Global Climate Cooperation without Sanctions*

Akira Okada[†]

February, 2022

Abstract

The negative result on the level of cooperation in the literature of international environmental agreements is caused by the static nature of a two-stage participation game. We present a dynamic coalition formation game where coalitions are irreversible, and countries negotiate transfer schemes to attract non-participants. We show that if countries are sufficiently patient, then there exists a Markov perfect equilibrium where the grand coalition forms immediately. We further show that if countries are impatient, the grand coalition gradually forms in finitely many rounds for every Markov perfect equilibrium under a super-additivity condition. The results hold for any transfer scheme satisfying individual rationality and coalitional efficiency.

JEL classification: C72, H41, Q54
Keywords: international environmental agreement; non-cooperative coalitional bargaining; stable coalition; transfer; participation

*I would like to thank seminar participants at Hitotsubashi University, Konan University, and Waseda University for their useful comments. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

[†]Professor Emeritus, Hitotsubashi University, 2-1 Naka, Kunitachi, Tokyo 186-8601, Japan. E-mail: aokada313@gmail.com

1 Introduction

The notion of a stable coalition has been extensively studied in the literature of oligopoly (D'Aspremont et al. 1983) and international environmental agreements (IEAs) (Hoel 1992, Carraro and Siniscalco 1993, Barrett 1994). A coalition of countries joining an IEA is stable if it has the two properties: no single participant has the incentive to leave it (internal stability), and no single non-participant has the incentive to join it (external stability). The general message of the literature on IEAs is pessimistic concerning the level of cooperation (Hovi et al. 2015 and Buchholz and Sandler 2021). A stable coalition is typically small, with only two participants in some cases. The grand coalition of all countries forms only if the potential benefit of climate cooperation is small. This negative result, termed as the "Paradox of International Agreements (IAs)" (Kolstad and Toman 2005), is caused by the incentive of free-riding. Rational and self-interested countries are motivated to free-ride on an IEA agreed by other countries.

To overcome the paradox, some authors propose trade sanctions to free-riders (Barrett 1997 and Nordhaus 2015 among others). While sanction, if successful, is an effective tool to prevent free-riding, it has some undesirable properties. Its success is not guaranteed at all. Trade sanctions have a risk of a trade war. Even a successful sanction causes another incentive problem that countries may be motivated not to participate in it, free-riding on the sanction implemented by other countries. This incentive problem is called the "second-order dilemma" of public goods (Ostrom 1990).

In this paper, we consider whether and how climate cooperation is possible without sanctions. In particular, we explore a negotiation process under which the grand coalition of countries can form for climate cooperation.

When countries are (ex ante) identical, it is well-known that a stable coalition of countries is a Nash equilibrium (more precisely, a subgame perfect equilibrium) of a two-stage game of coalition formation. In the first stage, all countries independently decide whether to participate in a coalition. In the second stage, all countries choose their abatement levels. It is assumed that a coalition of countries acts as a "single player" to maximize the total welfare of its members, given the abatement levels of non-participants. Non-participants maximize non-cooperatively their welfare, given the abatement levels of all other countries.

The paradox of international agreements is illustrated in a simple emission game. Every country i chooses an abatement $q_i \in [0, 1]$ and receives the welfare given by $b(q_1 + \cdots q_n) - cq_i$ where b and c represent marginal benefit and cost of abatement, respectively. If $b < c$, the optimal choice of every country is $q_i = 0$, regardless of other countries' choices. If the size of a coalition is greater than $\frac{c}{b}$, the optimal choice of it is the full abatement ($q_i = 1$), and otherwise the zero abatement. The size s^* of a stable coalition must be equal to the minimum integer which is greater than $\frac{c}{b}$. External stability holds since each county wants to free-ride. Internal stability also holds at s^* since if one participant opts out, then the smaller coalition of size $s^* - 1$ is not beneficial anymore, and it collapses. When the cost-benefit ratio $\frac{c}{b}$ is low, a stable coalition is small. The grand coalition is stable only when the ratio is so high that $n - 1 < \frac{c}{b} < n$. The stable grand coalition yields only a small amount of welfare, $bn - c$, to every country, being less than the marginal benefit b.

When countries are heterogeneous, those with high abatement costs may be worse off by joining a coalition than in the case of no cooperation, sacrificed by the maximization of the coalition surplus. To attract such reluctant non-participants to a coalition, participants have to propose a transfer scheme for them. The literature has emphasized the role of transfer for stable IEAs (Carraro and Siniscalco 1993 and 1998, and Carraro et al. 2006, for example). In the general case of heterogeneous countries, the two-stage game should be modified such that participants negotiate for forming a coalition with a transfer scheme after the first stage. As a collective choice rule, unanimous voting is natural and frequently employed in international negotiations. If a coalition is agreed on, it chooses the optimal abatement in the third stage. If the agreement fails, all countries choose their abatements non-cooperatively.

The timing of the game is critical. Although the profitability of a coalition depends on which countries participate, countries are uncertain about it when they decide whether to join in the first stage. For this reason, the participants should be able to decide whether to agree on the coalition in the second stage, after they know the composition of it. Selten (1973) presents a prototype of the three-stage game above for cartel formation of oligopolistic firms. In this paper, we consider a dynamic model of the three-stage game of coalition formation.

The negative result aforementioned on stable IEA's is caused by the static nature of the model that participation is once and for all. In contrast, international negotiation is a dynamic process where countries renegotiate an agreement many times. We present a dynamic game of coalition formation. Our analysis departs from the theory of repeated games. If countries play the coalition formation game over infinitely many rounds, then the whole game is just a standard repeated game. The celebrated folk theorem shows that every individually rational outcome can be attained in a subgame perfect equilibrium if countries are sufficiently patient (Fudenberg and Maskin 1986). In our context, the grand coalition can be formed by a suitable equilibrium strategy. An example of it is the grim-trigger strategy that punishes a non-participant forever. The folk theorem, however, does not provide a complete answer to the climate cooperation problem. A repeated game has multiple equilibria. The grand coalition is just one of them. More critically, it has been unanswered how countries can reach an agreement to cooperate. Every country may have the incentive not to participate in the grim-trigger strategy if the remaining countries cooperate. The participation problem should be solved in the folk theorem. We shall discuss the relation of this paper to the recent literature that attempts to explain full cooperation by a weakly renegotiation-proof equilibrium (Farrell and Maskin 1989) in Section 4.

