Energy Economics 151 (2025) 108952

Contents lists available at ScienceDirect

Energy Economics

journal homepage: www.elsevier.com/locate/eneco

Designing tax plans in international environmental agreements with heterogeneous benefits

Ping Qiu [a,b], Liang Mao [a,*]

[a] College of Economics, Shenzhen University, 1066 Xueyuan Avenue, Shenzhen, Guangdong, China
[b] Nottingham University Business School China, University of Nottingham Ningbo China, 199 Taikang Road, Ningbo, Zhejiang, China

ARTICLE INFO

Keywords:
International environmental agreements
Game theory
Simulation
Tax regulation

ABSTRACT

We present an analysis of carbon abatement regulation through a self-enforcing international environmental agreement (IEA) model featuring two types of countries with dissimilar abatement benefits. The IEA involves a tax plan function that allocates an emission tax rate to each type of country under every coalition of voluntary signatories. An efficient tax plan is one that maximizes social welfare under a stable coalition, while an optimal tax plan maximizes the average payoff of a stable coalition. We demonstrate that an efficient or optimal tax plan always exists, and that the corresponding value of social welfare or average coalition payoff is greater than that under certain traditional tax systems. If the benefit heterogeneity between the two types of countries is sufficiently small, full cooperation and social optimum can be achieved or approximated through an efficient or optimal plan. Conversely, a high degree of heterogeneity will result in a relatively small coalition and an inefficient outcome, regardless of the tax plan employed.

1. Introduction

The regulation of global public goods has been a popular and long-standing topic in the fields of public and environmental economics. Specifically, the issue of climate change has been attributed to excessive carbon emissions and has been widely recognized as a significant consequence. Due to the presence of externalities, implementing carbon abatement measures is not feasible without regulatory intervention. In practice, two regulation methods have been commonly used and discussed in the literature, namely quantity and price regulation. Quantity regulation involves establishing an emission level quota, while price regulation entails imposing an emission fee or tax.[1] Under certain conditions, price regulation has been found to perform better than quantity regulation in terms of both abatement level and social welfare. Additionally, price regulation has been shown to offer other advantages, such as reduced administrative costs and the generation of additional tax revenues.

A special type of price regulation is the uniform tax system in which the tax rates are the same across all regulated countries.[2] Compared to more complex tax systems, the merit of this tax system is that it reduces potential conflicts during international negotiations while remains the potential to achieve satisfactory regulatory outcomes. Weitzman (2014) proved that a tax rate exists that, if globally applied, can help achieve a socially optimal outcome even when countries are heterogeneous.

However, as highlighted by McEvoy and McGinty (2018), Weitzman (2014) did not consider the issue of voluntary participation. No country would voluntarily charge a certain emission tax unless it is in its best interests. Hence, we should formally examine the determination of the ratio of countries that agree to this tax. In the literature, this is sometimes formulated as a two-stage game, often referred to as the international environmental agreement (IEA) model. In stage one, some countries voluntarily form a coalition and sign a self-enforcing IEA. In stage two, all signatories (coalition members) should take action according to the agreement.[3] The stability concept by d'Aspremont et al. (1983) is commonly applied in this model to determine the coalition formed in stage one and the resulting ratio of signatory countries. After considering the issue of voluntary participation, a uniform tax system does not seem to work well compared to the optimistic results in Weitzman (2014). Notably, McEvoy and McGinty (2018) showed that in a simple model, in which all countries are ex-ante homogeneous but only signatories to an IEA should charge the emission tax, only a

* Corresponding author.
E-mail address: maoliang@szu.edu.cn (L. Mao).

[1] Both regulatory approaches enable the implementation of a system of tradable emission permits.
[2] For examples on uniform tax, see Pearce (1991), Hoel (1992), Nordhaus (2006), Weitzman (2014), Cramton et al. (2015), and McEvoy and McGinty (2018).
[3] Carraro and Siniscalco (1993) and Barrett (1994) are some of the early studies that considered the IEA model. See also the reviews of Finus (2001) and Carraro (2003).

https://doi.org/10.1016/j.eneco.2025.108952
Received 9 March 2023; Received in revised form 26 July 2025; Accepted 27 September 2025
Available online 3 October 2025

very small fraction of countries will sign the IEA. The resulting carbon abatement is far from a socially optimal level.

Naturally, the following question arises: can a more satisfactory result of carbon abatement be achieved through a proper (not necessarily uniform) tax system within the framework of voluntarily formed coalition? To answer this question, we address the following two relevant issues.

One issue that may affect the performance of an IEA is the heterogeneity of countries. Earlier works in the IEA literature found that the impact of heterogeneity may be complex and diversified (Fuentes-Albero and Rubio, 2010; Pavlova and de Zeeuw, 2013; Finus and McGinty, 2019; Bakalova and Eyckmans, 2019). On the one hand, heterogeneity sometimes makes coordinating the abatement levels of different countries more difficult, leading to a smaller coalition of signatory countries. On the other hand, heterogeneity provides a potential surplus from the cooperation among different countries, and hence is good for the formation and stability of large coalitions, especially when transfers exist among countries. Therefore, the impact of heterogeneity on an IEA crucially depends on the model setup. For analytical convenience, this study only analyzes a simple setup with two types of countries that differ in their abatement benefits[4] and does not allow payoff transfers among countries. Due to the development of international carbon trading markets, we also assume that the marginal abatement costs are the same across all countries.

Another important issue is the design of the IEA. A widely adopted assumption in the IEA model literature that signatories to an IEA should act collectively to maximize a given payoff objective (e.g., the joint payoff of the coalition), whether the coalition is stable or not. Although this assumption is reasonable for stable coalitions, why the corresponding payoff should be maximized for non-stable coalitions remains unclear.[5] To address this question, some recent studies accommodate more general classes of IEAs, and endogenously choose an "optimal" one among them (Carraro et al., 2009; Köke and Lange, 2017; Mao, 2020; Masoudi, 2022). For instance, Mao (2020) studied a class of IEA rules wherein a signatory's abatement level depends on the number of signatories, and the optimal IEA rule can be properly designed to maximize the payoffs to the signatories in stable coalitions. Mao showed that the result can be significantly improved relative to the traditional IEA model if we only aim for stable coalitions because the superfluous requirement on non-stable coalitions will hinder the formation of a large coalition. This motivates us to study whether a similar tax system (formally referred to as a tax plan in this paper), wherein the tax rate for a signatory depends on the coalition of signatories, can achieve greater social welfare than traditional tax systems. To reflect the heterogeneity among countries, we also allow a country's tax rate specified by an IEA to vary according to its abatement benefit.

In summary, the purpose of this study is to explore the performance of an IEA that uses a tax plan under heterogeneous abatement benefits. To this end, we extend the traditional IEA model to a three-stage game. In stage one, a regulator (for example, the United Nations) designs a tax plan in order to achieve a large objective payoff. In stage two, all countries simultaneously decide whether to sign the IEA, and those signatories form a coalition. In stage three, countries decide on their tax rates, where signatories should follow what is specified in the tax plan given the coalition formed while the choices of non-signatories are not subject to such limitations.

A tax plan is considered efficient (or optimal) if it maximizes the expected value of social welfare (or average coalition payoff, respectively) in stage one. We provide an algorithm for calculating an efficient (optimal) tax plan for each game (Theorem 1). Under symmetry, an efficient/optimal tax plan can result in full cooperation and social optimum (Propositions 1 and 2). Under small heterogeneity, social welfare (or coalition payoff) under an efficient (optimal) tax plan decreases with the degree of benefit heterogeneity, but it is unaffected by the average level of marginal benefit (Propositions 1 and 3). As the degree of heterogeneity becomes larger, simulations show that these results would generally hold, except that a smaller coalition may form. Overall, our results are more optimistic than McEvoy and McGinty (2018) but more pessimistic than Weitzman (2014) regarding the performance of global carbon tax system.

One feature that is worth pointing out in our model setup is the functional forms. Some studies in the literature adopt a simple setting that both benefit and cost functions are linear (Kolstad and Ulph, 2011; Ulph et al., 2019). Some other studies work on a more realistic but also more complicated assumption that the benefit function is concave and the cost function is convex (Barrett, 1994; Weitzman, 2014; McEvoy and McGinty, 2018). We adopt a compromise by assuming a linear benefit function and a convex cost function (Na and Shin, 1998; Fujita, 2004; Köke and Lange, 2017; Mao, 2020). Note that some of our conclusions depend on the linearity of the benefit function and may not necessarily apply for more general functional forms. Unlike the case of a general benefit function in which a country's abatement level may decrease in reaction to increased abatement by others, a non-signatory country has a dominant abatement level that does not depend on other countries' actions if its benefit function is linear.

The remainder of this paper is organized as follows. After providing the model setup in Section 2, we present and solve the three-stage IEA game in Section 3. Section 4 shows the existence of efficient and optimal tax plans and provides some numerical examples. We analyze the properties of these tax plans in Section 5 and examine some simulation examples in Section 6. Finally, Section 7 concludes the study.

2. The model

2.1. Basic setup

There are two types of countries: A and B. The set of type k countries are denoted as N_k, and $|N_k| = n_k$ is the number of type k countries, $k \in \{A, B\}$. Let $N = N_A \cup N_B$ and $n = n_A + n_B$ be the set and the number of all countries, respectively.

