EE

Game Theory and International Environmental Cooperation

Michael Finus

NEW HORIZONS IN ENVIRONMENTAL ECONOMICS

General Editors
WALLACE E. OATES
HENK FOLMER

Game Theory and International Environmental Cooperation

NEW HORIZONS IN ENVIRONMENTAL ECONOMICS

General Editors: Wallace E. Oates, *Professor of Economics, University of Maryland, USA* and Henk Folmer, *Professor of General Economics, Wageningen University and Professor of Environmental Economics, Tilburg University, The Netherlands*

This important series is designed to make a significant contribution to the development of the principles and practices of environmental economics. It includes both theoretical and empirical work. International in scope, it addresses issues of current and future concern in both East and West and in developed and developing countries.

The main purpose of the series is to create a forum for the publication of high quality work and to show how economic analysis can make a contribution to understanding and resolving the environmental problems confronting the world in the twenty-first century.

Recent titles in the series include:

Designing International Environmental Agreements
Incentive Compatible Strategies for Cost-Effective Cooperation
Carsten Schmidt

Spatial Environmental and Resource Economics
The Selected Essays of Charles D. Kolstad
Charles D. Kolstad

Economic Theories of International Environmental Cooperation
Carsten Helm

Negotiating Environmental Quality
Policy Implementation in Germany and the United States
Markus A. Lehmann

Game Theory and International Environmental Cooperation
Michael Finus

Sustainable Small-scale Forestry
Socio-economic Analysis and Policy
Edited by S.R. Harrison, J.L. Herbohn and K.F. Herbohn

Environmental Economics and Public Policy
Selected Papers of Robert N. Stavins, 1988–1999
Robert N. Stavins

International Environmental Externalities and the Double Dividend
Sebastian Killinger

Global Emissions Trading
Key Issues for Industrialized Countries
Edited by Suzi Kerr

The Choice Modelling Approach to Environmental Valuation
Edited by Jeff Bennett and Russell Blamey

Uncertainty and the Environment
Implications for Decision Making and Environmental Policy
Richard A. Young

Game Theory and International Environmental Cooperation

Michael Finus
University of Hagen, Germany

NEW HORIZONS IN ENVIRONMENTAL ECONOMICS

Edward Elgar
Cheltenham, UK · Northampton, MA, USA

Published by
Edward Elgar Publishing Limited
Glensanda House
Montpellier Parade
Cheltenham
Glos GL50 1UA
UK

Edward Elgar Publishing, Inc.
136 West Street
Suite 202
Northampton
Massachusetts 01060
USA

A catalogue record for this book
is available from the British Library

Library of Congress Cataloguing in Publication Data

Finus, Michael, 1965–
Game theory and international environmental cooperation/Michael Finus.
(New horizons in environmental economics)
Includes index.
1. Environmental economics. 2. Game theory. 3. Environmental law, International. I. Title. II. Series.
HD75.6 .F56 2000
333.7—dc21 00–039369

ISBN 1 84064 408 7

Printed and bound in Great Britain by MPG Books Ltd, Bodmin, Cornwall

To Kerstin

Contents

Figures

Tables

Acknowledgments

In writing this book I have benefitted from the help of many people. Most of all, I am indebted to my mentor at the University of Hagen, Professor Alfred Endres. First, I benefitted from many discussions on the subject and received many comments that substantially improved the selection of topics and the style and clarity of the exposition, and helped me to avoid many errors. Second, I learnt how to write papers in economics by being involved in many joint projects with him. Third, he provided me with an ideal working environment, reducing my workload for the department to the minimum, so that I found enough time to conduct my research and write this book. Fourth, I received plenty of encouragement during times in which I felt progress was slow.

I also feel obliged to my second mentor at the University of Hagen, Professor Arnold, for carefully reading through the manuscript and providing me with many comments and hints that helped me to avoid some mistakes.

I also would like to thank my former mentor at the University of Giessen, Professor Roland Herrmann, who was the first to stimulate my interest in research during my studies of Agricultural Economics. He encouraged me to orientate my research towards international standards.

Moreover, Bianca Rundshagen, my colleague and co-author of some papers, spent much time in explaining some difficult math with a great amount of patience. I enjoyed working with her very much. Some of the research on which this book is based would, undoubtedly, not have been possible without her. She also read parts of this book and forced me to be more precise with respect to the technical exposition of the subject, helped me to eliminate much nonsense and to avoid some murky prose.

I have also received many comments on preliminary work from scholars at international conferences, which helped to improve the quality of my research. Though I cannot name them all, I would like to particularly mention Professor Henk Folmer (University of Wageningen and Tilburg University), Professor Ayre Hillman (Bar-Ilan University), Professor Michael Hoel (University of Oslo), Professor Rüdiger Pethig (University of Siegen), Professor Sigve Tjøtta (University of Bergen), Professor Henry Tulkens (University of Louvain, CORE), Professor Ekko van Ierland (University of Wageningen) and Professor Heinz Welsch (University of Oldenburg).

