in \Omega^\delta(c)$. Suppose $T(M^*) = \infty$.

Similarly to the proof of Proposition 3 above, define η such that

$$0 < \eta < c - \beta(1-\delta) \ . \tag{10}$$

Define as before

$$t_{ik} = \{t | \ i \text{ plays role } k\},$$

$$\tau_\eta = \min\{t | \ \pi_i^t + \eta > \pi_i^{t'} \ \forall t, t' \in t_{ip}\},$$

and

$$\tau = \min_t \{t | \ q_j^t = q_j^{\tau_\eta - 1}\} \ .$$

First note that

$$q_j^t \neq q_j^{\tau_\eta} \ \ \forall t < \tau_\eta \ . \tag{11}$$

Otherwise $q_j^t = q_j^{\tau_\eta}$ for some $t < \tau_\eta$. But then, we have $\pi_i^t = \pi_i^{\tau_\eta}$ by Lemma 3. This contradicts the definition of τ_η.

Next, consider j using another machine $M_j' = \{M_{jp}', M_{jr}'\}$ where, for $k = p, r$, $M_{jk}' = (Q_{jk}', q_{jk}^{1\prime}, \lambda_{jk}', \mu_{jk}')$. This machine is identical to M_j^* except that:

- (similarly to the deviation in the proof of the previous proposition) the output function is such that $\lambda_{jp}'(q_j^{\tau_\eta - 1}, \emptyset) = (\pi_{ir}^{\max}, 1 - \pi_{ir}^{\max})$ (where $\pi_{ir}^{\max}$ is defined by (6) with η now given by (10) above); and additionally
- $q_j^{\tau_\eta}$ is dropped (i.e. $Q_{jr}' = Q_{jr}^* \backslash q_j^{\tau_\eta}$).

As in the proof of Proposition 3, such deviation results in an acceptance and would end the negotiation game at τ. By (11), dropping $q_j^{\tau_\eta}$ does not affect the outcome path up to τ. By the same argument as in the proof of previous Proposition, j's deviation payoff here is given by (8) above:

$$1 - \pi_{ir}^{\max} \geq 1 - (\pi_i^\tau + (1-\delta)\beta + \delta\eta) \ .$$

We know that $1 - \pi_i^\tau \geq \pi_j^\tau$. Thus, j's loss from such deviation is such that

$$\pi_j^\tau - (1 - \pi_{ir}^{\max}) \leq (1-\delta)\beta + \delta\eta \ . \tag{12}$$

[10] The case of $c \geq 1$ is trivial because then complexity cost (weakly) dominates any feasible average payoff for each player in the negotiation game and thus any equilibrium machine must be minimal. We can refer to BW Result 2 for SPEMc characterization in this case.

But the new machine M_j' has one less state than M_j^* (since $q_j^{\tau_\eta}$ has been dropped) which means that the deviation also results in a saving on complexity cost by $c > 0$. Since we fixed $c > (1-\delta)\beta + \eta$ and $\delta < 1$, we have

$$\pi_j^\tau - (1 - \pi_{ir}^{\max}) < c$$

implying that the deviation is in fact profitable. (More precisely, this implies that M_j^* is not a *c-best response* to M_i^*.) This contradicts the proposed SPEMc. Therefore, $T(M^*) < \infty$. But then, we know from Proposition 2 that $T(M^*) \leq 2$ and from Corollary 2 that M^* is efficient. $\|$

5 An Alternative Machine Specification

Since each stage game of the negotiation game has a sequential structure, we can have alternative machine specifications that employ more frequent transitions and hence account for finer partitions of histories and continuation strategies. Let us present a machine which consists of four sub-machines. This machine will sometimes be referred to as 4SM.

Definition 9 (Four sub-machine (4SM) specification) *A machine, $M_i = \{\tilde{M}_{ip}, M_{ip}, \tilde{M}_{ir}, M_{ir}\}$, consists of four sub-machines $\tilde{M}_{ik} = (\tilde{Q}_{ik}, \tilde{q}_{ik}^1, \tilde{\lambda}_{ik}, \tilde{\mu}_{ik})$ and $M_{ik} = (Q_{ik}, q_{ik}^1, \lambda_{ik}, \mu_{ik})$ for $k = p, r$. Each sub-machine consists of a set of states, an initial state, an output function and a transition function such that, $\forall \tilde{q}_{ik} \in \tilde{Q}_{ik}$, $\forall q_{ik} \in Q_{ik}$, $\forall x^i, x^j \in \triangle^2$, and $\forall a \in A$,*