Each country i has a representative firm F_i and a government G_i. Let $x_i \geq 0$ denote the carbon abatement of F_i below an initial emission level e_i. Naturally, x_i is also the abatement level of country i. Suppose that e_i is sufficiently large so that we need not consider the constraints $x_i \leq e_i$ throughout this study. Let $X = \sum_{i\in N} x_i$ be the total abatement.

Government G_i collects an emission tax from F_i at a rate $p_i \geq 0$. The tax revenue is $\tau_i(p_i, x_i) = p_i(e_i - x_i)$, which is retained within the country. Let $\boldsymbol{p} = (p_1, \dots, p_n)$ denote a tax rate combination, and let $\bar{p} = \frac{1}{n}\sum_{i\in N} p_k$ be the average tax rate.

Firm F_i's cost of abatement is $C(x_i) = \frac{1}{2}x_i^2$. Thus, the profit of F_i is

$$\pi_i = \pi_i^0 - C(x_i) - \tau_i(p_i, x_i), \tag{1}$$

where π_i^0 is F_i's baseline profit when there is no tax ($p_i = 0$) and no emission reduction ($x_i = 0$). Assume that all π_i^0 are the same and are normalized to 0.

Country i's benefit from abatement is $B_i(X) = \lambda_i X$, where λ_i is its marginal benefit. Suppose $\lambda_i = \lambda_A$ if $i \in N_A$, and $\lambda_i = \lambda_B$ if $i \in N_B$, where $\lambda_A \geq \lambda_B > 0$. Thus, the type of country i is characterized by λ_i. Let $\bar{\lambda} = \frac{1}{n}\sum_{i\in N} \lambda_i = \frac{n_A\lambda_A + n_B\lambda_B}{n}$ denote the average marginal benefit.

Since $\tau_i(p_i, x_i)$ is retained within the country, the payoff of country i is

$$u_i = B_i(X) + \pi_i + \tau_i(p_i, x_i) = \lambda_i X - \frac{1}{2}x_i^2. \tag{2}$$

[4] In practice, some countries (e.g. the Maldives) suffer more from global warming, and hence benefit more from carbon emission abatement, than other countries (e.g. Russia).

[5] See Mao (2018) for a counterexample.

Table 1
Notations.

ω	coalition structure	Ω	set of coalition structures
θ	tax plan	Θ	set of tax plans
$\overline{\Omega}(\theta)$	set of stable coalition structures under θ	$\overline{\Theta}(\omega)$	set of essential plans at ω
$\overline{\Theta}$	set of essential plans	$\omega(\theta)$	coalition structure under $\theta \in \overline{\Theta}$
$\overline{\Theta}^0$	set of essential plans: $\omega(\theta) \neq (0,0)$		

Given a tax rate combination $\boldsymbol{p}$, firm F_i chooses abatement level $x_i(\boldsymbol{p})$ to maximize its profit π_i. From (1), we have

$$x_i(\boldsymbol{p}) = p_i. \tag{3}$$

Thus, the total abatement is

$$X(\boldsymbol{p}) = \sum_{i\in N} x_i(\boldsymbol{p}) = n\bar{p}. \tag{4}$$

Using (2), (3) and (4), the payoff of country i under a given $\boldsymbol{p}$ is

$$u_i(\boldsymbol{p}) = \lambda_i X(\boldsymbol{p}) - \frac{1}{2}x_i(\boldsymbol{p})^2. \tag{5}$$

Finally, we define the social welfare

$$U(\boldsymbol{p}) = \frac{1}{n}\sum_{i\in N} u_i(\boldsymbol{p}) = \bar{\lambda}X(\boldsymbol{p}) - \frac{1}{2n}\sum_{i\in N} x_i(\boldsymbol{p})^2. \tag{6}$$

as the average payoff of all countries.

2.2. Two special tax rate combinations

The determination of tax rate combination $\boldsymbol{p}$ is a theme of this study. In this subsection, we consider two special cases. If $\boldsymbol{p}$ is entirely up to a global regulator to decide, then the chosen combination $\boldsymbol{p}$ must be socially optimal in the sense that it maximizes social welfare $U(\boldsymbol{p})$. On the other hand, if each p_i is decided by the government G_i so that country i's payoff is maximized given other countries' tax rates, then all governments are involved in a non-cooperative game, and the resulting combination $\boldsymbol{p}$ should be a Nash equilibrium of this game.

Let $\boldsymbol{p}^* = (p_1^*, \dots, p_n^*)$ denote a socially optimal combination, and let $\boldsymbol{p}^0 = (p_1^0, \dots, p_n^0)$ denote a Nash equilibrium combination. Then, $x_i(\boldsymbol{p}^*)$ and $x_i(\boldsymbol{p}^0)$ are country i's socially optimal and equilibrium abatement levels, respectively.

Lemma 1(a) and 1(b) explicitly characterize the tax rates and the corresponding abatement levels in the equilibrium case and the socially optimal case, respectively. In addition, Lemma 1(c) suggests that the equilibrium abatement level is less than the socially optimal level.

Lemma 1.
(a) *Let $p_i^0 = \lambda_i$, then $\boldsymbol{p}^0 = (p_1^0, \dots, p_n^0)$ is a dominant-strategy equilibrium and hence a Nash equilibrium, and $x_i(\boldsymbol{p}^0) = \lambda_i$ for all $i \in N$.*
(b) *Let $p_i^* = n\bar{\lambda}$, then $\boldsymbol{p}^* = (p_1^*, \dots, p_n^*)$ is socially optimal, and $x_i(\boldsymbol{p}^*) = n\bar{\lambda}$ for all $i \in N$.*
(c) *$x_i(\boldsymbol{p}^*) > x_i(\boldsymbol{p}^0)$ for all $i \in N$.*

Proof. (a) Given any $i \in N$ and $\boldsymbol{p} = (p_1, \dots, p_n)$ where $p_i \neq \lambda_i$, let $\boldsymbol{p}^i = (p_1^i, \dots, p_n^i)$ such that $p_j^i = p_j$ for all $j \neq i$ and $p_i^i = \lambda_i$. From (5), it follows that $u_i(\boldsymbol{p}^i) - u_i(\boldsymbol{p}) = \frac{1}{2}(\lambda_i - p_i)^2 > 0$. Therefore, $p_i = \lambda_i$ is a dominant strategy of G_i. Hence, $\boldsymbol{p}^0 = (p_1^0, \dots, p_n^0)$ is a dominant-strategy and Nash equilibrium. From (3), $x_i(\boldsymbol{p}^0) = \lambda_i$ for all $i \in N$.
(b) Since $U(\boldsymbol{p}) = \bar{\lambda}X(\boldsymbol{p}) - \frac{1}{2n}\sum_{i\in N} x_i(\boldsymbol{p})^2$, we have $\frac{\partial U}{\partial x_i} = \bar{\lambda} - \frac{x_i}{n}$. It follows from $\frac{\partial U}{\partial x_i} = 0$ that $x_i(\boldsymbol{p}^*) = n\bar{\lambda}$. From (3), $p_i^* = x_i(\boldsymbol{p}^*) = n\bar{\lambda}$.
(c) Evidently, $x_i(\boldsymbol{p}^*) = n\bar{\lambda} > \lambda_i = x_i(\boldsymbol{p}^0)$. □

3. The IEA game

In Table 1, we list some notations used in this section for readers' reference.

The result of too much equilibrium emission in Lemma 1(c) arises from the incentives of free-riding on other countries' abating efforts. To solve this problem, some countries may voluntarily join a coalition and sign an IEA to regulate their own actions. Specifically, an IEA specifies a function that assigns a tax rate to each coalition. We refer to this function as a tax plan.

Formally, let $m_A \in [0, n_A]$ and $m_B \in [0, n_B]$ denote the number of type A and B signatories to the IEA, respectively. A pair $\omega = (m_A, m_B)$ is called a coalition structure and can fully characterize the coalition. Let Ω denote the set of all coalition structures. A tax plan θ is a two-variable function that assigns a uniform tax rate $\theta(\omega)$ to each coalition structure $\omega \in \Omega\backslash(0,0)$. Let Θ denote the set of all tax plans.

The determination of the tax plan and coalition formation can be described by a three-stage IEA game $G\left(n_A, n_B, \lambda_A, \lambda_B\right)$. In stage one, a regulator designs a tax plan θ with the aim of maximizing an objective payoff (either expected social welfare or average coalition payoff, which will be formally defined in (23) or (24), respectively). In stage two, the governments of all countries decide simultaneously whether to join the coalition and sign the IEA to maximize the respective payoffs of their own countries, resulting in a coalition structure ω. In stage three, each signatory i follows the IEA and sets a tax rate that determined by the tax plan and the coalition formed:

$$p_i = \theta(\omega) + \lambda_i, \tag{7}$$

while each non-signatory j will choose $p_j = \lambda_j$ according to Lemma 1(a), resulting in a tax rate combination $\boldsymbol{p}$.

Intuitively, (7) shows that a signatory country i's tax rate specified by the tax system is the sum of the its dominant rate λ_i and a uniform rate $\theta(\omega)$. Note that the dominant rate λ_i is fixed because of the linear benefit of abatement function, and does not apply for more general functional forms of benefit functions.

Traditional IEA games usually assume that after any coalition is formed, a tax rate for signatories will be chosen to maximize the objective payoff. By contrast, our tax system is explicitly designed before the coalition forms and is more flexible since it only concerns the objective payoff under the coalition that forms in the equilibrium of game.[6] The flexibility of our tax system provides the potential to form a larger coalition and achieve a greater objective payoff.