I also received many useful comments from my colleagues Cornelia Ohl, Andreas Lüdeke, Jörn Martiensen and Volker Radke during research seminars at the University of Hagen. These seminars functioned as think-tanks and stimulated much of the research I conducted in recent years.

As to the ideal working environment, I was in the fortunate position to receive research support from many student assistants in the department. Frank Lobigs and Frank Brockmeier carefully checked the language, did the formatting of the text and also provided me with many useful comments in structuring the subject. Michael Prinz worked with great enthusiasm to improve the design of the manuscript and drew most of the graphs. Peter Limberger, Rainer Skulimma and Dimitrios Gotsis supported me in many ways, for instance researching literature and drawing graphs.

Another important 'human factor' responsible for the ideal working environment was Gabriele Debray. In doing all the organizational work at the department in Hagen, she reduced my bureaucratic workload, which enabled me to devote more time to this book.

Usually, authors have to master two steps before a book is put on the market: writing a manuscript and getting it ready for publication. However, due to the excellent work of the editors of this series, Professor Henk Folmer and Professor Wallace E. Oates as well as Julie Leppard (Managing Editor), Christine Boniface (Desk Editor), Dymphna Evans (Senior Commissioning Editor) and Alexandra Minton (Assistant Editor) at Edward Elgar, the second step was an easy one. The highly professional work of these people made it an extremely enjoyable task to jointly prepare the manuscript for publication at every stage.

Most parts of this book are based on a research project that was financed by the Volkswagen-Foundation, grant number II 69 982. Their generous support is gratefully acknowledged.

Finally, I owe a great debt to my friend Kerstin and to my parents who had to put up with the fact that I did not have very much time for them over the last couple of months, but still provided me with the psychological support I needed to finish this book.

Of course, needless to say, any blame for remaining mistakes and inaccuracies rests with me.

M. Finus
Hagen, March 2000

1. Introduction

International environmental problems have received increasing attention from economists in recent years. Two basic strands of literature may be distinguished. The first strand estimates the costs and benefits of various abatement targets under different cost allocation rules. It also discusses institutional issues and the design of treaties with respect to their efficiency. In particular the problem of global warming caused by the so-called greenhouse gases (see, for example, Brunner 1991; Cansier 1991; Chapman and Drennen 1990; Cline 1992a, b; Crosson 1989; Grubb 1989; IPCC 1996a, b; Manne and Richels 1991; Michaelis 1992; Nitze 1990; Nordhaus 1991a, b, c; Schelling 1991; and Welsch 1995) and the 'acid rain' problem due to sulfur and nitrogen oxides (for example, Crocker 1984; Førsund and Naevdal 1994; Foster 1993; Newbery 1990; Tahvonen *et al.* 1993; and Welsch 1990) have been studied in depth.

The second strand of literature has approached the problem of international pollution control from a game theoretical perspective (for example, Alho 1992; Andersson 1991; Buchholz and Konrad 1994; Chen 1997; Endres and Finus 1998a; Heister 1998; Kölle 1995; Kuhl 1987; Müller-Fürstenberger and Stephan 1997; Van der Ploeg and de Zeeuw 1992; and Welsch 1993). The incentive scheme of countries which sign a treaty and the stabilization of international environmental agreements (IEAs) are typical issues analyzed by this literature. This book is in the tradition of this second strand of literature.

Since, broadly speaking, game theory analyzes the interaction between agents and formulates hypotheses about their behavior and the final outcome in games, international environmental problems are particularly suited to analysis by these methods. Global and transboundary emissions exhibit a negative externality not only in the country of origin but in other countries too. Hence, there is a high interdependence between countries and strategic considerations enter the scene. Strategic aspects are particularly important in international pollution control since there is no 'world government' which could enforce IEAs. Therefore, the free-rider problem is a distinguishing feature of international environmental policy. Though countries are usually better off by coordinating their environmental policy, cooperation is often difficult to achieve. Since each country has only a marginal effect on aggregate emissions, it is always

better off letting others do the abatement job, thereby saving abatement costs.

Consequently, it is interesting to analyze the causes of this free-rider phenomenon and to derive conditions under which IEAs can, nevertheless, be stabilized. Thus, the central question of this book may be stated as follows: How can cooperation between countries to fight international pollution be established? The following analysis will focus on six aspects which crucially determine the prospects for cooperation. The first aspect concerns the *cost–benefit structure* of emission control. As a central result it will turn out that whenever cooperation generates high global welfare gains, IEAs achieve only little and are plagued by instability.