$$\begin{aligned}
\tilde{\lambda}_{ip}(\tilde{q}_{ip}, \emptyset) &\in \triangle^2; \\
\lambda_{ip}(q_{ip}, \emptyset) &\in A_i; \\
\tilde{\lambda}_{ir}(\tilde{q}_{ir}, x^j) &\in \{Y, N\}; \\
\lambda_{ir}(q_{ir}, \emptyset) &\in A_i;
\end{aligned}$$

and

$$\begin{aligned}
\tilde{\mu}_{ip}(\tilde{q}_{ip}, x^i, N) &\in Q_{ip}; \\
\mu_{ip}(q_{ip}, a) &\in \tilde{Q}_{ir}; \\
\tilde{\mu}_{ir}(\tilde{q}_{ir}, x^j, N) &\in Q_{ir}; \\
\mu_{ir}(q_{ir}, a) &\in \tilde{Q}_{ip} \ .
\end{aligned}$$

This machine maintains the role distinction and makes transition twice within each period - once after the bargaining and once after the disagreement game.[11] As a notational convention, we shall use $\tilde{q}_i$ to denote a state used by a sub-machine that plays the bargaining part of the negotiation game to distinguish it from q_i, a state associated with a sub-machine that plays the disagreement game.

As before $H^t_{ik}(= H^t_{jl})$ refers to the set of t-period histories. Here, we also denote the set of all possible histories at a disagreement game of period t in which i plays role k as $\tilde{H}^t_{ik} = H^t_{ik} \times \{(x, N) \mid \forall x \in \triangle^2\}$. Also, define $\tilde{H}^\infty_{ik} = \cup^\infty_{t=1} \tilde{H}^t_{ik}$.

A minimal machine in the 4SM specification corresponds to an alternative notion of stationarity. It implements a strategy f_i such that

- $f_i(h) = f_i(h') \; \forall h, h' \in H^\infty_{ip}$ and $\forall h, h' \in \tilde{H}^\infty_{ik}$ $(k = p, r)$; and
- $f_i(h, x^j) = f_i(h', x^j) \;\; \forall h, h' \in H^\infty_{ir}, \forall x^j \in \triangle^2$.

The definition of complexity captured by the size of a machine in the 4SM specification also needs to be modified. The size of a 4SM is measured by the cardinality $\sum_k |\tilde{Q}_{ik}| + \sum_k |Q_{ik}|$. Define $F_{ik}(f_i) = \{f_i|h : h \in H^\infty_{ik}\}$ as before and introduce $\tilde{F}_{ik}(f_i) = \{f_i|h : h \in \tilde{H}^\infty_{ik}\}$ to indicate the set of continuation strategies at a disagreement game of period t when i plays role k. It is straightforward to extend Proposition 1 to show that, for any $f_i \in F_i$, $\sum_k |F_{ik}(f_i)| + \sum_k |\tilde{F}_{ik}(f_i)|$ corresponds to the size of the smallest 4SM implementing f_i.

Given this foundation, analyzing the machine game in the 4SM specification is analogous to the previous 2SM case (though a little more cumbersome expositionally). Any NEMc profile in 4SM must be by definition a Nash equilibrium of the underlying negotiation game and every state belonging to an equilibrium 4SM must appear on the equilibrium path (Lemma 1).

The following three Lemmas correspond to Lemmas 2, 3, and 4 respectively. (We omit some of the proofs.) Note that while the game is being played bargaining alone does not generate any payoffs. Thus, $\pi^t_i(\cdot)$ equally represents i's continuation payoff at every subgame within the period (on the equilibrium path).

Lemma 7 *Assume that $M^* = (M^*_1, M^*_2)$ is a NEMc in the 4SM specification with $c \geq 0$. Then, $\forall i, j$ we have:*

$$\begin{aligned} &(i) \quad \pi^\tau_i(M^*) = \max_{f_i} \pi_i(f_i, M^*_j(\tilde{q}^\tau_j)) \;\; \forall \tau \leq T(M^*) \\ &(ii) \quad \pi^\tau_i(M^*) = \max_{f_i} \pi_i(f_i, M^*_j(q^\tau_j)) \;\; \forall \tau < T(M^*) \end{aligned}$$

[11] We can also construct a machine in which transition occurs at each decision node of the stage game. Six sub-machines will then be required (some of which will in fact serve only to make transition and not output). There are several other ways to divide each stage. But we conjecture that as long as we keep the role distinction for the bargaining part the central results will remain irrespective of the machine specification.

where $M_j^*(\tilde{q}_j^\tau)$ and $M_j^*(q_j^\tau)$ are the machines that are identical to M_j^* except that they start with the sub-machine which operates in the bargaining and disagreement game of period τ respectively with the initial states $\tilde{q}_j^\tau(\equiv \tilde{q}_j^\tau(M^*))$ and $q_j^\tau(\equiv q_j^\tau(M^*))$.