In the following part of this section, we will solve $G\left(n_A, n_B, \lambda_A, \lambda_B\right)$ by means of backward induction.

3.1. Stage three

In stage three, given a tax plan θ and a coalition structure $\omega = \left(m_A, m_B\right)$, it follows from (3), (4) and Lemma 1(a) that the abatement of a type $k \in \{A, B\}$ signatory (cooperator) and non-signatory (outsider) are

$$x_k^C(m_A, m_B; \theta) = \theta(m_A, m_B) + \lambda_k, \text{ if } m_k > 0, \tag{8}$$

and

$$x_k^O(m_A, m_B; \theta) = \lambda_k, \text{ if } m_k < n_k, \tag{9}$$

respectively, and the total abatement is

$$X(m_A, m_B; \theta) = (m_A + m_B)\theta(m_A, m_B) + n\bar{\lambda}. \tag{10}$$

When $m_k > 0$, the payoff of a type k signatory is

$$u_k^C(m_A, m_B; \theta) = \lambda_k X(m_A, m_B; \theta) - \frac{1}{2}x_k^C(m_A, m_B; \theta)^2. \tag{11}$$

[6] Later, we will define this coalition as a stable coalition.

When $m_k < n_k$, the payoff of a type k non-signatory is

$$u_k^O(m_A, m_B; \theta) = \lambda_k X(m_A, m_B; \theta) - \frac{1}{2} x_k^O(m_A, m_B; \theta)^2. \tag{12}$$

Social welfare under θ and $\omega = (m_A, m_B)$ is

$$\begin{aligned} V(m_A, m_B; \theta) = & \left[m_A u_A^C(m_A, m_B; \theta) + (n_A - m_A) u_A^O(m_A, m_B; \theta) \right. \\ & \left. + m_B u_B^C(m_A, m_B; \theta) + (n_B - m_B) u_B^O(m_A, m_B; \theta) \right] / n. \end{aligned} \tag{13}$$

When $(m_A, m_B) \neq (0, 0)$, average coalition payoff is

$$Y(m_A, m_B; \theta) = \left[m_A u_A^C(m_A, m_B; \theta) + m_B u_B^C(m_A, m_B; \theta) \right] / (m_A + m_B). \tag{14}$$

3.2. Stage two

In stage two of $G(n_A, n_B, \lambda_A, \lambda_B)$, given tax plan θ, the government of a country will choose to join the coalition unless it is strictly better off otherwise. Following (d'Aspremont et al., 1983), a coalition is considered stable relative to θ, if

$$u_A^C(m_A, m_B; \theta) \geq u_A^O(m_A - 1, m_B; \theta), \quad \text{when } m_A > 0; \tag{15}$$

$$u_B^C(m_A, m_B; \theta) \geq u_B^O(m_A, m_B - 1; \theta), \quad \text{when } m_B > 0; \tag{16}$$

$$u_A^O(m_A, m_B; \theta) > u_A^C(m_A + 1, m_B; \theta), \quad \text{when } m_A < n_A; \tag{17}$$

$$u_B^O(m_A, m_B; \theta) > u_B^C(m_A, m_B + 1; \theta), \quad \text{when } m_B < n_B. \tag{18}$$

Conditions (15) and (16) show that no signatory will unilaterally leave the coalition (internally stable), while (17) and (18) imply that no non-signatory will unilaterally join the coalition (externally stable).

If a coalition is stable relative to θ, we may equivalently say that its coalition structure is stable relative to θ. Let $\overline{\Omega}(\theta)$ denote the set of all stable coalition structures relative to θ.

Given a tax plan θ, only those coalitions that are stable relative to θ can be formed. However, stable coalitions do not always exist. That is, $\overline{\Omega}(\theta)$ may sometimes be empty. For example, suppose $n_A = n_B = 1$, $\lambda_A = \lambda_B = 2$. We define a tax plan θ_1, where $\theta_1(0, 1) = 0.1$ and $\theta_1(1, 0) = \theta_1(1, 1) = 0$. It is easy to verify that $u_A^O(0, 0; \theta_1) = u_A^C(1, 0; \theta_1) = 6$, $u_B^O(1, 0; \theta_1) = u_B^C(1, 1; \theta_1) = 6$, $u_A^O(0, 1; \theta_1) = 6.2 > u_A^C(1, 1; \theta_1) = 6$, $u_B^O(0, 0; \theta_1) = 6 > u_B^C(0, 1; \theta_1) = 5.995$. The four possible coalition structures form a deviation cycle $(0, 0) \to (1, 0) \to (1, 1) \to (0, 1) \to (0, 0)$; hence, no coalition structure is stable relative to θ_1. Additionally, even when $\overline{\Omega}(\theta)$ is not empty, it may contain multiple coalition structures. Given a tax plan, the non-existence and non-uniqueness of stable coalition structures make it difficult to anticipate which coalition will be formed.

To avoid this problem, consider a special type of tax plans relative to each of which a unique stable coalition structure exists. Given $\omega = (m_A, m_B)$, a tax plan θ is called an essential plan at ω, if

$$u_A^C(s_A + 1, s_B; \theta) \geq u_A^O(s_A, s_B; \theta), \quad \text{when } s_A < m_A; \tag{19}$$

$$u_B^C(s_A, s_B + 1; \theta) \geq u_B^O(s_A, s_B; \theta), \quad \text{when } s_B < m_B; \tag{20}$$

$$u_A^O(s_A - 1, s_B; \theta) > u_A^C(s_A, s_B; \theta), \quad \text{when } s_A > m_A; \tag{21}$$

$$u_B^O(s_A, s_B - 1; \theta) > u_B^C(s_A, s_B; \theta), \quad \text{when } s_B > m_B. \tag{22}$$

Let $\overline{\Theta}(\omega)$ denote the set of all essential plans at ω.

If $\theta \in \overline{\Theta}(\omega)$, then any other coalition structure $(s_A, s_B) \neq \omega$ cannot be stable relative to θ. Intuitively, θ will compel (s_A, s_B) to transform into $\omega = (m_A, m_B)$. In fact, when $s_A < m_A$ or $s_B < m_B$ holds, (19) or (20) indicates the existence of a type A or type B non-signatory who is willing to join the coalition and become a signatory. Thus, any coalition structure (s_A, s_B) cannot be externally stable if $s_A < m_A$ or $s_B < m_B$. Similarly, (21) and (22) show that (s_A, s_B) cannot be internally stable if $s_A > m_A$ or $s_B > m_B$.

Furthermore, we have:

Lemma 2. *For any $\theta \in \overline{\Theta}(\omega)$, $\overline{\Omega}(\theta) = \{\omega\}$.*

Proof. We have already established that $\omega' \notin \overline{\Omega}(\theta)$ for all $\omega' \neq \omega$. It remains to prove that $\omega \in \overline{\Omega}(\theta)$. If we let $(s_A, s_B) = (m_A - 1, m_B)$ in (19) and let $(s_A, s_B) = (m_A, m_B - 1)$ in (20), then (m_A, m_B) is internally stable relative to θ. Likewise, by letting $(s_A, s_B) = (m_A + 1, m_B)$ in (21) and $(s_A, s_B) = (m_A, m_B + 1)$ in (22), we see that (m_A, m_B) is externally stable relative to θ. Therefore, $\omega \in \overline{\Omega}(\theta)$. □

Lemma 2 establishes the existence and uniqueness of a stable coalition structure for any essential plan. Therefore, it is convenient to assume that the regulator will only choose an essential plan in stage one of the game so that the coalition structure results in the next stage can be uniquely determined. Our next lemma suggests that this is not a restrictive assumption, because the regulator can always find an essential plan at any coalition structure ω.

Lemma 3. *For any $\omega \in \Omega$, $\overline{\Theta}(\omega) \neq \emptyset$.*

Proof. Given any $\omega = (m_A, m_B)$, we construct a tax plan θ as follows. Let $\theta_k^C(m_A, m_B)$ be a tax rate such that $u_k^C(m_A, m_B; \theta)$ is maximized, $k = A, B$. Denote $\theta(m_A, m_B) = \arg\min_{k \in \{A,B\}} u_k^C(m_A, m_B; \theta)$. Fix this $\theta(m_A, m_B)$, then $u_k^C(m_A, m_B; \theta)$ and $u_k^O(m_A, m_B; \theta)$ are also given. Define $\theta(m_A + 1, m_B)$ and $\theta(m_A, m_B + 1)$ such that $u_A^C(m_A + 1, m_B; \theta)$ is slightly smaller than $u_A^O(m_A, m_B; \theta)$, and $u_B^C(m_A, m_B + 1; \theta)$ is slightly smaller than $u_B^O(m_A, m_B; \theta)$. Then, define $\theta(m_A + 1, m_B + 1)$ such that both $u_A^O(m_A, m_B + 1) > u_A^C(m_A + 1, m_B + 1)$ and $u_B^O(m_A + 1, m_B) > u_B^C(m_A + 1, m_B + 1)$ hold. Inductively, suppose we have already defined $\theta(s_A + 1, s_B)$ and $\theta(s_A, s_B + 1)$ where $s_A \geq m_A$ and $s_B \geq m_B$, then define $\theta(s_A + 1, s_B + 1)$ to ensure $u_A^O(s_A, s_B + 1) > u_A^C(s_A + 1, s_B + 1)$ and $u_B^O(s_A + 1, s_B) > u_B^C(s_A + 1, s_B + 1)$.