The dynamic coalition game in this paper has the critical property that a coalition is *irreversible*. Once countries participate in a coalition, they cannot leave it. In the next round, only non-participants decide whether to join it. Both incumbent and new participants negotiate for forming an enlarged coalition. The incumbent members employ a transfer scheme to attract non-participants. The negotiation process is repeated until the grand coalition forms.

We prove two results for the formation of the grand coalition. First, if countries are sufficiently patient, then there exists a Markov perfect equilibrium where the grand coalition forms in the first round of negotiations (Theorem 1). Second, for every discount factor of future welfare and every Markov perfect equilibrium, the grand coalition gradually forms in finitely many rounds under a supper-additivity condition of welfare functions (Theorem 2). The results hold for any transfer scheme satisfying individual rationality and coalitional efficiency.

The intuition for the results is as follows. In Theorem 1, we construct a Markov perfect equilibrium where all non-participants join a current coalition, no matter what it is. The grand coalition forms in the first round on the equilibrium play. If any country does not participate, all other countries negotiate for forming the coalition in the second stage. There are two cases to consider: (i) negotiation fails, and (ii) negotiation succeeds. In case (i), no coalition prevails in the first round, and the grand coalition with the equilibrium allocation will form in the second round. Non-participation causes only the delay of the grand coalition, and thus it is not beneficial to the non-participant. In case (ii), the sub-coalition forms in the first round, and it will be expanded to the grand one with a new allocation in the second round. When countries are patient, they are concerned about their future welfare. All countries except the non-participant form the coalition in the first round only if they are better off in the grand one with a new allocation in the second round than in that with the equilibrium allocation. Since the surplus of the grand coalition is constant, there is a trade-off of the welfare between participants and non-participant in it. Therefore, the non-participant is worse off in the grand coalition formed in the second round than in the equilibrium one formed in the first round. In either case, the non-participant is not better off.

Theorem 2 formalizes the well-known idea of the Coase "theorem" (Coase 1960). If the welfare function is supper-additive, then an enlarged coalition increases the total welfare of incumbent and new participants. Rational and self-interested countries renegotiate an inefficient allocation towards a Pareto-improving one in every Markov perfect equilibrium. In this way, Coasian bargaining internalizes externality. The grand coalition gradually forms in the renegotiations, independent of the discount factor of future welfare.

There exist a growing number of works on dynamic coalitional bargaining involving renegotiation without externality (Seidmann and Winter 1998, Okada 2000) and with externality (Gomes 2005, Hyndman and Ray 2007 and Ray 2007). These works employ a proposal-based model in which a proposer selected by a predetermined rule chooses a coalition with an allocation, and all members respond to it. If they all accept the proposal, the coalition forms, and it will be renegotiated towards a Pareto-improving one in the next round. Assuming that coalitions are irreversible, it has been shown that the gradual formation of the grand coalition holds in a supper-additive game. In the existing models, players can join a coalition only if they are invited by the proposer. Since the models describe the dynamic coalitional bargaining with exclusive membership, they are not suitable to address the Paradox of International Agreements with open membership. Unlike Theorem 1, the Markov perfect equilibrium where the grand coalition immediately forms does not exist in the proposal-based model when players are patient (Okada 2000). The reason for this is that the possibility of renegotiation decreases the acceptance thresholds of responders, and the proposer strategically chooses a small coalition first to exploit the surplus. The grand coalition equilibrium exists, provided that players are impatient. The extreme case is that countries have zero discount factors for future payoffs, which corresponds to the ultimatum bargaining.

In the companion paper (Okada 2022), we present a dynamic coalitional bargaining model for the Nash bargaining solution in an n-person strategic-form game and prove the results similar to those in this paper. We generalize the results to any allocation rule satisfying individual rationality and coalitional efficiency. The advantage of the approach of this paper is that the efficiency result does not depend on a specific bargaining solution for the allocation problem. Kováč and Schmidt (2021) consider a case of our dynamic coalition game under the restrictive conditions that countries are symmetric, the game ends once a coalition forms, however small it is, and transfer is not allowed. They show the existence of the grand coalition Markov perfect equilibrium. They need a specific membership rule to select a Nash equilibrium of the participation game and assume several properties of welfare functions to solve the model. Karp and Simon (2013) show that the Paradox of IAs is not robust to specifications of benefit and cost functions and identify conditions under which the equilibrium coalition size can be large. In contrast to their static approach, our dynamic approach shows that global climate cooperation can be achieved, provided that it is socially optimal.

We organize the remainder of the paper as follows. Section 2 presents the model. Section 3 provides the results. Section 4 discusses the implications and limitations of the results. Section 5 concludes.

2 The Model

There are n countries that emit greenhouse gases (GHGs). Let $N = \{1, \cdots, n\}$ denote the set of countries. Each country i chooses an abatement level $q_i \in [0, \bar{e}_i]$ of emissions where $\bar{e}_i$ is a baseline emission. For an abatement vector $q = (q_1, \cdots, q_n)$, the welfare of every country i is given by $\pi_i(q_1, \cdots, q_n)$. The literature of IEAs traditionally represent the welfare of a country in terms of cost and benefit so that $\pi_i(q_1, \cdots, q_n) = B_i(Q) - C_i(q_i)$ where B_i is a benefit function, C_i is a cost function, and $Q = \sum_{i=1}^{n} q_i$ is the total abatement. The result of this paper does not depend on a specific form of the welfare function. It is often assumed that countries are (ex ante) identical in that they have the identical benefit and cost functions. We do not need to assume it. In what follows, countries may be heterogeneous. We call a strategic situation between n countries regarding abatement decisions an *emission game*, formally defined by a triple (N, Σ, π) where $\Sigma = [0, \bar{e}_1] \times \cdots \times [0, \bar{e}_n]$ and $\pi = (\pi_1, \cdots, \pi_n)$.

Consider first that n countries choose non-cooperatively their abatements, that is, every country i chooses its abatement q_i to maximize its own welfare, given the abatements chosen by all other countries. The play of an emission game is determined by a Nash equilibrium $q^* = (q_1^*, \cdots, q_n^*)$ such that q_i^* maximizes $\pi_i(q_i, q_{-i}^*)$ for every $i \in N$ where q_{-i}^* is the equilibrium abatement vector of all countries except i.