Lemma 8 *Assume that $M^* = (M_1^*, M_2^*)$ is a NEMc in the 4SM specification with $c \geq 0$. Then, $\forall i, j$ and $\forall t, t' \leq T(M^*)$, we have the following:*

$$\textit{if } \tilde{q}_j^t = \tilde{q}_j^{t'} \textit{ or } q_j^t = q_j^{t'}, \textit{ then } \pi_i^t(M^*) = \pi_i^{t'}(M^*) \ .$$

Lemma 9 *Assume that $M^* = (M_1^*, M_2^*)$ is a NEMc in the 4SM specification with $c \geq 0$. Then, for any i and any $\tau \leq T(M^*)$, we have the following:*

$$\begin{array}{ll} (i) & \textit{if } \ \tilde{q}_i^\tau(M^*) \neq \tilde{q}_i^t(M^*) \ \forall t < \tau, \ \textit{then } \tilde{q}_j^\tau(M^*) \neq \tilde{q}_j^t(M^*) \ \forall t < \tau \\ (ii) & \textit{if } \ q_i^\tau(M^*) \neq q_i^t(M^*) \ \forall t < \tau, \ \textit{then } q_j^\tau(M^*) \neq q_j^t(M^*) \ \forall t < \tau \ . \end{array}$$

Proof. (i) Suppose not. So, there exists some $\tau \leq T$ such that $\tilde{q}_i^\tau \neq \tilde{q}_i^t \ \forall t < \tau$ and $\tilde{q}_j^\tau = \tilde{q}_j^{\tau'}$ for some $\tau' < \tau$. By Lemma 8, $\pi_i^\tau = \pi_i^{\tau'}$.

Let $M_i^* = \{\tilde{M}_{ip}^*, M_{ip}^*, \tilde{M}_{ir}^*, M_{ir}^*\}$ where, for $k = p, r$, $\tilde{M}_{ik}^* = (\tilde{Q}_{ik}, \tilde{q}_{ik}^{1*}, \tilde{\lambda}_{ik}^*, \tilde{\mu}_{ik}^*)$ and $M_{ik}^* = (Q_{ik}^*, q_{ik}^{1*}, \lambda_{ik}^*, \mu_{ik}^*)$.

But then, consider i using another machine $M_i' = \{\tilde{M}_{ip}', M_{ip}', \tilde{M}_{ir}', M_{ir}'\}$ where, for $k = p, r$, $\tilde{M}_{ik}' = (\tilde{Q}_{ik}', \tilde{q}_{ik}^{1\prime}, \tilde{\lambda}_{ik}', \tilde{\mu}_{ik}')$ and $M_{ik}' = (Q_{ik}', q_{ik}^{1\prime}, \lambda_{ik}', \mu_{ik}')$. This machine is identical to M_i^* except that:

- $\tilde{q}_i^\tau$ is dropped; and
- the transition function is such that $\mu_{il}(q_{ik}^{\tau-1}, e^{\tau-1}) = \tilde{q}_i^{\tau'}$ $(k \in \{p, r\})$.

Since $\tilde{q}_i^\tau \neq \tilde{q}_i^t \ \forall t < \tau$, this preserves the outcome path up to $\tau - 1$ while making the path between τ' and $\tau - 1$ repeat from τ on. Similarly to the proof of Lemma 4 above, we can show that this will not change i's payoff. But, since $\tilde{q}_i^\tau$ has been dropped, $M_i^* \succ M_i'$. We thus have a contradiction against NEMc.

(ii) This part can be proven similarly to (i) above. ‖

Using these Lemmas, it is straightforward to extend the agreement results in Section 4.2 to the 4SM case. If a NEMc outcome under this alternative specification ends at some finite period, the pair of states occurring in the last period must be distinct. (Notice that the sub-machines used for playing the disagreement game will not be operating in the final period.)