In this manner, we can define $\theta(s_A, s_B)$ for all $s_A \geq m_A$ and $s_B \geq m_B$, satisfying (19)–(22). Other parts of $\theta(s_A, s_B)$ where $s_A < m_A$ or $s_B < m_B$ can be similarly defined in an inductive way. Thus, we have constructed an essential plan $\theta \in \overline{\Theta}(\omega)$. □

Let $\overline{\Theta} = \cup_{\omega \in \Omega} \overline{\Theta}(\omega)$ be the set of essential plans. Given any $\theta \in \overline{\Theta}$, let $\omega(\theta)$ denote the corresponding stable coalition structure. Furthermore, let $\overline{\Theta}^0 = \{\theta \in \overline{\Theta} : \omega(\theta) \neq (0, 0)\}$ denote the set of essential plans that leads to nonempty coalitions. From Lemma 3, both $\overline{\Theta}$ and $\overline{\Theta}^0$ are not empty.

3.3. Stage one

If an essential plan $\theta \in \overline{\Theta}$ is chosen in stage one, then a coalition with structure $\omega(\theta)$ will be formed in stage two, and the expected social welfare from the perspective of the regulator is[7]

$$v(\theta) = V(\omega(\theta); \theta). \tag{23}$$

Similarly, if $\theta \in \overline{\Theta}^0$ is chosen in stage one, then the expected average coalition payoff is[8]

$$y(\theta) = Y(\omega(\theta); \theta). \tag{24}$$

The objective of the regulator in stage one is to maximize either $v(\theta)$ or $y(\theta)$ by selecting an appropriate essential plan. The choice of objective function hinges on the regulator's identity. If the regulator's goal is to enhance the overall well-being of humanity, the social welfare $v(\theta)$ is the appropriate objective function. Conversely, if the regulator represents the coalition and prioritizes the interests of its members, average coalition payoff $y(\theta)$ should be the objective function.

[7] For the completeness of the definition, if $\theta \notin \overline{\Theta}$ and $\overline{\Omega}(\theta) \neq \emptyset$, we define $v(\theta) = \min_{(m_A, m_B) \in \overline{\Omega}(\theta)} V(m_A, m_B; \theta)$. That is, we assume that the smallest value of social welfare will be realized if θ is not essential and there are multiple stable coalition structures relative to θ.

[8] If $\theta \notin \overline{\Theta}$ and there are multiple stable coalition structures $(m_A, m_B) \neq (0, 0)$ relative to θ, we define $y(\theta) = \min_{(m_A, m_B) \in \overline{\Omega}(\theta) \setminus (0,0)} Y(m_A, m_B; \theta)$.

If $\theta^* \in \overline{\Theta}$ exists such that $v(\theta^*) \geq v(\theta')$ for all $\theta' \in \overline{\Theta}$, then θ^* is called an efficient plan; if $\theta^{**} \in \overline{\Theta}^0$ exists such that $y(\theta^{**}) \geq y(\theta')$ for all $\theta' \in \overline{\Theta}^0$, then we call θ^{**} an optimal plan.

To derive efficient/optimal plans, it is convenient to introduce the concepts of locally efficient/optimal plans first. Given $\omega \in \Omega$, a tax plan $\theta \in \overline{\Theta}(\omega)$ is considered locally efficient at ω, if it maximizes the social welfare within $\overline{\Theta}(\omega)$, that is, $V(\omega;\theta) \geq V(\omega;\theta')$ for all $\theta' \in \overline{\Theta}(\omega)$. Likewise, a tax plan $\theta \in \overline{\Theta}^0(\omega)$ is said to be locally optimal at ω, if $Y(\omega;\theta) \geq Y(\omega;\theta')$ for all $\theta' \in \overline{\Theta}^0(\omega)$.

3.4. Discussion

To conclude this section, we examine a critical implicit assumption in the model: the tax plan (coalition rule) is designed by the organizer in stage one and remains fixed thereafter, especially after coalition formation in stage two. Below, we present arguments in support of this assumption.

First, international environmental agreements are fundamentally self-enforcing, meaning countries perpetually retain the right to re-assess their participation decisions. If we allow the regulator to adjust the coalition rule after coalition formation, we must likewise permit countries to reconsider their participation in response to such changes. As a result, altering the original coalition rule to different rules could prompt some coalition members to withdraw, ultimately reducing the regulator's objective value rather than enhancing it. This occurs because the rule change alters the relative payoff structure between being a member versus a non-member. Specifically, the advantage of remaining in the coalition compared to leaving diminishes, disrupting the equilibrium that previously maintained stability and making participation less attractive for some members.[9]

Second, some readers may wonder our tax plan involves incredible threats, because the out-of-equilibrium tax rates serve as punishments to support the equilibrium outcome. In practice, an off-equilibrium outcome may result from a signatory deviating from the coalition due to some perturbation (e.g., a less rational president is elected in a country). In that case, the regulator would eliminate the possibility of its future re-entry by altering the coalition rule to make rejoining unprofitable for that country. Conversely, maintaining the original tax plan preserves incentives for the country to rejoin in the future. For example, while a president's assumption of office might cause a country to exit the coalition, the next president could reverse this decision, provided the coalition rule remains unchanged and continues to offer participation incentives. Therefore, although adherence to the original tax plan may temporarily decrease the organizer's objective value due to external disturbances,[10] it can, in the long term, achieve a higher objective value by creating stronger participation incentives. In this sense, the punishments implicit in our tax plans are indeed credible.

Finally, our approach aligns with established literature in this field. Several existing studies, including (Carraro et al., 2009), Köke and Lange (2017), Mao (2020), and Masoudi (2022), adopt similar assumptions that contract terms remain unchanged after coalition formation, even when potentially more favorable cooperation prospects might exist. This methodological commitment reflects the practical constraints of international agreements where sovereign participants must retain certainty about the terms to which they have committed, regardless of hypothetical improvements that could theoretically be achieved through subsequent modifications.

[9] Mao (2018) presents an example where, after a coalition is formed, changing the original coalition rule to the MTP rule (which aims to maximize the total coalition payoff) leads to some members leaving the coalition. Consequently, this results in a smaller total coalition payoff.

[10] The design of the coalition rule in practice should take the impact of such disturbances into account. See Mao (2020) for a detailed discussion.

4. Efficient plan and optimal plan

The following theorem establishes the existence of efficient plans and optimal plans.

Theorem 1. *There exists an efficient plan and an optimal plan for any* $G\left(n_A, n_B, \lambda_A, \lambda_B\right)$.

Proof. Given (m_A, m_B), $V(m_A, m_B;\theta)$ is a continuous function of θ on the set $\{\theta$: (19) and (20) hold$\}$, which is evidently a nonempty closed set. From the forms of $B_i(X)$ and $C(x_i)$, the maximal value of $V(m_A, m_B;\theta)$, if exists, must be reached as $\theta(m_A, m_B)$ is not larger than a bounded value. Since a continuous function that is defined on a bounded closed set has a maximum value, we can find a tax plan, $\theta^*_{m_A,m_B}$, such that some parts of $\theta^*_{m_A,m_B}$ maximizes $V(m_A, m_B;\theta)$ under the constraints (19) and (20), while other parts of $\theta^*_{m_A,m_B}$ is constructed to ensure that (21) and (22) hold. Thus, we can construct a locally efficient plan $\theta^*_{m_A,m_B}$ at each coalition structure (m_A, m_B).

By comparing these locally efficient plans at different (m_A, m_B), we can find among them a tax plan θ^* such that $v(\theta)$ is maximized. It remains to be proven that θ^* is efficient. Suppose by contradiction that $\theta' \in \overline{\Theta}$ exists such that $v(\theta^*) < v(\theta')$. If $\theta' \in \overline{\Theta}(m'_A, m'_B)$, then $v(\theta') \leq v(\theta^*_{m'_A,m'_B}) \leq v(\theta^*)$, which contradicts $v(\theta^*) < v(\theta')$. Hence, θ^* is an efficient plan.

Similarly, we can prove the existence of an optimal plan by first constructing locally optimal plans $\theta^{**}_{m_A,m_B}$ at all $(m_A, m_B) \neq (0,0)$, and then among them identify the optimal plan as the one that maximizes $y(\theta)$. □

Inspired by the proof of Theorem 1, a two-step algorithm for finding an efficient or optimal plan is as follows. Step 1: construct locally efficient or locally optimal plans at all coalition structures. Step 2: among these tax plans, find the one that maximizes the objective payoff $v(\theta)$ or $y(\theta)$.