We next consider a situation where a coalition S of countries choose their abatements cooperatively. Following the literature of IEA's, we assume that the coalition S acts as a single-player to maximize the total welfare of the participants, given the abatements of non-participants. We further assume that countries' utility is transferable. Every non-participant behaves in the same manner as in a Nash equilibrium, and it maximizes its welfare, given all other countries. In this paper, we call such an equilibrium an S-*equilibrium* and denote it by $q^S = (q_1^S, \cdots, q_n^S)$.

For a coalition S, $q_S = (q_i : i \in S)$ denotes an abatement vector of all participants in S, and $q_{N \setminus S} = (q_i : i \notin S)$ denotes that of all non-participants. Formally, an S-equilibrium $q^S = (q_S^S, q_{N \setminus S}^S)$ of an emission game satisfies the two conditions: for every $q_S = (q_i : i \in S)$,

$$\sum_{i \in S} \pi_i(q_S^S, q_{N \setminus S}^S) \geq \sum_{i \in S} \pi_i(q_S, q_{N \setminus S}^S)$$

and, for every non-participant $j \notin S$ and every q_j,

$$\pi_j(q_j^S, q_{N\setminus\{j\}}^S) \geq \pi_j(q_j, q_{N\setminus\{j\}}^S).$$

In an S-equilibrium, the participants of S jointly choose the coalition-optimal abatement to maximize their total welfare. Since an S-equilibrium generally differs from a Nash equilibrium, some participants may have the incentive to deviate from it. It is customary in the literature of IEA's to assume that any optimal decision of a coalition is binding under some external enforcement mechanism. We also employ this approach and assume that the participants can enforce any group decision effectively, *provided that they agree upon it.* We shall discuss this point in Section 4.

There are two polar cases for the level of cooperation. The first case is that no country participates in a coalition ($S = \emptyset$). In this non-cooperation case, the $\emptyset$-equilibrium is equal to a Nash equilibrium. A remark may be helpful. If a coalition has only one participant, then the participant maximizes its welfare, given all other countries' abatements. It means that a single-member coalition results in the same outcome as in the Nash equilibrium. A coalition is substantial only if it has more than one participant. The second case is that all countries participate in the grand coalition N. In the full cooperation case, the N-equilibrium produces the socially optimal outcome that maximizes the total welfare of all countries.

To analyze the problem of coalition formation in the emission game, we assume the following standard assumptions in the literature of IEA's:

A1. For every coalition S, there exists a unique S-equilibrium q^S.

A2. The N-equilibrium uniquely maximizes the total welfare of all countries.

A1 is a technical assumption that makes the analysis of the emission game tractable. By definition, the grand coalition is socially optimal. A2 assumes that any other coalition S is not socially optimal. It means that when S forms, there exists some welfare transfer under which all countries are better off by forming the grand coalition N than S. Most cost-benefit models employed in the literature satisfy A1 and A2.

While we are mainly concerned with positive-externality games such as emission games, n-person prisoner's dilemma, and public goods games, the result of the paper can be applied to other games satisfying A1 and A2. Examples are international trade games with negative externality where global free trade maximizes the total welfare in the world and non-participants are worse off.

If all countries are identical, as is often assumed in the literature of IEA's, then every participant of every coalition S is better off in the S-equilibrium than in the Nash equilibrium. However, if countries are heterogeneous, this is not always the case. For example, countries with high abatement costs may be worse off in a coalition, sacrificed by the agreement to maximize the coalition welfare, and thus may be unwilling to participate. To increase participants, the coalition members can use welfare transfer schemes. To analyze whether the possibility of transfer schemes may lead to global climate cooperation, we assume that every coalition S can make the transfer among its members to give them the incentive to participate.

Let $\lambda(S) = (\lambda_i(S) : i \in S)$ be a transfer vector in S where each participant i receives a positive or negative welfare transfer $\lambda_i(S)$. Transfer $\lambda(S)$ should be self-financed so that

$$\sum_{i \in S} \lambda_i(S) = 0. \tag{1}$$

Under a transfer scheme $\lambda(S)$, the welfare of every participant $i \in S$ is given by $\pi_i(q^S) + \lambda_i(S)$. We denote by $\tilde{\pi}_i(q^S)$ the welfare of country $i \in S$ after transfer. For non-participant $i \notin S$, we simply set $\tilde{\pi}_i(q^S) = \pi_i(q^S)$.

Participants of a coalition S negotiate a transfer between them. An agreement of a transfer is assumed to be made by unanimous voting. If at least one participant rejects it, negotiation fails and thereafter, the Nash equilibrium q^* prevails. Thus, an agreeable transfer in S should satisfy *individual rationality* which requires that it makes all participants better off than in the Nash equilibrium. The participants negotiate how to allocate the coalition surplus $\sum_{j\in S}\pi_j(q^S)-\sum_{j\in S}\pi_j(q^*)$. Let α_i^S denote the share of the surplus that each participant i receives. Then, the welfare of participant i is given by

$$\pi_i(q^*)+\alpha_i^S(\sum_{j\in S}\pi_j(q^S)-\sum_{j\in S}\pi_j(q^*)),$$

and a transfer $\lambda_i(S)$ to participant i is given by

$$\lambda_i(S)=\pi_i(q^*)+\alpha_i^S(\sum_{j\in S}\pi_j(q^S)-\sum_{j\in S}\pi_j(q^*))-\pi_i(q^S).$$

The literature of game theory has extensively studied bargaining solutions, both in non-cooperative bargaining games and in cooperative games. The (symmetric) Nash bargaining solution and the equity allocation are the most well-known bargaining solutions which require that all participants should equally allocate the coalitional surplus, that is $\alpha_i^S=\frac{1}{|S|}$ where $|S|$ is the number of participants. Besides these bargaining solutions, the literature of IEA's has proposed various types of proportional allocations where the sharing ratios of participants are based on population, GDP, emission, abatement benefit, and costs, among others. To focus the analysis on the coalition formation problem, we fix an arbitrary transfer scheme for every coalition based on some bargaining solution. The results of this paper do not depend on the choice of a transfer scheme, provided that it satisfies *individual rationality* (all participants are better off in a coalition than when it fails) and *coalitional efficiency* (the coalition surplus is allocated efficiently).

We consider a dynamic process of coalition formation in an emission game. In the process, all countries play the following three-stage game over infinitely many rounds $t = 1, 2, \cdots$.

Stage 1 (participation stage): All countries that have not joined a coalition simultaneously decide whether to participate in it.