The second aspect is related to the *time dimension* of a game. Most generally, it will turn out that it is conducive to the stability of an IEA if IEAs are based on a long-term relationship between governments, if treaty obligations are regularly monitored and if instant reactions to a violation of a treaty are possible. In contrast, environmental projects suffer from high instability if they are connected with high sunk costs that involve a substantial time-lag to alter abatement strategies.

The third aspect concerns the *punishment options* in a game and the *credibility of threats*. Of course, the harsher the available punishments are, the easier it is to neutralize the free-rider incentive. However, it will turn out that, in the context of international environmental problems, punishments are normally not only detrimental to the punished but to the punishers as well. Hence, the question arises: Will a government be deterred by a threat of punishment which it believes will never be carried out since the punishers would hurt themselves? To use a metaphor: do we believe that a soldier would pull the pin out of a hand grenade if he will kill himself? The issue of the credibility of threat strategies will be given particular attention in this book. It is closely related to the definitions of equilibrium concepts. We shall discuss several equilibrium concepts which have emerged over recent years as refinements of the central concept in game theory: the Nash equilibrium.

The fourth aspect concerns the *enlargement of the strategy space* and is closely related to the term *issue linkage*. In reality it can frequently be observed that environmental issues are not negotiated in isolation but in connection with other policy fields. Concessions on one issue are traded against concessions on other issues. Such package deals may have several explanations. One is that they avoid possible asymmetric distributions of the gains from cooperation. Another is that issue linkage may ease the enforcement of an agreement. A government may threaten to withdraw from all agreements if one treaty is violated. This may increase the threat potential and thereby the stability of a treaty. It will be interesting to find

out in which situations issue linkage is conducive to cooperation (as is commonly believed) and whether there are conditions under which issue linkage should not be recommended to international negotiators.

The fifth aspect deals with the *institutional framework* in which negotiations take place. This issue evolves around the *instrumental choice* in international pollution control. We shall present a simple bargaining game in which governments bargain on the level of a uniform emission reduction quota and a uniform effluent charge. It will be shown in various chapters that the commonly believed superiority of a tax (market-based instrument) over a quota regime (command and control instrument) no longer holds in a second-best world where this second-best world is constituted by the following restrictions: the accession to an IEA must be voluntary; governments settle for the lowest common denominator proposal and agree on that institutional framework which a majority of governments favor; and the stability of an IEA must be enforced. The model provides many reasons for the fact that emission reduction quotas are part of most of the IEAs signed so far, but as yet no effluent charge has been applied in international pollution control, although economists strongly advocate market-based instruments on efficiency grounds.

The sixth aspect is concerned with *coalition formation* in international pollution control. The central three questions to be answered are: (a) how many and which countries sign an IEA? (b) will there be one or several coexisting agreements? and (c) how effective will agreements be?

The answers turn out to be highly complex since the set of possible government strategies in an N-country world is almost infinitely large. This is particularly true in negative externality games since the strategies of a coalition depend on the strategies of all other coalitions, which in turn depends on the overall coalition structure in the game. The final objective is to explain the entire coalition formation process endogenously.

The subsequent analysis assumes transboundary or global emissions which are summarized under the term *international environmental problems*. Typical examples of *transboundary pollutants* include the aforementioned acid rain or the salination of the river Rhine due to the potash mines in France (upstream country) from which Germany (downstream country) suffers. The first example is a *multilateral externality*, the second example a *unilateral externality*. Typical examples of *global pollutants* include the aforementioned greenhouse gases or substances which deplete the ozone layer.[1]

Transboundary pollutants are also classified as *impure public bads* since the distribution of pollutants is usually uneven across countries. Accordingly, global pollutants belong to the group of *pure public bads* since emissions mix uniformly in the atmosphere and all countries suffer from the

externality approximately equally.[2] According to this classification it comes as no surprise that, apart from the environmental economics literature, the literature on the economics of public goods has also investigated the problem of cooperation and free-riding in the provision of public goods, though this literature does not explicitly refer to international environmental problems. The provision of a public good, say x, depends on the sum of contributions of each agent i, that is, $x = \Sigma x_i$, and agent i derives utility from x, that is, $u_i(x)$. If x_i is interpreted as abatement or emission reduction from some status quo, then it is obvious that a public goods model can be used to describe the structure of international environmental problems studied in this book. Therefore, some of the approaches in the public goods literature to explain the voluntary provision of public goods will be investigated in Chapter 10 in a game theoretical context.

The basic situation which will be analyzed subsequently may be described as follows. Two or more countries emit a transboundary or global pollutant. They are currently not cooperating but are considering doing so. Each government pursues its interests non-cooperatively, that is, it behaves individually rationally. This implies that governments only cooperate if it is in their interest, and that they always take a free-ride whenever this seems profitable to them. Implicitly, we assume that governments maximize some kind of welfare function that measures the gains from emissions accruing from the production and consumption of goods and the losses of emissions accruing from environmental damages. Thereby, the aggregation of welfare is treated as a black box. That is, we abstract from problems mentioned in the public choice literature concerning the interaction of interest groups, voters and bureaucracy and their effect on governmental decisions (Endres and Finus 1996a, b; Ursprung 1992).