To illustrate this algorithm, consider a numerical example $G(3, 2, \lambda_A, \lambda_B)$ where $\bar{\lambda} = 2$. First, let $\lambda_A = 2.2$, $\lambda_B = 1.7$. At $(m_A, m_B) = (2, 1)$, we can construct a locally efficient plan $\theta^*_{2,1}$ by solving the maximization problem $\max_\theta V(2, 1;\theta)$, or a locally optimal plan $\theta^{**}_{2,1}$ by solving $\max_\theta Y(2, 1;\theta)$, both subject to the following constraints derived from (19)–(22):

$$\begin{aligned} &u^O_A(0,0;\theta) \leq u^C_A(1,0;\theta),\ u^O_A(1,0;\theta) \leq u^C_A(2,0;\theta),\\ &u^O_A(0,1;\theta) \leq u^C_A(1,1;\theta),\ u^O_A(1,1;\theta) \leq u^C_A(2,1;\theta),\\ &u^O_A(0,2;\theta) \leq u^C_A(1,2;\theta),\ u^O_A(1,2;\theta) \leq u^C_A(2,2;\theta); \end{aligned} \tag{25}$$

$$\begin{aligned} &u^O_B(0,0;\theta) \leq u^C_B(0,1;\theta),\ u^O_B(1,0;\theta) \leq u^C_B(1,1;\theta),\\ &u^O_B(2,0;\theta) \leq u^C_B(2,1;\theta),\ u^O_B(3,0;\theta) \leq u^C_B(3,1;\theta); \end{aligned} \tag{26}$$

$$\begin{aligned} &u^O_A(2,0;\theta) > u^C_A(3,0;\theta),\ u^O_A(2,1;\theta) > u^C_A(3,1;\theta),\\ &u^O_A(2,2;\theta) > u^C_A(3,2;\theta); \end{aligned} \tag{27}$$

$$\begin{aligned} &u^O_B(0,1;\theta) > u^C_B(0,2;\theta),\ u^O_B(1,1;\theta) > u^C_B(1,2;\theta),\\ &u^O_B(2,1;\theta) > u^C_B(2,2;\theta),\ u^O_B(3,1;\theta) > u^C_B(3,2;\theta). \end{aligned} \tag{28}$$

We list a pair of solutions to these two problems in Table 2.[11] The corresponding objective payoffs are $v(\theta^*_{2,1}) = 36.60$ and $y(\theta^{**}_{2,1}) = 26.51$, respectively.

In this manner, we can derive the values of $v(\theta^*_{m_A,m_B})$ for all locally efficient plans $\theta^*_{m_A,m_B}$, and the values of $y(\theta^{**}_{m_A,m_B})$ for all locally optimal plans $\theta^{**}_{m_A,m_B}$ (see column 1 of Table 3). Since $v(\theta^*_{3,2}) = 49.97 > v(\theta^*_{m_A,m_B})$ and $y(\theta^{**}_{3,2}) = 49.97 > y(\theta^{**}_{m_A,m_B})$ for all $(m_A, m_B) \neq (3, 2)$, $\theta^*_{3,2}$ is efficient and $\theta^{**}_{3,2}$ is optimal. Thus, for $G(3, 2, 2.2, 1.7)$, $\omega(\theta^*) = \omega(\theta^{**}) = (3, 2)$.

Similarly, column 2 and 3 of Table 3 show that for $G(3, 2, 3.06, 0.41)$, $\omega(\theta^*) = (3, 2)$, $\omega(\theta^{**}) = (3, 0)$, and for $G(3, 2, 3.23, 0.16)$, $\omega(\theta^*) = \omega(\theta^{**}) = (3, 0)$. Given $\bar{\lambda}$, the coalition formed under efficient/optimal plans depends on λ_A/λ_B. Specifically, the coalition formed under θ^* may be larger than that under θ^{**}. We will explain this result in Section 6.

[11] All numerical simulations in this paper were conducted using Wolfram Mathematica. Interested readers may request the codes via correspondence.

Table 2
Locally efficient/optimal plans at $(2,1)$ for $G(3,2,2.2,1.7)$.

(m_A,m_B)	(0,1)	(0,2)	(1,0)	(1,1)	(1,2)	(2,0)	(2,1)	(2,2)	(3,0)	(3,1)	(3,2)
$\theta^{*}_{2,1}(\cdot)$	0	0	0	1.48	0.56	0	6.80	8.29	0	5.25	5.97
$\theta^{**}_{2,1}(\cdot)$	0	0	0	1.20	1.45	1.59	4.07	6.39	1.95	5.54	6.86

Table 3
Looking for efficient/optimal plans for $G(3,2,\lambda_A,\lambda_B)$, $\bar{\lambda}=2$.

(m_A,m_B)	$\lambda_A=2.2,\ \lambda_B=1.7$		$\lambda_A=3.06,\ \lambda_B=0.41$		$\lambda_A=3.23,\ \lambda_B=0.16$	
	$v(\theta^{*}_{m_A,m_B})$	$y(\theta^{**}_{m_A,m_B})$	$v(\theta^{*}_{m_A,m_B})$	$y(\theta^{**}_{m_A,m_B})$	$v(\theta^{*}_{m_A,m_B})$	$y(\theta^{**}_{m_A,m_B})$
(0,0)	17.97	–	17.16	–	16.87	–
(1,0)	17.97	19.58	17.16	25.90	16.87	27.06
(2,0)	27.83	22	26.66	30.58	26.03	32.26
(3,0)	36.22	29.26	31.62	**44.60**	**30.64**	**47.87**
(0,1)	19.10	15.29	17.16	4.05	16.88	1.60
(1,1)	26.61	19.47	19.76	16.07	17.92	14.82
(2,1)	36.60	26.51	24.10	24.45	19.77	21.21
(3,1)	43.09	37.95	29.78	35.19	22.34	27.36
(0,2)	26.95	17	20.19	4.13	18.12	1.61
(1,2)	37.28	23.87	24.97	14.25	20.16	11.40
(2,2)	43.89	34.68	31.09	24.81	22.93	18.78
(3,2)	**49.97**	**49.97**	**38.14**	38.14	26.36	26.36

The bold numbers in this table are the corresponding payoffs under efficient/optimal plans.

Table 4
Comparing θ^{*}, θ^{**} to θ^{a}, θ^{b} in $G(3,2,2,2)$.

(m_A,m_B)	$\theta^{*}=\theta^{**}$			θ^{a}			θ^{b}		
	$\theta(\cdot)$	$u^{C}_{k}(\cdot;\theta)$	$u^{O}_{k}(\cdot;\theta)$	$\theta(\cdot)$	$u^{C}_{k}(\cdot;\theta)$	$u^{O}_{k}(\cdot;\theta)$	$\theta(\cdot)$	$u^{C}_{k}(\cdot;\theta)$	$u^{O}_{k}(\cdot;\theta)$
(0,0)	–	–	18	–	–	34	–	–	18
(1,0)	0	18	18	8	−14	34	0	18	18
(2,0)	0.58	18.37	18.78	8	2	50	2	20	26
(3,0)	1.56	19.92	21.00	8	18	66	4	26	42
(0,1)	0	18	18	8	−14	34	0	18	18
(1,1)	0.58	18.37	18.78	8	2	50	2	20	26
(2,1)	1.56	19.92	21.00	8	18	66	4	26	42
(3,1)	3.23	33.17	46.96	8	34	82	6	36	66
(0,2)	0.58	18.37	18.78	8	2	50	2	20	26
(1,2)	1.56	19.92	21.00	8	18	66	4	26	42
(2,2)	3.23	33.17	46.96	8	34	82	6	36	66
(3,2)	8	50	–	8	50	–	8	50	–
$\overline{\Omega}(\theta)$	{(3,2)}			{(0,0)}			{(3,0), (2,1), (1,2)}		
$v(\theta)$	50			18			23.6		
$y(\theta)$	50			–			20		

Now, we compare efficient plan θ^{*} and optimal plan θ^{**} with some other tax plans. Define θ^{a} to be the tax plan that maximizes $V(m_A,m_B;\theta)$ for all (m_A,m_B), and define θ^{b} as the tax plan that maximizes $Y(m_A,m_B;\theta)$ for all $(m_A,m_B)\neq(0,0)$. The key difference between θ^{*} and θ^{a} is that the latter concerns the values of social welfare under all coalitions, whether stable or not, while the former only concerns those under a stable coalition. The difference between θ^{**} and θ^{b} is similar. Note that θ^{b} can also be defined as maximizing the joint payoff of the coalition for all coalitions formed, as assumed by many existing studies.[12]

To compare the difference between these tax plans, we consider an example with symmetric countries $G(3,2,2,2)$ and list the corresponding outcomes in Table 4. From this table, we can see that no country will join the coalition under θ^{a}, while three countries will join the coalition under θ^{b}.[13] In contrast, all countries choose to join the coalition under θ^{*} and θ^{**} (in this example, $\theta^{*}=\theta^{**}$), leading to a larger value of objective payoff than those under θ^{a} and θ^{b}. An advantage of θ^{*} over θ^{a} (or θ^{**} over θ^{b}) is that it abandons the unnecessary constraints on non-stable coalitions, and thus can be more flexibly designed to attract more countries to join the coalition.

[12] For example, Barrett (1994).

[13] Note that θ^{b} is not an essential plan, since $\overline{\Omega}(\theta^{b})=\{(3,0),(2,1),(1,2)\}$. We can define $v(\theta^{b})$ and $y(\theta^{b})$ according to footnote .

5. Properties under small heterogeneity

In this section, we explore the properties of efficient and optimal plans and the corresponding objective payoffs when the degree of heterogeneity is small, that is, when λ_A and λ_B are sufficiently close to each other.

The following proposition shows when $\lambda_A-\lambda_B$ is sufficiently small, efficient and optimal plans lead to full cooperation (all countries join the coalition), and $v(\theta^{*})$ and $y(\theta^{**})$ decrease with the degree of heterogeneity but increase with the level of average benefit.