Stage 2 (negotiation stage): All (incumbent and new) participants decide whether to form the coalition with a predetermined allocation (or transfer) by unanimous voting. The coalition forms if and only if all participants agree to it (vote for YES).

Stage 3 (action stage): All n countries, both participants and non-participants, choose their abatements. If a coalition S forms in Stage 2, then the S-equilibrium q^S is played. S implements the agreed-upon transfer scheme. If a coalition does not form, then the Nash equilibrium is played.

Countries maximize their discounted welfare sum where $\delta < 1$ is a common discount factor of future welfare. The result of the paper can be easily extended to the case of heterogeneous discount factors.

A critical property of the model is that a coalition is *irreversible*. It means that once a coalition forms, all participants are committed to becoming members of the current and future ones (if any). Participants have no choice to leave coalitions in future negotiations. We discuss this property more in Section 4.

Formally, the dynamic process of coalition formation is described as follows. Let S_{t-1} be the coalition formed in round $t-1$. As the initial condition, we set $S_0 = \emptyset$. In the first stage of round t, all countries outside S_{t-1} decide independently whether to join S_{t-1}. Let P_t denote the set of new participants. The second stage of negotiation is played by all countries in S_{t-1} and P_t. They independently decide to accept or reject the extended coalition $S_t = S_{t-1} \cup P_t$. If negotiation succeeds, then S_t forms. Otherwise, $S_t = S_{t-1}$ prevails. In the third stage of actions, the S_t-equilibrium with the transfer prevails. By the rule of the game, a sequence of coalitions $\{S_t\}$ is weakly and monotonically increasing (in terms of set inclusion). If the negotiation succeeds, the coalition expands from S_{t-1} to S_t. Otherwise, the incumbent coalition S_{t-1} remains effective, and the S_{t-1}-equilibrium with the transfer prevails. Once the grand coalition N forms, the game ends substantially, and the N-equilibrium prevails forever. Let Γ denote the dynamic coalition game described above.

In the second stage, suppose that the incumbent coalition S_{t-1} and new participants P_t negotiate for the new coalition $S_t = S_{t-1} \cup P_t$. If the negotiation fails, then every country i receives the welfare $\tilde{\pi}_i(q^{S_{t-1}})$. Specifically, every incumbent participant $i \in S_{t-1}$ receives the welfare after transfer, $\pi_i(q^{S_{t-1}}) + \lambda_i(S_{t-1})$, agreed by S_{t-1}, and every non-participant i receives the free-riding welfare $\pi_i(q^{S_{t-1}})$. If the negotiation succeeds, then the coalitional surplus that countries produce by expanding S_{t-1} to S_t, denoted by $v(S_t|S_{t-1})$, is given by

$$v(S_t|S_{t-1}) = \sum_{i \in S^t} \pi_i(q^{S_t}) - \sum_{i \in S^t} \pi_i(q^{S_{t-1}}).$$

The participants of the new coalition S_t negotiate how to allocate $v(S_t|S_{t-1})$. To attract non-participants to the coalition, a transfer $\lambda(S_t|S_{t-1})$ with the self-financing condition (1) is used, depending on S_t and S_{t-1}. We denote the welfare of every participant i of S_t after transfer by $\tilde{\pi}_i(q^{S_t}|S_{t-1})$,

$$\tilde{\pi}_i(q^{S_t}|S_{t-1}) = \pi_i(q^{S_t}) + \lambda_i(S_t|S_{t-1}).$$

For it to be agreed upon by unanimous voting, a transfer $\lambda(S_t|S_{t-1})$ must satisfy the individual rationality condition for S_t,

$$\tilde{\pi}_i(q^{S_t}|S_{t-1}) \geq \tilde{\pi}_i(q^{S_{t-1}}), \forall i \in S_t. \tag{2}$$

If $v(S_t|S_{t-1}) < 0$, then there exists no individually rational transfer in S_t, and thus the negotiation in S_t fails. By A2, it holds that $v(N|S_{t-1}) > 0$. The grand coalition N can always produce a positive surplus.

In this paper, we consider a Markov perfect equilibrium (MPE) of Γ. Roughly speaking, an MPE is a subgame perfect equilibrium where countries' choices depend only on "payoff-relevant" histories. In Γ, an MPE satisfies the following three conditions. For every round t, (i) countries' participation decisions depend only on the coalition S_{t-1} formed in the last round $t-1$, and (ii) participants' voting choices depend only on S_{t-1} (incumbent participants) and P_t (new participants), and (iii) countries actions depend only on S_t where $S_t = S_{t-1} \cup P_t$ if negotiation succeeds in the second stage, and otherwise $S_t = S_{t-1}$. An important implication of an MPE is to rule out punishments against free-riders. In the standard model of an infinitely repeated emission game, an MPE is only the repetition of the one-shot Nash equilibrium where no cooperation is possible. In contrast, the stage game of Γ includes a coalition formation phase, and transfer schemes make cooperation possible in equilibrium.

Finally, we remark that as in the static game of voluntary participation, there exists a trivial MPE in Γ where no country participates in a coalition in each round. We eliminate such a trivial equilibrium from our analysis and consider under what condition the grand coalition of countries form.

3 The Results

In the dynamic coalition game Γ, the discounted welfare sum of country i from round t is decomposed as $\pi_i + \delta \times \Pi_i$ where π_i is the present welfare in round t and Π_i is the discounted welfare sum from round $t+1$, called the *continuation value*. In a Markov perfect equilibrium, the continuation value Π_i depends only on a current coalition. We denote by $\Pi_i(S_{t-1})$ the continuation value of country i from each round t where coalition S_{t-1} formed in round $t-1$.

The first result shows the existence of an MPE where the grand coalition immediately forms when countries are sufficiently patient.

Theorem 1. Under A1 and A2, there exists an MPE of Γ for every sufficiently large $\delta < 1$ where all countries participate in the grand coalition N in the first round and agree to the socially optimal abatement.

We explain the intuition for the theorem. We construct an MPE where all countries form the grand coalition N, no matter what incumbent coalition exists. In particular, all n countries participate in N in the first round. If any country i does not participate, then all other participants negotiate for forming the coalition $N/\{i\}$. The following two cases are possible:

Case 1. $N/\{i\}$ does not form in round 1, and N will do in round 2.

Case 2. $N/\{i\}$ forms in round 1, and it will expand to N in round 2.