Lemma 10 *Assume that $M^* = (M_1^*, M_2^*)$ is a NEMc in the 4SM specification with $c \geq 0$ and $T(M^*) < \infty$. Then, $\tilde{q}_i^t \neq \tilde{q}_i^T \ \forall t < T$ and $\forall i$.*

Proof. Suppose not. So, suppose that $\tilde{q}_i^t = \tilde{q}_i^T$ for some i and some $t < T$. Let $z = (z_1, z_2) \in \triangle^2$ be the agreement at T. There are two possible cases to consider.

Case A: Player i is the proposer at T.
Define $\tau = \min\{t | \tilde{q}_i^t = \tilde{q}_i^T\}$. By Lemma 8, $\pi_j^\tau = \pi_j^T$. Since there is an agreement on z at T, we have $\pi_j^\tau = z_j$.

Now consider player j using another machine $M_j' = \{\tilde{M}_{jp}', M_{jp}', \tilde{M}_{jr}', M_{jr}'\}$ where, for $k = p, r$, $\tilde{M}_{jk}' = (\tilde{Q}_{jk}', \tilde{q}_{jk}^{1\prime}, \tilde{\lambda}_{jk}', \tilde{\mu}_{jk}')$ and $M_{jk}' = (Q_{jk}', q_{jk}^{1\prime}, \lambda_{jk}', \mu_{jk}')$. This machine is identical to M_j^* except that:

- $\tilde{q}_j^\tau$ is dropped (i.e. $\tilde{Q}_{jr}' = \tilde{Q}_{jr}^* \backslash q_j^\tau$); and
- the transition function is such that $\mu_{jp}'(q_j^{\tau-1}, e^{\tau-1}) = \tilde{q}_j^T$.

Since, by Lemma 9, $\tilde{q}_j^\tau$ (as does $\tilde{q}_i^\tau$ by definition) appears for the first time at τ on the original equilibrium path, this new machine (given M_i^*) generates an identical outcome path as the original machine M_j^* up to τ and then induces the agreement z at τ. We know $\pi_j^\tau = z_j$, and thus, it follows that $\pi_j(M_i^*, M_j') = \pi_j(M_i^*, M_j^*)$. But since $\tilde{q}_j^\tau$ is dropped, $M_j^* \succ M_j'$. This contradicts NEMc.

Case B: Player i is the responder at T.
We can show contradiction similarly to Case A above. ||

Again, this implies that the agreement must occur within the first stage of the game; otherwise the states in the final period can be "replaced" thereby yielding a saving on complexity cost. (We shall omit the proof of the following result. It is almost identical to that of Proposition 2.)

Proposition 5 *If $M^* = (M_1^*, M_2^*)$ is a NEMc in the 4SM specification with $c \geq 0$ and $T(M^*) < \infty$, then (i) $T(M^*) \leq 2$; (ii) M_1^* and M_2^* are minimal; and thus M^* is efficient.*

What we gain from using this alternative machine specification is in the case of perpetual disagreement. Specifically, the SPEMc results in Section 4.3 no longer depend on the discount factor. Let $\tilde{\Omega}^\delta(c)$ denote the set of SPEMc in 4SM given discount factor δ and complexity cost c.

Proposition 6 *Consider any $c \geq 0$, any $\delta \in (0,1)$, and any $M^* \in \tilde{\Omega}^\delta(c)$ such that $T(M^*) = \infty$. Then, for any $\epsilon > 0$, $\exists\ \tau < \infty$ such that $\sum_i \pi_i^\tau(M^*) > 1 - \epsilon$.*

Proof. Fix any ϵ. Consider any $c \geq 0$, any $\delta \in (0,1)$, and any $M^* = (M_1^*, M_2^*) \in \Omega^\delta(c)$ such that $T(M^*) = \infty$. As before, let $M_i^* = \{\tilde{M}_{ip}^*, M_{ip}^*, \tilde{M}_{ir}^*, M_{ir}^*\}$ where, for $k = p, r$, $\tilde{M}_{ik}^* = (\tilde{Q}_{ik}, \tilde{q}_{ik}^{1*}, \tilde{\lambda}_{ik}^*, \tilde{\mu}_{ik}^*)$ and $M_{ik}^* = (Q_{ik}^*, q_{ik}^{1*}, \lambda_{ik}^*, \mu_{ik}^*)$.