Proposition 1. *Suppose that θ^{*} is efficient, and θ^{**} is optimal, then $\sigma>0$ exists, such that when $\lambda_A-\lambda_B<\sigma$:*
(a) $\omega(\theta^{*})=\omega(\theta^{**})=(n_A,n_B)$;
(b) $v(\theta^{*})=y(\theta^{**})$;
(c) $v(\theta^{*})$ *and* $y(\theta^{**})$ *decrease with* $\lambda_A-\lambda_B$;
(d) $v(\theta^{*})$ *and* $y(\theta^{**})$ *increase with* $\bar{\lambda}$.

Proof. (a) When $\lambda_A-\lambda_B$ is sufficiently small,

$$V(m_A,m_B;\theta)\approx\frac{m_A+m_B}{n}u^{C}_{A}(m_A,m_B;\theta)+\frac{n-m_A-m_B}{n}u^{O}_{A}(m_A,m_B;\theta),$$

which, from (11) and (12), is a quadratic function of θ. Driving the maximum value of this function, we obtain $v(\theta^{*}_{m_A,m_B})\approx\frac{\bar{\lambda}^2}{2n}[(m_A+m_B)(n-1)^2+2n^2-n]$, where $\theta^{*}_{m_A,m_B}$ is locally efficient at (m_A,m_B).

Because $v(\theta^*_{m_A,m_B})$ increases with m_A and m_B, $\theta^* = \theta^*_{n_A,n_B}$ is efficient, and $\omega(\theta^*) = (n_A, n_B)$.

Similarly, we can prove that when $\lambda_A - \lambda_B$ is very small, $y(\theta^{**}_{m_A,m_B}) \approx \frac{1}{2}\bar{\lambda}^2(m_A+m_B-1)^2+(n-\frac{1}{2})\bar{\lambda}^2$, where $\theta^{**}_{m_A,m_B}$ is locally optimal at (m_A, m_B). Because $y(\theta^{**}_{m_A,m_B})$ increases with m_A and m_B, $\theta^{**} = \theta^{**}_{n_A,n_B}$ is optimal, and $\omega(\theta^{**}) = (n_A, n_B)$.

(b): From (a), when $\lambda_A-\lambda_B$ is small enough, $v(\theta^*) = \frac{n_A}{n}u^C_A(n_A, n_B; \theta^*)+\frac{n_B}{n}u^C_B(n_A, n_B; \theta^*)$. From (8), (10), and (11), we have $v(\theta^*) = (2\lambda^2_A n^3_A n_B + 2\lambda_A n^3_A \lambda_B n_B + \lambda^2_A n^2_A n^2_B + n^2_A \lambda^2_B n^2_B + 4\lambda_A n^2_A \lambda_B n^2_B - \lambda^2_A n_A n_B + 2n_A\lambda^2_B n^3_B - n_A\lambda^2_B n_B + 2\lambda_A n_A \lambda_B n^3_B + 2\lambda_A n_A \lambda_B n_B + \lambda^2_A n^4_A + \lambda^2_B n^4_B)/2n^2 = n^2\bar{\lambda}^2/2 - n_A n_B(\lambda_A - \lambda_B)^2/2n^2$. Similarly, $y(\theta^{**}) = n^2\bar{\lambda}^2/2 - n_A n_B(\lambda_A - \lambda_B)^2/2n^2 = v(\theta^*)$.

(c)(d): From (b), $v(\theta^*) = y(\theta^{**}) = n^2\bar{\lambda}^2/2 - n_A n_B(\lambda_A - \lambda_B)^2/2n^2$. Therefore, $v(\theta^*)$ and $y(\theta^{**})$ decrease with $\lambda_A - \lambda_B$, and increase with $\bar{\lambda}$. □

The last paragraph in Section 4 explains the reason for the full cooperation outcome of Proposition 1(a), which naturally leads to Proposition 1(b). The intuition behind Proposition 1(c) is simple. From (7), with a larger benefit difference $\lambda_A - \lambda_B$, a larger gap between the corresponding tax rates $p_A - p_B$ will be created. Meanwhile, according to Lemma 1(b), socially optimal tax rates $p^*_A = p^*_B$. Therefore, a larger degree of heterogeneity makes coordinating the interests of different types of countries more difficult and obtaining a large value of $v(\theta)$ or $y(\theta)$ less likely.

As for the value of social welfare, the socially optimal level $U(\boldsymbol{p}^*)$ is the largest level that can ever be reached, while $v(\theta^*)$ is the largest level that can be obtained through our three-stage IEA game. We have $U(\boldsymbol{p}^*) \geq v(\theta^*)$, and the equality holds if the socially optimal level can be implemented by an efficient plan. Thus, an important question is under what conditions $v(\theta^*) = U(\boldsymbol{p}^*)$ holds. According to our next result, it only holds when all countries are symmetric.

Proposition 2. *If and only if $\lambda_A = \lambda_B$, $v(\theta^*) = U(\boldsymbol{p}^*)$.*

Proof. Suppose $v(\theta^*) = U(\boldsymbol{p}^*)$. Then, the socially optimal tax rate $\boldsymbol{p}^*$ coincides with efficient plan θ^*. According to Proposition 1(a), and Lemma 1(b), an efficient plan θ^* exists such that $\theta^*(n_A, n_B) = n\bar{\lambda} - \lambda_A = n\bar{\lambda} - \lambda_B$, and thus $\lambda_A = \lambda_B$.

Conversely, if $\lambda_A = \lambda_B$, then from the proof of Proposition 1(b), $v(\theta^*) = n^2\bar{\lambda}^2/2$. On the other hand, from Lemma 1(b), we have $U(\boldsymbol{p}^*) = n^2\bar{\lambda}^2/2$. Thus, $v(\theta^*) = U(\boldsymbol{p}^*)$. □

Further, we can use the index $\mu = \frac{U(\boldsymbol{p}^*)-v(\theta^*)}{v(\theta^*)}$ to characterize how much better the socially optimal level of social welfare is compared to the efficient level. The following proposition concerns the factors that may affect μ.

Proposition 3. *There exists $\sigma > 0$, such that when $\lambda_A - \lambda_B < \sigma$:*
(i) *Given $\bar{\lambda}$, μ increases with $\lambda_A - \lambda_B$;*
(ii) *Given λ_A/λ_B, μ is invariant to $\bar{\lambda}$.*

Proof. (i) It follows from Lemma 1(b) that $U(\boldsymbol{p}^*) = n^2\bar{\lambda}^2/2$. From the proof of Proposition 1(b), we have $v(\theta^*) = n^2\bar{\lambda}^2/2 - n_A n_B(\lambda_A - \lambda_B)^2/2n^2$. Since $v(\theta^*)$ decreases with $\lambda_A-\lambda_B$, $\mu = \frac{U(\boldsymbol{p}^*)-v(\theta^*)}{v(\theta^*)}$ increases with $\lambda_A-\lambda_B$.

(ii) Again, use $U(\boldsymbol{p}^*) = n^2\bar{\lambda}^2/2$, $v(\theta^*) = n^2\bar{\lambda}^2/2 - n_A n_B(\lambda_A - \lambda_B)^2/2n^2$. For simplicity, we write $\lambda_{AB} = \lambda_A/\lambda_B$. Then, $\lambda_A = \lambda_{AB}\frac{n\bar{\lambda}}{n_A\lambda_{AB}+n_B}$, $\lambda_B = \frac{n\bar{\lambda}}{n_A\lambda_{AB}+n_B}$, and $\lambda_A - \lambda_B = \frac{n\bar{\lambda}(\lambda_{AB}-1)}{n_A\lambda_{AB}+n_B}$. For fixed λ_{AB}, $\mu = \frac{U(\boldsymbol{p}^*)-v(\theta^*)}{v(\theta^*)}$ is independent of $\bar{\lambda}$. □

Part (i) of Proposition 3 shows that if $\bar{\lambda}$ is fixed, μ increases with the degree of heterogeneity. To reiterate, this result holds because larger heterogeneity results in more difficulty in reconciling different countries. Provided that λ_A/λ_B is fixed, part (ii) of this proposition shows that average benefit $\bar{\lambda}$ has the same impact on $v(\theta^*)$ and $U(\boldsymbol{p}^*)$, and hence does not affect μ.

The results in this section are distinct from the existing literature in two respects. On the one hand, McEvoy and McGinty (2018) found that only a small fraction of countries will choose to join the coalition under a tax system that applies a uniform tax rate to all signatories to maximize the coalition payoff. However, we show that by introducing a tax system that maximizes the corresponding payoff only under a stable coalition, full cooperation can result when heterogeneity is small enough. By removing unnecessary requirements on non-stable coalitions, our tax plan has more flexibility in attracting more countries to participate in the coalition. On the other hand, contrary to Weitzman (2014), we find that it is impossible to achieve the socially optimal outcome through our tax system when countries are asymmetric.

Note that Weitzman (2014) and McEvoy and McGinty (2018) assume more general benefit function $B_i(X) = \lambda_i X - \frac{\beta}{2}X^2$ and cost function $C_i(x_i) = c_i x_i + \frac{\gamma}{2}{x_i}^2$ than our functions $B_i(X) = \lambda_i X$ and $C_i(x_i) = \frac{1}{2}{x_i}^2$. Whether our results hold under a more general model setup remains a question. Nevertheless, at least in a special setting ($\beta = 0$, $c_i = 0$, $\gamma = 1$), our findings contrast with the conclusions of these studies. On all accounts, this distinction between results in different models is worthy of more discussion on the respective application conditions of these models.