Since there is no international agency which could enforce an agreement among sovereign countries, any IEA must be *self-enforcing*. The methods used to analyze such a situation will be taken from game theory, particularly from non-cooperative game theory.[3] The approach followed in this book thereby strongly adheres to the principles of neoclassical economics.

This kind of reasoning has often been attacked as being unrealistic. In particular, the postulate of rationality has been subjected to severe criticism. The critics question whether human beings are as rational as game theory typically assumes them to be. However, this critique has also been raised against neoclassical economics in general. Since arguments in defense of the postulate of rationality are laid down in mighty words elsewhere, it is not necessary to restate them again here. We would only like to point out that we believe that governments approximately act in a rational manner. One can, however, disagree about the appropriate arguments which should be part of governments' objective function.

Another popular charge against game theory may be summarized in the following slogan: 'With game theory one can prove *any* hypothesis – and its antithesis at the same time, too.' Although this slogan has the charm of being humorous, it nevertheless is false. In game theory, as in any other theory, assumptions drive results. Sticking to assumptions once made, one will never – not even in game theory – find contradictory results. If the slogan is interpreted in a friendlier way, it can be seen as a claim that assumptions should be chosen very carefully; that is, they should be based on some plausibility and should, ideally, be empirically supported. Obviously, this (pretty trivial) claim is not only valid for game theory but applies to all theoretical reasoning.

To refute this last general critique, this book will strongly emphasize the underlying assumptions of the models analyzed. Moreover, the way in which they drive the results will be revealed explicitly. With a 'critical distance', the assumptions will be evaluated with respect to their aptness to reflect a particular problem. An overview of the most important assumptions is given in Chapter 2, Section 2.3.

There are two types of game theoretical literature with respect to the mathematical level of the presentation. The first type is highly technical, which makes it difficult for the mathematically less interested reader to access the material. This literature has dominated the game theoretical scene for a long time, which probably explains the fact that widespread applications to economics have only occurred within the last twenty years, though the roots of game theory can be traced back to J. von Neumann and O. Morgenstern (1944). The second type of literature could be responsible for the rapid expansion of applications of game theory to economics in recent years.[4] It explains the central results of game theory intuitively but applies less rigor to formal proofs. We draw attention in particular to the books of Eichberger (1993); Gibbons (1992); Holler and Illing (1993); Kreps (1990); Myerson (1991); and Rasmusen (1995). Simple applications to environmental economics may be found in Burger (1994); Fees (1995); and Weimann (1995).

This book is intended to bridge the gap between these two strands of literature. It starts from the basics of non-cooperative game theory (Chapter 3) and then progresses step by step to the most advanced topics of coalition formation of recent years (Chapter 15). It is aimed at readers with a basic knowledge of microeconomics and mathematics[5] and virtually no knowledge of game theory. Therefore, Chapter 2 introduces the reader to the most important terms used subsequently; other terms will be introduced in the course of reading the book. The proofs of most propositions are provided either in the text or in appendices: exceptions are proofs which are either obvious or would need too much space. Though the intuition of

all proofs is emphasized, those readers who are less interested in the technical part of a proof should be able to understand the central idea of all results from reading the text. Thus, this book has the, surely, heroic aim of being self-contained.

In what follows, Chapter 2 introduces some important terms of game theory (Section 2.1) and some frequently used notations in this book (Section 2.2). The second chapter also provides a taxonomy of game theory (Section 2.3). In the course of the discussion of this taxonomy it will be emphasized which parts of game theory are covered in this book. Therefore, the outline of the book has been left to Section 2.4.

NOTES

1. For a complete classification of environmental problems, see Siebert (1985, 1992).
2. However, this does not imply that environmental damages are equally perceived and evaluated.
3. The details of this setting and the terms mentioned will be explained in Chapter 2.
4. In the environmental economics context, see, for instance, the volumes edited by Hanley and Folmer (1998) and Pethig (1992).
5. Though some proofs may be very long and therefore may look tricky, basic algebra is sufficient to follow all the proofs.

2. Important terms, notation and classification of games

2.1 TERMS

The essential elements of a game are the players, actions, strategies, outcomes, payoffs, equilibria, the information and the order of the game (see, for example, Rasmusen 1989, pp. 23ff.). The *players* are the actors in the game who take decisions. In the international environmental context the players are countries or the political representatives of countries, such as politicians or diplomats. Players can take *actions*, such as making catalytic converters for automobiles mandatory or not.

In contrast, a strategy is a complete plan of action for each contingency which might arise during the game. In a game comprising several stages, a strategy specifies how a player reacts at each point in time to all possible actions of fellow players. For instance, a participant to an IEA must specify an answer if a signatory to an IEA complies with its obligations but also if it violates the agreement.