Define η such that $0 < \eta < \epsilon$. Define also

$$t_{ik} = \{t |\ i \text{ plays role } k\}$$

and

$$\tau = \min\{t |\ \pi_i^t + \eta > \pi_i^{t'}\ \forall t, t' \in t_{ir}\}\ .$$

Notice that, given M_j^*, if j offers $(\pi_i^\tau + \eta, 1 - \pi_i^\tau - \eta) \in \triangle^2$ at any $t \in t_{ir}$, i must accept. Since

- every state of M_j^* appears on the equilibrium path of M^* (Lemma 1)
- now transition also occurs at the end of bargaining within each period
- $\pi_i^t = \max_{f_i} \pi_i(f_i, M_j^*(q_j^t))\ \forall t$ (Lemma 7)

(given M_j^*) the maximum continuation payoff i can obtain if he rejects such offer is equal to $\sup_{t \in t_{ir}} \pi^t$ which is less than $\pi_i^\tau + \eta$.

Consider now player j using another machine $M_i' = \{\tilde{M}_{ip}', M_{ip}', \tilde{M}_{ir}', M_{ir}'\}$ where, for $p = k, r$, $\tilde{M}_{ik}' = (\tilde{Q}_{ik}', \tilde{q}_{ik}^{1\prime}, \tilde{\lambda}_{ik}', \tilde{\mu}_{ik}')$ and $M_{ik}' = (Q_{ik}', q_{ik}^{1\prime}, \lambda_{ik}', \mu_{ik}')$. This machine is identical to M_j^* except for the output function which is such that $\tilde{\lambda}_{jp}'(\tilde{q}_j^\tau, \emptyset) = (\pi_i^\tau + \eta, 1 - \pi_i^\tau - \eta)$.

Now, note that $\tilde{q}_j^\tau \neq \tilde{q}_j^t\ \forall t < \tau$. Otherwise, $\pi_i^t = \pi_i^\tau$ by Lemma 8, which contradicts the definition of τ. Thus, (M_i^*, M_j') would end the game at τ.

Since M^* is a SPEMc, it must be that $\pi_j^\tau \geq 1 - \pi_i^\tau - \eta$, implying that $\pi_i^\tau + \pi_j^\tau \geq 1 - \eta$. But we fixed $\eta < \epsilon$, and thus, at $\tau < \infty$ we have $\sum_i \pi_i^\tau > 1 - \epsilon$ as in the claim. $\|$

Proposition 7 *For any $c > 0$ and any $\delta \in (0,1)$, every $M^* \in \tilde{\Omega}^\delta(c)$ is efficient (and stationary) with $T(M^*) \leq 2$.*

Proof. Suppose not. So, consider any $c > 0$, any $\delta \in (0,1)$, and any $M^* \in \tilde{\Omega}^\delta(c)$; suppose $T(M^*) = \infty$.

Define η such that $0 < \eta < c$, and also

$$t_{ik} = \{t |\ i \text{ plays role } k\}$$

and

$$\tau = \min\{t |\ \pi_i^t + \eta > \pi_i^{t'}\ \forall t, t' \in t_{ir}\}$$

as before.

Consider now player j using another machine $M'_i = \{\tilde{M}'_{ip}, M'_{ip}, \tilde{M}'_{ir}, M'_{ir}\}$ where $\tilde{M}'_{ik} = (\tilde{Q}'_{ik}, \tilde{q}^{1\prime}_{ik}, \tilde{\lambda}'_{ik}, \tilde{\mu}'_{ik})$ and $M'_{ik} = (Q'_{ik}, q^{1\prime}_{ik}, \lambda'_{ik}, \mu'_{ik})$.

This machine is identical to M^*_j except that:

- (similarly to the deviation in the proof of the previous proposition) the output function $\tilde{\lambda}'_{jp}(\tilde{q}^\tau_j, \emptyset) = (\pi^\tau_i + \eta, 1 - \pi^\tau_i - \eta)$; and additionally
- q^τ_j is dropped (i.e. $Q'_{jp} = Q^*_{jp} \backslash q^\tau_j$).

First note that we have $\tilde{q}^\tau_j \neq \tilde{q}^t_j$ and $q^\tau_j \neq q^t_j$ $\forall t < \tau$. Otherwise, $\pi^t_i = \pi^\tau_i$ by Lemma 8, which contradicts the definition of τ. Thus, dropping q^τ_j would not affect the outcome up to τ when the deviation would end the game (before reaching the disagreement game stage of the period).