6. Simulations

It is difficult to solve the model analytically when the degree of heterogeneity is relatively large. Instead, using the algorithm stated in Section 4, we may study the performance of efficient and optimal plans through simulations. An important question is whether the results in Propositions 1 and 3 still hold for a large degree of heterogeneity.

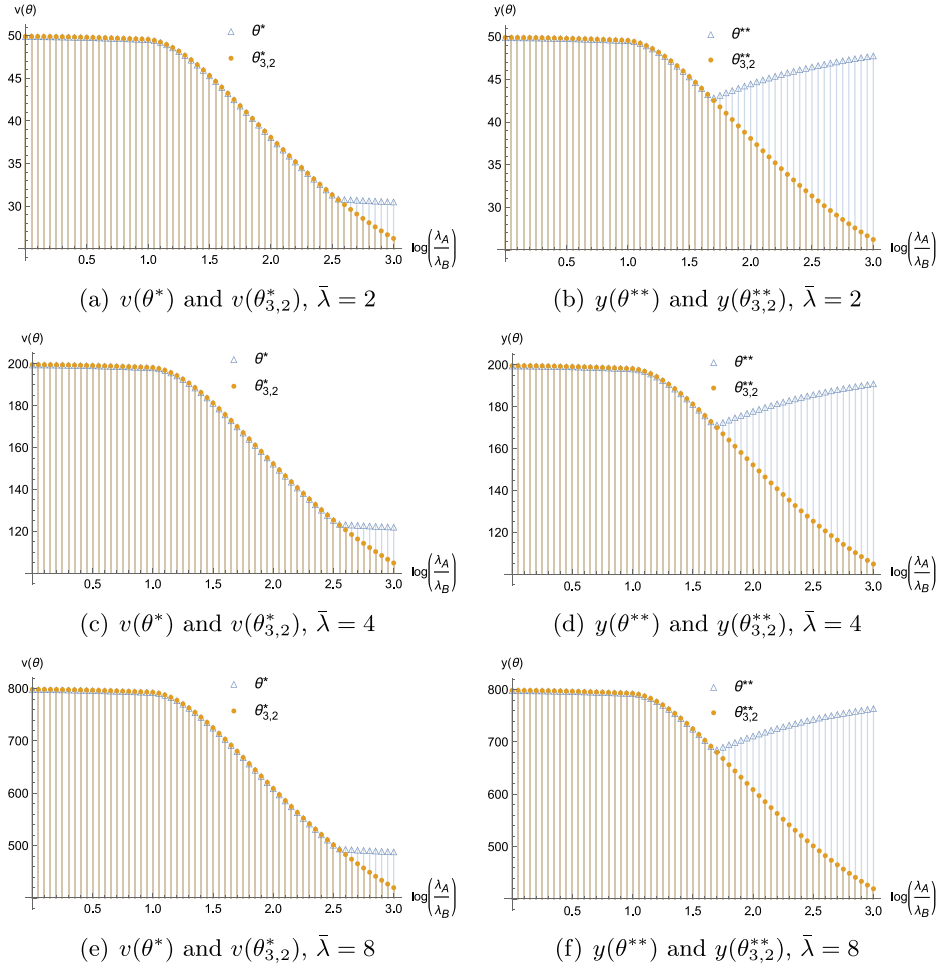
First, we examine the impact of the degree of heterogeneity (measured by $\lambda_A - \lambda_B$ or $\frac{\lambda_A}{\lambda_B}$) on $v(\theta^*)$ and $y(\theta^{**})$. Consider example $G\left(3, 2, \lambda_A, \lambda_B\right)$ where $\bar{\lambda}$ is fixed at 2, 4, or 8. In Fig. 1, we show that for all $\bar{\lambda}$, $v(\theta^*) = v(\theta^*_{3,2})$ and $y(\theta^{**}) = y(\theta^{**}_{3,2})$ when heterogeneity is small enough. This confirms the full cooperation result in Proposition 1(a). However, if heterogeneity is relatively large, then $v(\theta^*) > v(\theta^*_{n_A,n_B})$, $y(\theta^{**}) > y(\theta^{**}_{n_A,n_B})$, implying that a coalition smaller than the grand coalition will be formed under θ^* or θ^{**}. Additionally, Fig. 1 also confirms the monotonicity in Proposition 1(c) and shows that it may no longer hold for θ^{**} when heterogeneity is sufficiently large.

Next, we investigate the impact of the average marginal benefit $\bar{\lambda}$ on $v(\theta^*)$ and $y(\theta^{**})$. Fig. 2 illustrates the simulation outcomes of example $G\left(3, 2, \lambda_A, \lambda_B\right)$ where λ_A/λ_B is fixed at 2, 4, 6 or 8. This figure shows that for any given λ_A/λ_B, both $v(\theta^*)$ and $y(\theta^{**})$ are increasing with $\bar{\lambda}$. Thus, Proposition 1(d) can be extended to more general degrees of heterogeneity.

The examples in Figs. 1 and 2 also show the relationship between efficient plan and optimal plan. Simply put, these two tax plans coincide only when the degree of heterogeneity is sufficiently small. For example, we can learn from Fig. 2(a)(b) that with small heterogeneity λ_A/λ_B, $v(\theta^*) = y(\theta^{**})$ for all $\bar{\lambda}$, which is consistent with Proposition 1(b). However, with large λ_A/λ_B, we see from Fig. 2(c)(d) that $v(\theta^*) < y(\theta^{**})$. Intuitively, this is because an optimal plan only addresses the interests of signatories, while an efficient plan concerns the overall payoffs of all countries. Under our tax system, there is a large gap between the payoffs of different types of signatories when heterogeneity is large. Sometimes, an optimal plan will ensure that only type A countries join the coalition, while an efficient plan has the incentive to attract type B countries to participate as well, resulting in $v(\theta^*) < y(\theta^{**})$.

Finally, we discuss the impact of $\frac{\lambda_A}{\lambda_B}$ and $\bar{\lambda}$ on μ, the relative welfare gap between socially optimal level $U(\boldsymbol{p}^*)$ and efficient level $v(\theta^*)$. Again, we consider example $G\left(3, 2, \lambda_A, \lambda_B\right)$. From Fig. 3, we can see that μ increases with λ_A/λ_B and is invariant to $\bar{\lambda}$, which generalizes the condition of Proposition 3 to a large range of λ_A/λ_B.

In summary, with a large degree of heterogeneity, the simulations in this section suggest that most results in Propositions 1 and 3 are still true, with some notable exceptions: (a) a partial coalition may be

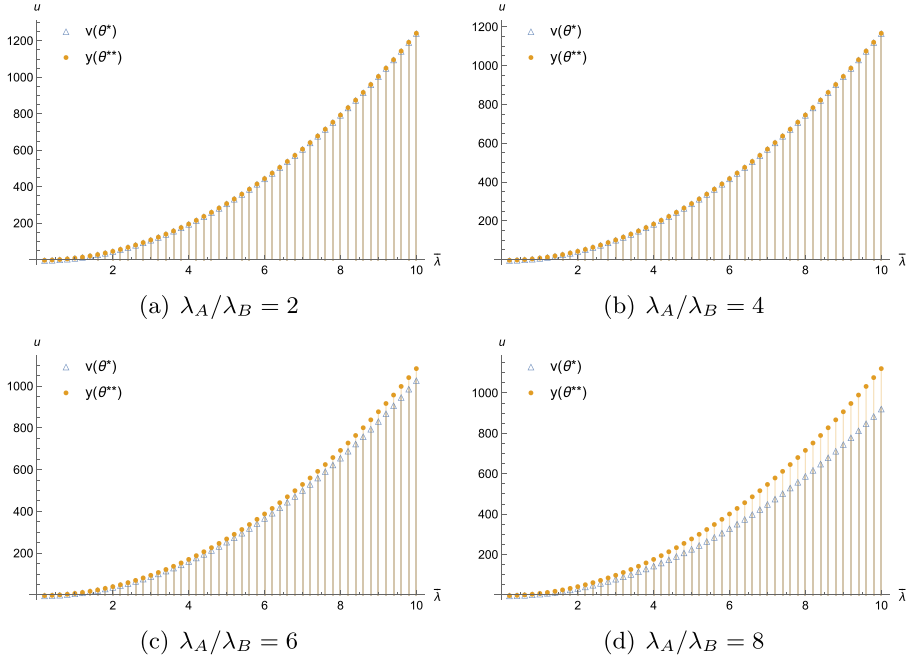


(a) $v(\theta^*)$ and $v(\theta^*_{3,2})$, $\bar{\lambda} = 2$ (b) $y(\theta^{**})$ and $y(\theta^{**}_{3,2})$, $\bar{\lambda} = 2$

(c) $v(\theta^*)$ and $v(\theta^*_{3,2})$, $\bar{\lambda} = 4$ (d) $y(\theta^{**})$ and $y(\theta^{**}_{3,2})$, $\bar{\lambda} = 4$

(e) $v(\theta^*)$ and $v(\theta^*_{3,2})$, $\bar{\lambda} = 8$ (f) $y(\theta^{**})$ and $y(\theta^{**}_{3,2})$, $\bar{\lambda} = 8$

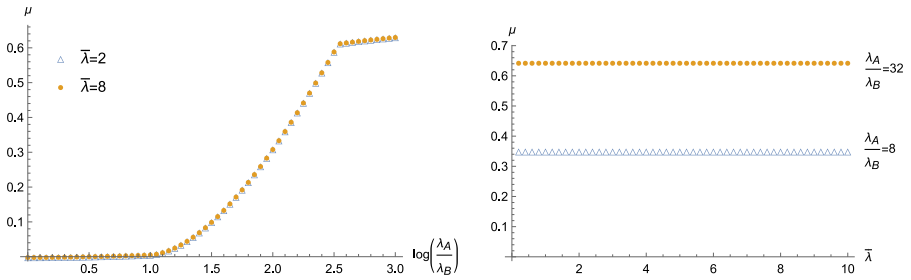
Fig. 1. Impact of λ_A/λ_B on $v(\theta)$ and $y(\theta)$: $G\left(3, 2, \lambda_A, \lambda_B\right)$, $\bar{\lambda}$ fixed.