A particular combination of actions (resulting from the play of some strategy combination) leads to the *outcome* of the game. For instance, in the catalytic converter example outcomes could be measured as the nitrogen oxide concentration in the air, which depends on how many and which governments introduce stricter car regulations. Alternatively, the outcome could also be measured with respect to some other environmental index. The choice will depend on the focus of the analysis (Rasmusen 1989, p. 25).

However, since game theory is mainly concerned with predicting which of the possible outcomes will emerge in a game and which strategies will be played in equilibrium, it is more important to have information on how players evaluate these outcomes. Therefore, the outcomes must be transformed into some form of utility. The utility derived from an outcome is called the *payoff* to a player.

If all action combinations and their associated payoffs are known in a game, the possible strategies for each player can be determined. If each player chooses an equilibrium strategy, this strategy combination leads to the *outcome* and the *equilibrium of the game*, which is sometimes also called

the *equilibrium point of a game*.[1] The prediction of an equilibrium depends on the assumptions regarding the behavior of players. For this it is important to define the exact notion of an equilibrium. For example, it is not sufficient to say that an equilibrium strategy is a 'best strategy' for each player, delivering the highest payoff. It is also necessary to specify whether it is a best strategy independently of what other players do, a best strategy if others also choose their 'best strategy', a best strategy with respect to the overall game, at each point in time during the game, along the equilibrium path or with respect to each situation which might arise during the play. Typical *equilibrium concepts* which will be encountered in this book are Nash equilibrium, subgame-perfect equilibrium and renegotiation-proof equilibrium.

Important elements which influence the play in a game are *information* and the *order of the game*.[2] Information refers to what a player knows about own payoffs and strategies and those of fellow players. This also covers whether players can observe actions of fellow players and, if not, how the players conjecture about unknowns. The order of a game refers to the sequence in which actions are taken. Assumptions with respect to both elements crucially influence the formulation of strategies, which in turn affect the equilibrium in a game.

2.2 NOTATION

We refer to a particular player as player i, j or k where $i, j, k \in \{1, ..., N\}$ and $i \neq j \neq k$. The set of all players is denoted by I, that is, $\mathrm{I} = \{1, ..., N\}$.[3]

A particular action is denoted a_{ik}, $k \in \{1, ..., \mathrm{K}_i\}$ and player i's action set is $\mathrm{A}_i = \{a_{i1}, ..., a_{i\mathrm{K}_i}\}$. A player i has K_i strategies where the subscript is necessary since players may have a different number of actions available (for example, $\mathrm{K}_1 \neq \mathrm{K}_2$). This notation implies that the number of actions is finite: that is, actions can be counted. The catalytic converter example mentioned above belongs to such games with *discrete action sets*. In contrast, if the decision is to choose the level of a fuel tax, then players face a *continuous action set*. The tax rate could range from zero to some upper bound which might be given by the choke price at which demand is cut off.

In the case of an infinite or *continuous action set* we write $\mathrm{A}_i = [a_{i1}, a_{i\mathrm{K}_i}]$ where a_{i1} denotes the lower bound and $a_{i\mathrm{K}_i}$ the upper bound of player i's action space.[4] Regardless whether A_i is finite or infinite we denote the set of all actions A, that is, $\mathrm{A} = \cup \mathrm{A}_i$. Since A is the *Cartesian product* of all players' *action spaces*, it is sometimes also written as $\mathrm{A} = \mathrm{A}_1 \times \mathrm{A}_2 \times ... \times \mathrm{A}_N$.

For reference reasons, we shall use the same notational style throughout

the book. The set of elements is denoted by a capital letter and the elements themselves with the same lower-case letter. For indices we use lower-case letters and the last element of an ordered set is denoted by the same capital letter as the set itself. The only exception is the index *i* which runs from 1 to *N* (and not from 1 to I). This is because the letter 'I' also denotes the set of players and it has become a notational convention in game theory and microeconomics to use *N* for the number of players.

In line with this convention, we denote a strategy by s_{ir}, $r \in \{1, ..., R_i\}$, where the first subscript refers to the player and the second to the strategy. S_i refers to country *i*'s strategy space and S is the strategy space of the entire game where $S = S_1 \times ... \times S_N$ or $S = \cup S_i$ (see, for example, Friedman 1986, pp. 23ff.; and Eichberger 1993, p. 64). The number of strategies may be finite or infinite.

If each player plays a particular strategy s_i^* in equilibrium, we write $s^* = (s_1^*, ..., s_N^*)$. If there is a single equilibrium point in a game, we have $S^* = \{s^*\}$ and if there is more than one equilibrium point this may be expressed by $S^* = \{s^{*(1)}, s^{*(2)}, ...\}$. Generally, asterisks will be used throughout the book to indicate an equilibrium. If the emphasis is on a particular concept, initials will be used; for example, s^{N*} for a 'Nash equilibrium'.