Consider first case 1. The non-participation of country i causes only the delay of N, and it is not beneficial to i. Next consider case 2. The discounted welfare sum of every country $j \neq i$ is given by $\tilde{\pi}_j(q^{N/\{i\}}) + \frac{\delta}{1-\delta}\tilde{\pi}_j(q^N|N/\{i\})$. When the discount factor δ is sufficiently large, the members j of $N/\{i\}$ are more concerned about their future welfare $\tilde{\pi}_j(q^N|N/\{i\})$ than the present welfare $\tilde{\pi}_j(q^{N/\{i\}})$. For them to agree to $N/\{i\}$, it must be the case that $\tilde{\pi}_j(q^N|N/\{i\}) > \tilde{\pi}_j(q^N)$, that is, they must be better off by forming first $N/\{i\}$ and expanding it to N than by forming N immediately. However, since the total welfare of N is constant, this implies that non-participant i has the opposite preference so that i prefers to form N immediately. Thus, non-participation is not the optimal choice.

In the MPE in Theorem 1, the grand coalition N forms immediately on and off the equilibrium path. Thus, the equilibrium is *renegotiation-proof* in the sense that the strategy is not renegotiated towards a Pareto improving one, independent of the history of the game.

Theorem 1 tells us that when countries are concerned about their future welfare, all may agree to form the grand coalition N in the first round in Γ. The static model in the literature of IEA's is a case of Γ where the discount factor is zero or very small. According to the Paradox of International Agreements, Theorem 1 does not hold when countries are impatient.

Theorem 1 poses two further questions. Are there any other types of MPE where a small coalition forms in the first round? What happens when countries are not patient? One may conjecture that the participants could attract non-participants by designing a suitable transfer scheme, and thus that a small coalition gradually expands to the grand one N. In what follows, we show that this conjecture is true for every discount factor under a supper-additivity condition of countries' welfare. We assume the following.

A3. For every coalition S and every country $i \notin S$, it holds that

$$\sum_{j \in S \cup \{i\}} \pi_j(q^{S \cup \{i\}}) > \sum_{j \in S} \pi_j(q^S) + \pi_i(q^S).$$

The supper-additivity condition is standard in cooperative game theory. A3 means that if a non-participant joins a coalition, then the total welfare of all incumbent members and the new participant increases. Under A3, the coalition can attract a non-participant to join them with some transfer scheme.

Theorem 2. Assume that A1-A3 hold and that $\delta \in (0, 1)$. In every MPE of Γ, the grand coalition N gradually forms in at most $n - 1$ rounds.

Theorem 2 shows that for every discount factor of future welfare, every MPE has the property that the grand coalition gradually forms in finitely many rounds under the supper-additivity condition A3, provided that the participation in a coalition is irreversible. The key result in the theorem is that there exists at least one new participant in a coalition in every round unless the grand coalition forms. The intuition for this is as follows. On the contrary, suppose that there exist no new participants in the current coalition $S_t \neq N$. Then, the S_t-equilibrium q^{S_t} is played and every country i receives the welfare $\tilde{\pi}_i(q^{S_t})$. From the next round, the same play prevails in every MPE. Thus, country i receives the discounted welfare sum $\frac{1}{1-\delta}\tilde{\pi}_i(q^{S_t})$. If any new country j participates and the extended coalition $S_t \cup \{j\}$ forms, then the supper-additivity of the coalitional surplus implies that there exists some individually rational welfare transfer that makes all members of $S_t \cup \{j\}$ better off than $\tilde{\pi}_i(q^{S_t})$ in the current and future rounds. Thus, the coalition $S_t \cup \{j\}$ is agreed upon and the new participant j is better off than in the equilibrium. This is a contradiction.

4 Discussion

There are four main problems to be solved for the success of international cooperation in a broad range of conflicts (Okada 2017): common knowledge problem, agreement problem, compliance problem, and participation problem. We discuss the results of the paper with these problems.

The *common knowledge problem* addresses the following difficulty. Countries engaging in climate negotiations are uncertain about the causes and effects of global warming, benefits, and damages of GHGs, technological innovation, and many others. Without complete information and mutual understanding of these uncertain factors, it is hard for countries to reach an IEA. In terms of game theory, the rule of the game and the goals of players should be common knowledge of players (they know that they know that ...). Although scientific reports of the Intergovernmental Panel on Climate Change (IPCC) have helped us solve the common knowledge problem, there are strategic problems that countries may be motivated to report false information on their own cost and benefit of abatement.

The *agreement problem* addresses whether and how participants can reach an IEA. Countries negotiate for various terms such as a total emission, an allocation of emission permits, a commitment period, flexible methods such as emission trading, clean development mechanism, and joint implementation, and so on. Participants also determine a collective choice rule and a bargaining protocol.

The *compliance problem* addresses whether and how countries comply with and implement an IEA that constrains their behavior. If they do not comply with an IEA, the negotiation becomes cheap talk. The literature on stable IEA coalitions traditionally abstracts from the compliance problem. It assumes that an IEA is effectively implemented by some external enforcement mechanism. The literature tells us that even if an IEA is enforceable, climate cooperation is hardly achieved due to free-riding. Although the notion of a stable coalition is often called a "self-enforcing" agreement (Barrett 1994), its meaning should be taken with care. The standard two-stage game of participation assumes that the agreement of the optimal abatement strategy of a coalition is exogenously enforced in the second stage. What is self-enforcing is the participation strategies of countries since they compose a Nash equilibrium of the first stage. An IEA itself is enforced by some unspecified mechanism in the model.

The *participation problem* addresses whether autonomous countries voluntarily participate in an IEA. In the international world, there exists no central institution that enforces countries to participate. Participation should be voluntary. A participation rule is either open or closed. Under the open-access one, any country freely joins an IEA. Under the closed-access one, countries can participate in an IEA only if their participation is invited and approved by incumbent members. The two-stage game in the literature of IEA's employs the open-access rule.

In this paper, we have focused the analysis on the participation problem of IEA's, following the literature of stable coalitions. To simplify the model, we leave the common knowledge and the compliance problems out of the scope of the paper. So, we have assumed that the welfare functions of countries are public information and that an IEA is effectively enforced if it is agreed upon. We consider the agreement problem in the simplest form. A coalition of countries is assumed to maximize the total welfare of participants, and a transfer is predetermined so that it satisfies individual rationality and coalitional efficiency. We have analyzed whether and how self-interested countries form the grand coalition for climate cooperation. In the companion paper (Okada 2022), we consider how countries agree on the Nash bargaining solution, and show the same results as Theorems 1 and 2. The approach of this paper has the advantage that the results do not depend on a specific bargaining solution.