Since $\pi^\tau_i \leq 1 - \pi^\tau_j$, j's loss from such deviation cannot be greater than η. But the new machine M'_j has one less state than M^*_j and there is also a saving in complexity cost by c. We fixed $\eta < c$ and thus the deviation is profitable. This contradicts the proposed SPEMc; therefore, $T(M^*) < \infty$. It then follows from Proposition 5 that $T(M^*) \leq 2$ and M^* is efficient and stationary. $\|$

6 Conclusion

When players care for complexity of a strategy as well as payoffs, the negotiation game can only display equilibria that are *efficient*. Thus, complexity and bargaining together offer an explanation for co-operation in two-person repeated interactions.

Independently of complexity cost, discount factor and the choice of machine specification, the negotiation game cannot have a NEMc in which an agreement takes place after delay beyond the first stage. If an agreement were to be part of an equilibrium outcome, then it must be so in the very first stage of the game, and the associated strategy profile must be stationary. Consequently, any SPEMc that induces an agreement must be efficient.

In fact, if complexity cost is strictly positive (and also discount factor is sufficiently close to one when we have the two sub-machine specification) there cannot be any other type of SPEMc outcome however small that complexity cost is. Thus, we have a very strong selection result in this case. If complexity cost is zero, and hence we have lexicographic preferences, it is also possible to have an equilibrium in which disagreement persists indefinitely. But this case still has to be (almost) efficient in the long run. It also follows here that perpetual disagreement cannot occur in cases where disagreement is strictly dominated by agreement.

There are several channels to further generalize the analysis in this paper. Especially, we can reinforce the repeated game flavor of the negotiation game by considering a broader set of payoffs that can be associated with bargaining and agreement. We can, for instance, let the space of offers be some arbitrary set $P \subset R^2$ such that $u(a) \subseteq P$ for all $a \in A$, thereby allowing an offer to be any (inefficient) disagreement game payoff vector as well as a partition of the maximum surplus available. We conjecture that complexity will still select the efficient outcomes in this case.

7 Appendix A: Relegated Proofs

Proof of Proposition 1. Let $M_i = \{M_{ip}, M_{ir}\}$ be implementation of some strategy f_i where $M_{ik} = (Q_{ik}, q^1_{ik}, \lambda_{ik}, \mu_{ik})$ for $k = p, r$.

First, we show that $|Q_{ik}| \geq |F_{ik}(f_i)|\ \forall k$.

For any $q_i \in Q_{ik}$ and $k = p, r$, let $M_i(q_i) = \{M_{ik}(q_i), M_{il}\}$ be the machine that is identical to M_i except that

- it starts with the sub-machine M_{ik}; and
- $M_{ik} = (Q_{ik}, q_i, \lambda_{ik}, \mu_{ik})$.

Note that for every $\bar{f}_i \in F_{ik}(f_i)$ and $k = p, r$, there exists some $h \in H^\infty_{ik}$ such that $\bar{f}_i = f_i|h$. Now define a function $\Gamma_{ik} : Q_{ik} \to F_{ik}(f_i)$ such that $\Gamma_{ik}(\bar{q}_i)$ is the strategy implemented by $M_i(\bar{q}_i)$ for any $\bar{q}_i \in Q_{ik}$. It then follows that for every $\bar{f}_i \in F_{ik}(f_i)$, there must exist a distinct state $\bar{q}_i \in Q_{ik}$ such that $\Gamma_{ik}(\bar{q}_i) = \bar{f}_i$. Simply let $\bar{q}_i = q_i(h)$ (as defined inductively above) where h is the history such that $\bar{f}_i = f_i|h$.[12]

Second, we show that there exists a machine implementation of f_i which only uses $F_{ik}(f_i)$ and $F_{il}(f_i)$ as the set of states for its sub-machines.

Define $\bar{M}_i = \{\bar{M}_{ip}, \bar{M}_{ir}\}$ such that, for $k = p, r$, $\bar{M}_{ik} = (F_{ik}(f_i), f^1_{ik}, \bar{\lambda}_{ik}, \bar{\mu}_{ik})$ where

- $f^1_{ik} \in F_{ik}(f_i)$ is the initial state and if i plays role k at the initial history then $f^1_{ik} = f_i$;
- for any $\bar{f}_i \in F_{ik}(f_i)$ and $d \in D_{ik}$, $\bar{\lambda}_{ik}(\bar{f}_i, d) = \bar{f}_i(\emptyset, d)$ where $\emptyset$ is the empty history;

[12] It is critical here that each sub-machine uses its own distinct set of states. Otherwise, a single state can be used to activate two distinct continuation strategies, one in each role.