The payoff to a player derived from some action combination is denoted $\pi_{il}, l \in \{1, ..., L\}$. More explicitly, we may write $\pi_{il}(a)$, where *a* denotes some action combination. This stresses that a payoff function maps action combinations on payoffs. Since the actions which are chosen depend on the strategies players pursue, we may also write $\pi_{il}(s)$, with *s* denoting some strategy combination.

The payoff set (or payoff space) of player *i* is Π_i, and the set of all payoffs Π. If the payoff space is infinite we write $\Pi_i = [\pi_i^L, \pi_i^U]$, where π_i^L denotes the lower bound and π_i^U the upper bound of player *i*'s payoff space.

2.3 TAXONOMY OF GAME THEORY

In this section a brief overview of the taxonomy of game theory is given as far as it is relevant to this book. Table 2.1 lists the major criteria according to which the subsequent models can be structured (Finus 1997).

Character of Games

Though not uniquely defined in the literature, in principle a *cooperative game* is one in which *binding agreements* can be signed (Eichberger 1993, p. 31; Friedman 1986, pp. 20, 89–90, 112; and Rasmusen 1989, pp. 30ff.). In contrast, if there is no outside party, for example, an international

Table 2.1 Taxonomy of game theory

1. Character of game	(a) cooperative↔(b) non-cooperative
2. Cost–benefit structure	(a) constant sum↔(b) non-constant sum
3. Number of players	(a) 2↔(b) N
4. Strategy space	(a) discrete↔(b) continuous
5. Time horizon	(a) static↔(b) dynamic: (i) finite, (ii) infinite
6. Time dimension	(a) discrete↔(b) continuous
7. Time structure	(a) independent↔(b) dependent
8. Information requirement	(a) complete↔(b) incomplete
9. Sequence of moves	(a) simultaneous↔(b) sequential

organization, which can enforce an agreement the game belongs to the category of *non-cooperative games*.

It is important to note that the term 'non-cooperative' should not be misinterpreted as implying a general conflict between players and that of cooperative games as an absence of conflict. For instance, most bargaining problems in game theory, like the division of a cake between players, are analyzed using concepts of *bargaining theory*, which belong (predominantly) to cooperative game theory, though an obvious conflict between players exists.

Another instance of a non-cooperative game where there is no conflict is a positive externality game. Imagine customers in a restaurant sitting on a terrace, enjoying cows grazing in a meadow nearby. Hence, the production of milk by the farmer exhibits a positive externality for the restaurant owner (by attracting more customers). If the farmer stops producing milk, the restaurant owner can negotiate with the farmer about transfer payments but cannot force him to continue production.

As we stated in the Introduction, since there is no agency at the global level that is empowered to sanction the breach of a contract (which makes cooperation in international pollution control so difficult), almost all aspects of international environmental problems belong to the realm of non-cooperative games (Congleton 1992; Heister 1997, p. 4). We depart from this 'rule' only in Chapters 10 and 13 since some important contribu-

tions based on the assumptions of cooperative game theory have emerged in the literature.[5]

Generally, it is assumed that players pursue their interests non-cooperatively. To phrase it differently, players are assumed to behave *individually rationally* (in the sense of neoclassical economics). Nevertheless, under certain circumstances, it might be possible to reach a '*cooperative solution*' (for example, joint implementation of an emission tax in several countries to curb greenhouse gases) in a non-cooperative game. This is particularly true in dynamic games where players may use threats to enforce a 'cooperative outcome' (Friedman 1986, pp. 90ff.). Thus, one has to be aware of the distinction between a cooperative game and a cooperative solution. The former refers to the *institutional setting* of a game; the latter to the *objective of an agreement*.

The term *cooperative solution* will be used whenever an agreement among some countries is sought to improve joint welfare compared to some non-cooperative benchmark, as for instance the *laissez-faire* status quo. If a solution is called *fully cooperative* or *socially optimal* we mean an agreement which maximizes aggregate welfare according to a *globally rational* strategy. These definitions also comprise solutions where some players lose compared to the status quo (although aggregate welfare increases due to the agreement), and hence we employ the Kaldor–Hicks welfare criterion when evaluating an outcome (Kaldor 1939; Hicks 1940; see also Feldman 1980, pp. 142ff.).[6] It goes without saying that this requires transferable utility between players and a cardinal measurement of utility.

The assumption that utility can be aggregated across players and measured cardinally will *only* be made in the following instances when evaluating the equilibrium of a game, however, *not* when deriving an equilibrium. Exceptions are those non-cooperative games in which mixed strategies are played (see Sections 3.3, 3.6 and 3.7) and the cooperative coalition formation games in Chapter 13. In the former case, only cardinal measurement of utility is required; in the latter case an aggregation of utility must also be possible.