The dynamic coalition game in the paper has two restrictive assumptions: (i) only one coalition forms, and (ii) a coalition is irreversible in that participants are not allowed to withdraw from it. The first assumption has been employed in the literature of IEA's. It seems reasonable in the context of climate change since there exists only one type of global public goods (climate). In other contexts such as trade unions and R&D cartels of oligopolistic firms, a general model of multiple coalitions is more appropriate.

The second assumption is admittedly restrictive, while it has been employed in several works of dynamic coalitional bargaining games (Seidmann and Winter 1998, Okada 2000, Gomes 2005). In climate negotiations, the US withdrew from the Kyoto protocol. In another area, the UK exited from the EU. However, these withdrawals were controversial in the exiting countries, and it is not clear whether or not the withdrawal decisions were based on rational considerations. Countries are constrained not to withdraw from an international agreement by various factors such as domestic politics, moral obligation, public opinion, trust, and reputation.

To relax the assumption of irreversible coalitions, we modify the rule of Γ so that all countries can decide whether to participate (or remain) in the coalition in every round. Participants in an IEA have the option to withdraw from it. Theorem 1 holds in the modified game as well. The reason is as follows. The MPE in the theorem can be applied to the new game. In the equilibrium, the grand coalition forms in all rounds. If any member withdraws, that country is not better off for the same reason as a non-participant is not so in Γ. However, Theorem 2 does not hold. When participants are allowed to leave the coalition,

Γ is equivalent to the repeated game where the three-stage game is played over infinitely many rounds without any coalitional commitment. Then, it is well-known that the repetition of a subgame perfect equilibrium of the one-shot game is an MPE of the repeated game. It implies that an inefficient coalition forms in every round in an MPE when coalitions are reversible.

Finally, we discuss the relationship of this paper with the existing works that consider the possibility of the grand coalition of the IEA by the notion of a weakly renegotiation-proof equilibrium in the repeated emission game (Barrett 1999, Froyn and Hovi 2008, and Asheim and Holtsmark 2009). Among others, Asheim and Holtsmark (2009) show that there exists a weakly renegotiation-proof equilibrium implementing a Pareto efficient abatement for sufficiently high discount factors (greater than $1 - \frac{1}{n}$), and characterize the maximum depth of cooperation in the grand coalition for low discount factors. A subgame perfect equilibrium is called weakly renegotiation-proof if there exist no two histories such that all players prefer the continuation values after one history to those after the other history (Farrell and Maskin 1986). If the condition does not hold, players will renegotiate the continuation strategy (punishments) to the Pareto-improving one.

The approach of this paper can be compared with that of weakly renegotiation-proof equilibria in the following three aspects.

First, both approaches are complementary. This paper considers the participation problem of IEA's, while the works of weakly renegotiation-proof equilibria consider the compliance problem. The former investigates how countries can agree on an efficient IEA, and the latter does how they comply with such an agreement under credible punishments. It is promising for future work to unify two approaches. See Maruta and Okada (2012) for preliminary work.

Second, the literature of weakly renegotiation-proof equilibria presumes the number of participants in an IEA as given and shows that the agreement is credibly implemented by the equilibrium strategy. Every country, however, may not participate in the equilibrium strategy. For example, if a country publicly announces and commits to non-participation before the game, hoping that all others play the weakly renegotiation-proof equilibrium, then the non-participant is better off by free-riding. Then, the Paradox of International Agreements arises. In a general context, the problem is referred to as the "second-order" dilemma of public goods (Ostrom 1990), which says that rational and self-interested players may have the incentive to free-ride on the mechanism, being the second-order public goods, that is designed and provided to solve the first-order public goods. The result of this paper shows how countries can overcome the dilemma in the context of IEA's by engaging in the dynamic participation game with coalition formation.

Third, as the incentive to cooperate, we employ welfare transfers, while the works of weakly renegotiation-proof equilibria do punishments. In real-world IEA negotiations, it may not be acceptable politically or morally to damage the global environments to punish free-riders.

5 Conclusion

International negotiation is a dynamic process. An agreement is negotiated in sequential steps and renegotiated to a better one. A coalition of countries participating in a treaty changes through participation and withdrawal. The Kyoto Protocol has been renegotiated to the Paris Agreement with more participants.

We have presented a dynamic coalition game for IEAs where a coalition is irreversible and participants negotiate transfer schemes to attract non-participants. We have shown that when countries are sufficiently patient, there exists an MPE where all countries form the grand coalition in the first round of negotiations. Even if the counter-factual events that some countries do not participate in the IEA happen off the equilibrium play, the grand coalition will be recovered in renegotiations. Free-riding is deterred by two different motives. If the participants do not form their coalition, then the delay of the grand coalition hurts a non-participant. Otherwise, the coalition will be expanded to the grand coalition in future negotiations, and the non-participant will be worse off due to the trade-off of the welfare between the incumbent participants and the non-participant. We have further shown that for every discount factor of future welfare, the grand coalition gradually forms in finitely many rounds of negotiations. We conclude that the Paradox of IAs in the literature can be resolved in the dynamic coalition game.

The results of the paper provide several policy recommendations for real-world climate cooperation. Countries should continue to negotiate until they agree on an efficient IEA. Ultimatums and deadlines are not effective for climate cooperation. They cause free-riding. Participants should design an efficient and fair transfer scheme to attract non-participants. Countries should have long-term concerns about their welfare. They should be motivated not to withdraw from an agreement during negotiations so that other countries can have prospects in future negotiations. Political constraints and public opinion have a critical role to provide politicians and bureaucrats with the right incentives in climate negotiations.

Appendix

Proof of Theorem 1.

For every sufficiently large $\delta < 1$, we construct the desired MPE of Γ as follows. In the first round, all countries participate in N in the first stage, and agree to form it in the second stage, and play the N-equilibrium with transfer λ^N. Off the equilibrium path, the countries behave as follows in every round $t > 1$ where the incumbent coalition $S_{t-1} \neq N$ exists.

(i) In the first stage, all non-participants join S_{t-1}.