- $\bar{\mu}_{ik}(\bar{f}_i, e) = \bar{f}_i|h, e$ for any $h \in H_{ik}^{\infty}$ and $e \in E$.

This machine has $\sum_k |F_{ik}(f_i)|$ states (each k sub-machine with $|F_{ik}(f_i)|$ states) and implements f_i. $\|$

8 Appendix B: An Alternative Equilibrium Concept

The following defines the general preference ordering over machine profiles proposed by Abreu and Rubinstein [1] (AR).

Definition 10 *Let $\succ_i^s$ (and $\sim_i^s$) denote player i's preference ordering over the set of machines profiles. For any pair of machine profiles $M = (M_i, M_{-i})$ and $M' = (M_i', M_{-i}')$, we have $M \succ_i^s M'$ if one of the following holds:*

$$\begin{array}{ll} (i) & \pi_i(M_i, M_{-i}) > \pi_i(M_i', M_{-i}') \textit{ and } ||M_i|| \leq ||M_i'|| \\ (ii) & \pi_i(M_i, M_{-i}) \geq \pi_i(M_i', M_{-i}') \textit{ and } ||M_i|| < ||M_i'|| \ . \end{array}$$

A Nash equilibrium can then be written in terms of machines that are most preferred against each other.

Definition 11 *A machine profile $M^* = (M_1^*, M_2^*)$ constitutes a Nash equilibrium of the machine game (NEM) if, $\forall i$, there exists no M_i' such that*

$$(M_i', M_{-i}^*) \succ_i^s (M_i^*, M_{-i}^*) \ .$$

The following Lemma extends Lemma 168.2 in Osborne and Rubinstein [15] (also part (a) of AR's Theorem 1) to the negotiation game. Any NEM must be such that each player's machine uses an equal number of states, and consequently, must correspond to a Nash equilibrium of the negotiation game.

Lemma 11 *Suppose that A is compact and $u_i(\cdot)$ is continuous for all i. Then, if $M^* = (M_1^*, M_2^*)$ is a NEM, we have*

$$\begin{array}{ll} (i) & ||M_1^*|| = ||M_2^*||; \textit{ and} \\ (ii) & \pi_i(M^*) = \max\limits_{f_i} \pi_i(f_i, M_{-i}^*) \ \ \forall i. \end{array}$$

Proof. Consider machines in the 2SM specification. (The 4SM case can be treated similarly.)

Let S_{ik} define the set of player i's *one-period strategies* in the extensive form game that he plays in role $k \in \{p, r\}$ every other period of the negotiation game. We denote its element by $s_{ik} \in S_{ik}$. With slight abuse of notation, let $u_i(s_{ik}, s_{jl})$ denote player i's (one-period) payoff given the pair of strategies.

(i) Fix player j's machine $M_j = \{M_{jp}, M_{jr}\}$ where, for $k = p, r$, $M_{jk} = (Q_{jk}, q^1_{jk}, \lambda_{jk}, \mu_{jk})$. Then, suppose that player i solves his dynamic optimization problem for the machine game ignoring complexity such that

$$\max_{\{s^t_{ik}\}^{\infty}_{t=1}} \sum_{t=1}^{\infty} \delta^{t-1} u_i(s^t_{ik}, M_j(q^t_j)) \tag{13}$$

where q^t_j is defined inductively as before (by the transition functions of j's machine).

This is a (deterministic) Markovian problem with the transition of states given by the other player's machine, and therefore, i's optimal action(s) in each period depends at most on the state of the other player's machine and the partial history within the period. (For finite state space, this statement is established by the Blackwell's theorem. For a general (countable) state space, the case we consider in the paper, see Hinderer [12] and the references therein. Also, such solution exists if S_{ip} and S_{ir} are compact (which is true if A is compact) and $u_i(\cdot)$ is continuous.) Let $s^*_{ik}(q, d)$ denote the optimal action for player i in role k given $q \in Q_{jl}$ and $d \in D_{ik}$.