(a) $\lambda_A/\lambda_B = 2$ (b) $\lambda_A/\lambda_B = 4$

(c) $\lambda_A/\lambda_B = 6$ (d) $\lambda_A/\lambda_B = 8$

Fig. 2. Impact of $\bar{\lambda}$ on $v(\theta^*)$ and $y(\theta^{**})$: $G\left(3, 2, \lambda_A, \lambda_B\right)$, λ_A/λ_B fixed.



(a) $\bar{\lambda}$ is fixed (b) λ_A/λ_B is fixed

Fig. 3. Impact of $\frac{\lambda_A}{\lambda_B}$ and $\bar{\lambda}$ on μ: $G\left(3, 2, \lambda_A, \lambda_B\right)$.

formed under θ^* or θ^{**}; (b) an optimal plan θ^{**} may induce only type A countries to participate, while an efficient plan θ^* sometimes attracts more signatories than θ^{**} does.

7. Conclusion

We examine the regulation of carbon abatement through an endogenously designed tax system in an IEA model with two types of countries that differ in their abatement benefits. To this end, we extend the traditional IEA game and add a preceding stage in which a tax plan is designed, and develop an algorithm to find a tax plan that maximizes social welfare or average coalition payoff under a stable coalition. In contrast, some early IEA models concern the corresponding payoffs for all possible coalitions and thus impose more restrictions on the design of IEA rules than our model does. By abandoning these redundant requirements on non-stable coalitions, our tax plans are more flexible than traditional IEA rules. This leads to two main theoretical contributions of this study.

First, our tax plans are better at creating more cooperation and can result in a preferable outcome than traditional tax systems in terms of coalition size and abatement level. Specifically, full cooperation results when heterogeneity is sufficiently small. However, the performance of the IEA may degrade when heterogeneity increases, mainly because heterogeneity somehow creates difficulty in reconciling different countries when they voluntarily sign the IEA (especially in the absence of international transfers).

Second, if a tax plan is properly designed in stage one of the game, the coalition structure that results in stage two can be uniquely identified. These tax plans (named essential plans in this study) could help us avoid some technical difficulties caused by the non-existence and non-uniqueness of stable coalitions.

Our conclusions have some implications for practical climate policy. To induce a proper ratio of countries to cooperate, a proper tax system should not consist of a single tax rate but instead a list of tax rates that are contingent on the coalition formed and on country's abatement benefit. Except for the one implemented in equilibrium, all other tax rates in this list serve as a rule for punishing free-riding behavior. Also note that even if we can, sometimes it is not efficient to get all countries involved in cooperation, especially for those that hardly benefit from carbon abatement.

Overall, the results of this study provide a highly optimistic assessment of the role a proper tax system can play in international environmental cooperation. However, many factors that could pose practical challenges to the theory are not considered in this paper. For instance, designing a tax plan requires coordination among a large number of diverse countries globally and consideration of the impacts of various uncertainties on the plan.[14] Therefore, the design of a proper tax system in practice remains challenging. On the one hand, we believe that more complex and realistic model setups (for example, with more general benefit function rather than linear benefit function) and tax systems (for example, those that allow for more general tax plan functions, such as $p_i = \theta(\omega) + f\left(\lambda_i, m_i\right)$, where $f\left(\lambda_i, m_i\right)$ is also designed by the organizer) are worthy of future research. On the other hand, given real-world complexities, it is also advisable to further explore simpler and more practical tax plan designs. For instance, a tax rate could be based solely on the ratio of the total emissions of all signatory countries to global emissions, instead of on the coalition structure.

[14] For example, parameter uncertainty, participation uncertainty, and so on. See Na and Shin (1998), Fujita (2004), Kolstad (2007), Dellink et al. (2008), Hong and Karp (2014), Nkuiya et al. (2015), Meya et al. (2018), and Mao (2020) for analyses of different types of uncertainty in IEAs.

CRediT authorship contribution statement

Ping Qiu: Writing – review & editing, Software, Data curation. **Liang Mao:** Writing – review & editing, Writing – original draft, Formal analysis, Conceptualization.

References

Bakalova, I., Eyckmans, J., 2019. Simulating the impact of heterogeneity on stability and effectiveness of international environmental agreements. European J. Oper. Res. 277, 1151–1162.
Barrett, S., 1994. Self-enforcing international environmental agreements. Oxf. Econ. Pap. 46, 878–894.
Carraro, C. (Ed.), 2003. The Endogenous Formation of Economic Coalitions. Edward Elgar.
Carraro, C., Marchiori, C., Oreffice, S., 2009. Endogenous minimum participation in international environmental treaties. Environ. Resour. Econ. 42 (3), 411–425.
Carraro, C., Siniscalco, D., 1993. Strategies for the international protection of the environment. J. Public Econ. 52, 309–328.
Cramton, P., Ockenfels, A., Stoft, S., 2015. An international carbon-price commitment promotes cooperation. Econ. Energy Environ. Policy 4, 51–64.
d'Aspremont, C., Jacquemin, A., Gabszewicz, J.J., Weymark, J., 1983. On the stability of collusive price leadership. Can. J. Econ. 16, 17–25.
Dellink, R., Finus, M., Olieman, N., 2008. The stability likelihood of an international climate agreement. Environ. Resour. Econ. 39 (4), 357–377.
Finus, M., 2001. Game Theory and International Environmental Cooperation. Edward Elgar.
Finus, M., McGinty, M., 2019. The anti-paradox of cooperation: Diversity may pay!. J. Econ. Behav. Organ. 157, 541–559.
Fuentes-Albero, C., Rubio, S.J., 2010. Can international environmental cooperation be bought? European J. Oper. Res. (ISSN: 0377-2217) 202 (1), 255–264.
Fujita, T., 2004. Design of international environmental agreements under uncertainty. Environ. Econ. Policy Stud. 6, 103–118.
Hoel, M., 1992. Carbon taxes: An international tax or harmonized domestic taxes? Eur. Econ. Rev. 36 (2), 400–406.
Hong, F., Karp, L., 2014. International environmental agreements with endogenous or exogenous risk. J. Assoc. Environ. Resour. Econ. 1 (3), 365–394.
Köke, S., Lange, A., 2017. Negotiating environmental agreements under ratification constraints. J. Environ. Econ. Manag. 83, 90–106.
Kolstad, C., 2007. Systematic uncertainty in self-enforcing international environmental agreements. J. Environ. Econ. Manag. 53 (1), 68–79.
Kolstad, C., Ulph, A., 2011. Uncertainty, learning and heterogeneity in international environmental agreements. Environ. Resour. Econ. 50, 389–403.
Mao, L., 2018. A note on stable cartels. Econ. Bull. 38, 1338–1342.
Mao, L., 2020. Designing international environmental agreements under participation uncertainty. Resour. Energy Econ. 61, 101167.

Masoudi, N., 2022. Designed to be stable: international environmental agreements revisited. Int. Environ. Agreements: Politics, Law Econ. 22, 659–672.
McEvoy, D.M., McGinty, M., 2018. Negotiating a uniform emissions tax in international environmental agreements. J. Environ. Econ. Manag. (ISSN: 0095-0696) 90, 217–231.
Meya, J.N., Kornek, U., Lessmann, K., 2018. How empirical uncertainties influence the stability of climate coalitions. Int. Environ. Agreements: Politics, Law Econ. 18, 175–198.
Na, S., Shin, H., 1998. International environmental agreements under uncertainty. Oxf. Econ. Pap. 50, 173–185.
Nkuiya, B., Marrouch, W., Bahel, E., 2015. International environmental agreements under endogenous uncertainty. J. Public Econ. Theory 17, 752–772.
Nordhaus, W.D., 2006. After kyoto: Alternative mechanisms to control global warming. Am. Econ. Rev. 96 (2), 31–34.
Pavlova, Y., de Zeeuw, A., 2013. Asymmetries in international environmental agreements. Environ. Dev. Econ. 18 (1), 51–68.
Pearce, D., 1991. The role of carbon taxes in adjusting to global warming. Econ. J. 101, 938–948.
Ulph, A., Pintassilgo, P., Finus, M., 2019. Uncertainty, learning and international environmental agreements: The role of risk aversion. Environ. Resour. Econ. 73, 1165–1196.
Weitzman, M.L., 2014. Can negotiating a uniform carbon price help to internalize the global warming externality? J. Assoc. Environ. Resour. Econ. 1, 29–49.