Cost–Benefit Structure

The second criterion in Table 2.1 classifies games as either *constant sum* or *non-constant sum games* (Friedman 1986, pp. 30ff.). In constant sum games players can only gain at the expense of other players and hence cooperation cannot generate additional welfare. This is why they are sometimes also called *strictly competitive games*. A typical example of a constant sum game is matching pennies: player 1 gains if head–head or tail–tail appears, otherwise player 2 gains. Since we are concerned with the formation and

stability of IEAs in this book, international environmental policy games must belong to the category of non-constant sum games. As will become apparent, in negative externality games cooperation can generate additional welfare.

Number of Players

Though most international environmental problems involve several countries, extending the analysis from *two players* to N *players* ($N > 2$) introduces a bundle of complications. The reason is twofold. First, two-player games usually ease the graphical exposition of a problem. Second, in an N-player game the number of possible strategies which have to be considered when searching for the solution of a game increases more than proportionally. Once players do not behave as singletons and are allowed to form sub-coalitions, strategies of all possible sub-coalitions also have to be analyzed. For instance, in a game with only three players, already five possible coalition structures can emerge:

$$\{\{1\}, \{2\}, \{3\}\}, \{\{1, 2\}, \{3\}\}, \{\{1, 3\}, \{2\}\}, \{\{1\}, \{2, 3\}\}, \{\{1, 2, 3\}\}.$$

The first possibility assumes that each player forms a coalition by him- or herself. The next three possibilities are coalitions among two players. Finally, we have the *grand coalition*, which is a coalition comprising all players. Hence, it is easy to perceive that things get rather complicated if N is large (for example, for $N = 4$ there are 15 possible permutations). Therefore, some simplifying assumptions may be necessary to determine an equilibrium in a coalition game. For instance, many papers on coalition formation in international pollution control assume symmetric countries (see Chapters 13 and 15). Moreover, other plausible assumptions, for example, about the behavior of countries, can also reduce the complexity of a game, making it possible to solve the game eventually (see Chapters 13–15).

Due to these complications we shall proceed stepwise. In Chapters 3–12 we either contemplate only two players or, in the case of N players, strategic aspects of the formation of sub-coalitions are discarded. Subsequently, in Chapters 13–15, the aspect of coalition formation is taken up separately.

Strategy Space

The fourth criterion refers to whether a game has a *discrete* or a *continuous strategy space*. The difference has already been discussed in Section 2.2 and therefore no further explanation is needed. A discrete strategy space allows

one to illustrate many results and concepts of game theory with the help of simple matrix games. This is why Chapters 3–8 exclusively focus on discrete decisions, and only from Chapter 9 onward is the analysis extended to cover continuous choices as well. However, both assumptions are less distinct than one would assume. By increasing the number of actions, which also raises the number of strategies in a game, a discrete strategy space transforms gradually into a continuous strategy space. Moreover, when players are allowed to *randomize* between strategies (mixing of strategies), that is, playing several strategies with some probability, a discrete strategy space turns into a continuous strategy space. In other words, the strategy space can be *convexified* through mixing strategies (see Section 3.6).

Time Horizon

One can basically distinguish between *static* and *dynamic games*. Static games are also called *one-shot games* (Chapters 3 and 9). Dynamic games can further be divided into *finite* (Chapters 4, 6 and 10) and *infinite* (Chapters 5, 7, 12 and 14) games. In finite games the game is played over some (limited) time and the termination of the game is known with certainty. In contrast, in infinite games the game either lasts until perpetuity or the end of the game is not known with certainty. There is also a particular type of game which does not belong to either of these categories. Basically, these are *static representations of dynamic games*, but where a *time explicit* story is missing. *Time implicit* models are treated in Chapters 10, 13 and 15.

The complexity of dynamic games is higher than that of static games. This is immediately apparent when recalling the definition of a strategy. The longer the time horizon, the more contingencies can arise during a game for which a strategy must specify an action. Hence, it is sometimes useful to approximate a situation as a one-shot game provided this simplification 'preserves' the structure of the game. For instance, if the investment in a new and cleaner power plant is associated with high sunk costs, the investment decision might be modeled as a one-shot game. This decision is not reviewed regularly by politicians and might be based on the net present value of the payoff stream of this project.

In contrast, if an industrial country has to decide whether and how much to contribute to an international environmental fund designed to foster environmental projects in developing countries, the decision might be better modeled as a dynamic game.[7] Each year parliament debates the budget for the forthcoming year. If this fund has been set up only for a certain number of years (for example, because it was earmarked only for the initial stage of an agreement), then the game is finite. However, if this

fund has been established without any further specification of its resolution, the game should be viewed as an infinite game.