(ii) In the second stage with new participants P_t, every participant i of

$S_t = S_{t-1} \cup P_t$ votes for YES if

$$\tilde{\pi}_i(q^{S_t}|S_{t-1}) + \frac{\delta}{1-\delta}\tilde{\pi}_i(q^N|S_t) \geq \tilde{\pi}_i(q^{S_{t-1}}) + \frac{\delta}{1-\delta}\tilde{\pi}_i(q^N|S_{t-1}),$$

and otherwise, votes for NO.

(iii) In the third stage with the new coalition S_t, all countries play the S_t-equilibrium. If the negotiation fails in the second stage, all countries play the S_{t-1}-equilibrium.

Let σ^* denote the strategy constructed above. Clearly, σ^* is a Markov strategy. On the equilibrium play of σ^*, all countries i participate in the first round, and agree to form N since $\tilde{\pi}_i(q^N) \geq \pi_i(q^*)$. We shall show that σ^* is a subgame perfect equilibrium of Γ. For this, it is sufficient to show that the choice of every country at every move in σ^* is optimal to all other countries' choices, provided that σ^* will be played in all future moves.

By backward induction, consider first the third stage where coalition S_t forms in the second stage. By the rule of the game, the members of S_t are bound to play the S_t-equilibrium q^{S_t}. Every non-participant i also plays q^{S_t} in σ^*. Then, i receives the discounted welfare sum $\pi_i(q^{S_t}) + \delta \times \Pi_i(S_t)$ where $\Pi_i(S_t)$ is the continuation value that i will receive in round $t+1$. If i chooses an abatement q_i different from $q_i^{S_t}$, then i receives the discounted welfare sum $\pi_i(q_i, q^{S_t}) + \delta \times \Pi_i(S_t)$. Note that i's continuation value $\Pi_i(S_t)$ is independent of q_i. Since $q_i^{S_t}$ is the S_t-equilibrium, it holds that $\pi_i(q^{S_t}) \geq \pi_i(q_i, q^{S_t})$. thus, non-participant i is not better off by deviating from $q_i^{S_t}$.

Consider next the second stage with the incumbent coalition S_{t-1} and new participants P_t. If they agree to form $S_t = S_{t-1} \cup P_t$, then every participant i of S_t receives the discounted welfare sum $\tilde{\pi}_i(q^{S_t}|S_{t-1}) + \frac{\delta}{1-\delta}\tilde{\pi}_i(q^N|S_t)$. Note that S_t will expand to the grand coalition N in round $t+1$ by (i). In the same way, if negotiation fails, then every participant i of S_t receives the discounted welfare sum $\tilde{\pi}_i(q^{S_{t-1}}) + \frac{\delta}{1-\delta}\tilde{\pi}_i(q^N|S_{t-1})$. Therefore, (ii) gives the optimal choice of every participant of S_t.

Finally, consider the first stage. Following σ^*, all non-participants i join S_{t-1} to form N, and receive the discounted welfare $\frac{1}{1-\delta}\tilde{\pi}_i(q^N|S_{t-1})$. Suppose that any country i does not participate. Then, the remaining participants negotiate for forming the coalition $N/\{i\}$ in the second stage. The following two cases are possible.

case 1. $N/\{i\}$ does not form:

Non-participant i receives the discounted welfare sum $\tilde{\pi}_i(q^{S_{t-1}})+\frac{\delta}{1-\delta}\tilde{\pi}_i(q^N|S_{t-1})$. Note that N will form in the next round $t+1$ in σ^*. Since $\tilde{\pi}_i(q^N|S_{t-1}) \geq \tilde{\pi}_i(q^{S_{t-1}})$ by the individual rationality (2), non-participant i is not better off. Non-participation only delays the formation of N.

case 2. $N/\{i\}$ forms:

Non-participant i receives the discounted welfare sum $\tilde{\pi}_i(q^{N/\{i\}})+\frac{\delta}{1-\delta}\tilde{\pi}_i(q^N|N/\{i\})$. Since $N/\{i\}$ forms, it must hold for every $k \in N/\{i\}$ that

$$\tilde{\pi}_k(q^{N/\{i\}}) + \frac{\delta}{1-\delta}\tilde{\pi}_k(q^N|N/\{i\}) \geq \tilde{\pi}_k(q^{S_{t-1}}) + \frac{\delta}{1-\delta}\tilde{\pi}_k(q^N|S_{t-1}). \qquad (3)$$

Assume that $\tilde{\pi}_k(q^N|N/\{i\}) < \tilde{\pi}_k(q^N|S_{t-1})$ for some $k \in N/\{i\}$. Then, (3) does not hold for any sufficiently large $\delta < 1$ since $\lim_{\delta\to 1}\frac{\delta}{1-\delta} = +\infty$. Thus, without loss of generality, we can assume that $\tilde{\pi}_k(q^N|N/\{i\}) \geq \tilde{\pi}_k(q^N|S_{t-1})$ for all $k \in N/\{i\}$. Since

$$\sum_{j\in N}\tilde{\pi}_j(q^N|N/\{i\}) = \sum_{j\in N}\tilde{\pi}_j(q^N|S_{t-1}),$$

it holds that $\tilde{\pi}_i(q^N|N/\{i\}) \leq \tilde{\pi}_i(q^N|S_{t-1})$. By the individual rationality of the transfer in N, we have $\tilde{\pi}_i(q^N|N/\{i\}) \geq \tilde{\pi}_i(q^{N/\{i\}})$. Then, it holds that

$$\begin{aligned}\frac{1}{1-\delta}\tilde{\pi}_i(q^N|S_{t-1}) &\geq \frac{1}{1-\delta}\tilde{\pi}_i(q^N|N/\{i\}) \\ &\geq \tilde{\pi}_i(q^{N/\{i\}}) + \frac{\delta}{1-\delta}\tilde{\pi}_i(q^N|N/\{i\}).\end{aligned}$$

Thus, non-participant i is not better off by deviating from σ^*. $\square$

Proof of Theorem 2.

Let σ be any MPE of Γ. We show that on the equilibrium play of σ, there exists at least one country that joins an incumbent coalition S_t unless $S_t = N$. By way of contradiction, suppose not. Then, there are no new participants in S_t and as a result, the S_t-equilibrium is played in round t. The next round starts with the same coalition S_t. Since σ is an MPE, the same play prevails forever. Thus, every country $i \in N$ receives the discounted welfare sum $\frac{1}{1-\delta}\tilde{\pi}_i(q^{S_t})$.