Now, consider a machine for player i $M_i = \{M_{ip}, M_{ir}\}$ defined by, for $k = p, r$,

- $Q_{ik} = Q_{jl}$;
- $q^1_{ik} = q^1_{jl}$;
- $\lambda_{ik}(q, d) = s^*_{ik}(q, d)$ $\forall q \in Q_{ik}$ and $\forall d \in D_{ik}$; and
- $\mu_{ik}(q, e) = \mu_{jl}(q, e)$ $\forall e \in E$.

This machine solves the maximization problem (13) above using only the states used by the other player's machine.

Thus, if $M^* = (M^*_1, M^*_2)$ is a NEM profile, then $||M^*_i|| \leq ||M^*_{-i}||$ $\forall i$. It follows that $||M^*_1|| = ||M^*_2||$.

(ii) This follows from part (i). $\|$

Lemma 11 connects our notion of NEMc (Definition 7) with AR's equilibrium notion (Definition 11). Effectively, both definitions take the set of Nash equilibria of the negotiation game and select outcomes that capture some measure of "trade-off" between payoffs and complexity. In this sense, the equilibrium notions used in this paper closely parallel those of AR.

References

[1] Abreu, D., and A. Rubinstein (1988): "The Structure of Nash Equilibria in Repeated Games with Finite Automata," *Econometrica*, 56, 1259-82.

[2] Binmore, K., M. Piccione, and L. Samuelson (1998): "Evolutionary Stability in Alternating-Offers Bargaining Games," *Journal of Economic Theory*, 80, 257-91.

[3] Bloise, G. (1998): "Strategic Complexity and Equilibrium in Repeated Games," unpublished doctoral dissertation, University of Cambridge.

[4] Busch, L-A., and Q. Wen (1995): "Perfect Equilibria in a Negotiation Model," *Econometrica*, 63, 545-65.

[5] Chatterjee, K., and H. Sabourian (1999): "N-Person Bargaining and Strategic Complexity," *mimeo*, University of Cambridge.

[6] Chatterjee, K., and H. Sabourian (2000): "Multiperson Bargaining and Strategic Complexity," *Econometrica*, 68, 1491-1509.

[7] Fernandez, R., and J. Glazer (1991): "Striking for a Bargain Between Two Completely Informed Agents," *American Economic Review*, 81, 240-52.

[8] Fudenberg, D., and J. Tirole (1991): *Game Theory*, Cambridge, Massachusetts: MIT Press.

[9] Gale, D., and H. Sabourian (2003): "Complexity and Competition, Part I: Sequential Matching," *mimeo*, University of Cambridge.

[10] Gale, D., and H. Sabourian (2003): "Complexity and Competition, Part II: Simultaneous Endogenous Matching," *mimeo*, University of Cambridge.

[11] Haller, H., and S. Holden (1990): "A Letter to the Editor on Wage Bargaining," *Journal of Economic Theory*, 52, 232-6.

[12] Hinderer, K. (1970): *Foundations of Non-stationary Dynamic Programming with Discrete Time Parameter*. Berlin: Springer-Verlag.

[13] Kalai, E., and W. Stanford (1988): "Finite Rationality and Interpersonal Complexity in Repeated Games," *Econometrica*, 56, 397-410.

[14] Nash, J. (1953): "Two-Person Cooperative Games," *Econometrica*, 21, 128-40.

[15] Osborne, M., and A. Rubinstein (1994): *A Course in Game Theory*, Cambridge, Massachusetts: MIT Press.

[16] Pearce, D. (1992): "Repeated Games: Cooperation and Rationality," in *Advances in Economic Theory, Sixth World Congress*, ed. J-J. Laffont, Cambridge: Cambridge University Press.

[17] Piccione, M. (1992): "Finite Automata Equilibria with Discounting," *Journal of Economic Theory*, 56, 180-93.

[18] Piccione, M., and A. Rubinstein (1993): "Finite Automata Play a Repeated Extensive Game," *Journal of Economic Theory*, 61, 160-8.

[19] Rubinstein, A. (1982): "Perfect Equilibrium in a Bargaining Model," *Econometrica*, 50, 97-109.

[20] Rubinstein, A. (1986): "Finite Automata Play the Repeated Prisoner's Dilemma," *Journal of Economic Theory*, 39, 83-96.

[21] Sabourian, H. (2003): "Bargaining and Markets: Complexity and the Competitive Outcome," forthcoming in *Journal of Economic Theory.*