Suppose that some country $i \notin S_t$ participates in S_t. By A3, it holds that

$$\begin{aligned}
\sum_{j \in S_t \cup \{i\}} \tilde{\pi}_j(q^{S_t \cup \{i\}}) &= \sum_{j \in S_t \cup \{i\}} \pi_j(q^{S_t \cup \{i\}}) \\
&> \sum_{j \in S_t} \pi_j(q^{S_t}) + \pi_i(q^{S_t}) \\
&= \sum_{j \in S_t} \tilde{\pi}_j(q^{S_t}) + \pi_i(q^{S_t}).
\end{aligned}$$

Therefore, the extended coalition $S_t \cup \{i\}$ can find some individually rational transfer for both the incumbent participants and the new participant i.

Every country $j \in S_t \cup \{i\}$ receives the discounted welfare sum

$$\tilde{\pi}_j(q^{S_t \cup \{i\}}) + \delta \times \Pi_j(S_t \cup \{i\}) \tag{4}$$

where $\Pi_j(S_t \cup \{i\})$ is the continuation value after $S_t \cup \{i\}$ forms. Since country j has the veto power in the negotiation, j receives at least $\tilde{\pi}_j(q^{S_t \cup \{i\}})$ forever, simply by rejecting any new coalition, after $S_t \cup \{i\}$ forms. Thus, $\Pi_j(S_t \cup \{i\}) \geq \frac{1}{1-\delta}\tilde{\pi}_j(q^{S_t \cup \{i\}})$. This implies that (4) is larger than $\frac{1}{1-\delta}\tilde{\pi}_i(q^{S_t \cup \{i\}})$, which is further larger than $\frac{1}{1-\delta}\tilde{\pi}_i(q^{S_t})$ by the individual rationality of the transfer in $S_t \cup \{i\}$. In the second stage, the coalition $S_t \cup \{i\}$ is agreed on in σ. Thus, i is better off by joining S_t. This is a contradiction. $\square$

References

Asheim, G. B. and Holtsmark (2009), "Renegotiation-proof climate agreements with full participation: Conditions for Pareto-efficiency." *Environment and Resource Economics*, 43, 519-533.

Barrett, S. (1994), "Self-enforcing international environmental agreements." *Oxford Economic Papers*, 46, 878-894.

Barrett, S. (1997), "The strategy of trade sanctions in international environmental agreements." *Resource and Energy Economics*, 19, 345-361.

Barrett S. (1999), "A theory of full international cooperation." *Journal of Theoretical Politics*, 11, 519-541.

Buchholz, W. and Sandler, T. (2021), "Global public goods: A survey." *Journal of Economic Literature*, 59, 488–545.

Carraro, C., Eyckmans, J. and Finus, M. (2006), "Optimal transfers and participation decisions in international environmental agreements." *Review of International Organizations*, 1, 379–396.

Carraro, C. and Siniscalco, D. (1993), "Strategies for the international protection of the environment." *Journal of Public Economics*, 52, 309-328.

Carraro, C. and Siniscalco, D. (1998), "International environmental agreements: Incentives and political economy." *European Economic Review*, 42, 561-572.

Coase, R. H. (1960), "The problem of social cost." *Journal of Law and Economics*, 3, 1–44.

D'Aspremont, C., Jacquemin, A., Gabszewicz, J.J. and Weymark, J.A. (1983), "On the stability of collusive price leadership." *Canadian Journal of Economics*, 16, 17-25.

Farrell, J. and Maskin, E. (1989), "Renegotiation in repeated games." *Games and Economic Behavior*, 1, 327–360.

Froyn, C. B. and Hovi, J. (2008), "A climate agreement with full participation." *Economics Letters*, 99, 317319.

Fudenberg, D. and E. Maskin (1986), "The folk theorem in repeated games with discounting or with incomplete information." *Econometrica*, 54, 533-554.

Gomes, A. (2005), "Multilateral contracting with externalities." *Econometrica*, 73, 1329–1350.

Hoel, M. (1992), "International environment conventions: The case of uniform reductions of emissions." *Environmental and Resource Economics*, 2, 141-159.

Hovi, J., Ward, H., and Grundig, F. (2015) "Hope or despair? Formal models of climate cooperation." *Environmental and Resource Economics*, 62, 665-688.

Hyndman, K. and Ray, D. (2007), "Coalition formation with binding agreements." *Review of Economic Studies*, 74, 1125–1147.

Karp, L. and Simon, L. (2013), "Participation games and international environmental agreements: A non-parametric model." *Journal of Environmental Economics and Management*, 65, 326-344.

Kolstad, C.D. and Toman, M.(2005), "The economics of climate policy," In: Mler, K.-G. and Vincent, J.R. (Eds.), *Handbook of Environmental Economics*, vol. 3. Elsevier, Amsterdam, 15611618.

Kováč, E. and Schmidt, R.C. (2021), "A simple dynamic climate cooperation model." *Journal of Public Economics*, 194, 1-24.

Maruta, T. and Okada, A. (2012), "Dynamic group formation in the repeated prisoner's dilemma." *Games and Economic Behavior*, 74, 269–284.

Nordhaus, W. (2015), "Climate clubs: Overcoming free-riding in international climate policy." *American Economic Review*, 105, 1339-1370.

Okada, A. (2000), "The efficiency principle in non-cooperative coalitional bargaining." *Japanese Economic Review*, 51, 34–50.

Okada, A. (2017), "International cooperation and institution formation: A game theoretic perspective. " In M. Suzuki and A. Okada (eds.), *Games of Conflict and Cooperation in Asia,* Springer, 137-158.

Okada, A. (2022), "Dynamic bargaining with voluntary participation and externality." *Economic Theory*, https://doi.org/10.1007/s00199-021-01410-4.

Ostrom, E. (1990), '*Governing the Commons: The Evolution of Institutions for Collective Action.* Cambridge: Cambridge University Press.

Ray, D. (2007), *A Game-Theoretic Perspective on Coalition Formation.* Oxford University Press, Oxford.

Seidmann, D.J. and Winter, E. (1998), "A theory of gradual coalition formation." *Review of Economic Studies*, 65, 793–815.

Selten, R. (1973), "A simple model of imperfect competition, where 4 are few and 6 are many." *International Journal of Game Theory*, 2, 141-201.