In the context of dynamic games one can distinguish between two classes of strategies: strategies that account for past actions, which are called *closed loop strategies* (sometimes also called *feedback strategies*), and those that ignore the past, which are called *open loop strategies* (Rasmusen 1989, p. 100). Though open loop strategies are simpler to analyze, in most games there is no reason why players should not use historical information. Therefore, we do not restrict our attention to open loop strategies.

An immediate prerequisite for closed loop strategies is the assumption of *perfect recall* of all past actions, that is, complete memory of the *history of the game* (Brandenburger and Dekel 1989). This assumption seems to be justified in the context of international environmental agreements because all relevant data, as for instance historical emission levels, is statistically reported.[8, 9]

Time Dimension

The time dimension refers to the feature whether time is counted in *discrete time intervals* ($t = 0, 1, 2, ..., T$), also called *periods*, or whether a game is viewed in *continuous time*. Discrete time intervals have two implications. First, if the violation of a treaty is detected, punishment can take place only one period later at the earliest and will therefore be delayed. Second, a country which violates a treaty receives a transitory gain before being punished, which makes free-riding attractive. In contrast, in continuous time actions can be taken at each instant (if no time lag is explicitly introduced into the decision process) and hence free-riding is less of a problem. Since political decisions usually take time and since we believe free-riding is a feature by which IEAs are typically plagued, we assume in all time-explicit models discrete time intervals. Only the time implicit models in Chapters 10 and 13 require a continuous time interpretation.

Time Structure

There is a particular class of games called *repeated games*. In repeated games the 'basic game', which is called the *constituent game* or *stage game*, is played over several rounds (Taylor 1987, pp. 60ff.; Sabourian 1989, p. 64). When the same constituent game is played an infinite number of times, the whole game is called a *supergame*.[10]

In contrast, if the *game structure* changes over time, that is, a payoff at a point in time depends on the payoffs received and actions taken in previous rounds, this is called a *differential game* (Rasmusen 1989, p. 81). This

implies that repeated games assume structural time *independence*, whereas differential games exhibit a structural time *dependence* (Friedman 1986, p. 72). However, this does *not* imply *strategic time independence* of repeated games. Though the same game might be repeated infinitely, strategies can be based on past actions. For instance, a player may threaten a fellow player to punish him/her in future rounds if s/he does not comply with some agreed strategy.

Generally, the analysis of supergames is much simpler than that of differential games. In fact, differential games have to be analyzed with the methods of dynamic programming. Since these methods are quite involved and warrant an extensive treatment in their own right, differential games are not covered in this book.[11]

In Chapters 12 and 14 we model a supergame in the context of an agreement to reduce greenhouse gases. This framework may raise some objections because greenhouse gases are *stock* pollutants and not *flow* pollutants, that is, they accumulate in the atmosphere. Depending on the amount of greenhouse gases released each year and the rate of decay in the atmosphere, the environmental situation changes over time. Moreover, economic parameters and environmental preferences might alter over time. All together, this would suggest that a differential game approach is more suitable than a supergame framework. However, because the impact of most pollutants with respect to their time dimension is rather uncertain, political agents may well approximate future payoffs using payoffs received today, at least if they are risk neutral. For such a supergame the check on whether a government complies with the terms of an IEA is simplified since the game looks the same in each period.[12] The check is independent of the point in time because if it pays for a country to defect, say, at time $t=4$, then it will also pay at time $t=0$. All we have to do is to discount the complete payoff stream of both alternatives (compliance and non-compliance) to time $t=0$. By following this approach the number of potential equilibrium strategies to be considered in a game is reduced and the task of selecting an equilibrium strategy becomes manageable.

To summarize, under the assumption that the situation is the same at each point in time t, $t=0, 1, ..., \infty$, one can determine whether a country will comply with the terms of an IEA at $t=0$, though in reality the situation might well change at a later stage of the game (but this is not relevant to the consideration at $t=0$).

Information Requirement

Games in which all information is known to all players are called games of *complete information*. If one piece of information is not available to at least

one player, then the game is classified as a game of *incomplete information* (Selten 1982). Usually, it is assumed that the information allocation in a game is *common knowledge* (Brandenburger and Dekel 1989; Fudenberg and Tirole 1996, pp. 541ff.; Myerson 1991, pp. 63ff.). That is, every player knows what all other players know and which information is not available to them.

In incomplete information games it is typically assumed that players know how their fellow players form expectations about unknowns and how they process the information which gradually becomes available to them in the course of the game. This is necessary to determine the strategies of all players – a prerequisite to solving the equilibrium of a game (Kreps 1989, 1990). If the conjectures of players about unknowns are not evident from the underlying problem itself, they have to be introduced explicitly into the game. Obviously, this implies some *ad hoc* flavor. This is one reason – apart from the conceptual difficulties of modeling incomplete information – why most papers analyzing international pollution problems choose a complete information framework.[13,14] We shall stay in this tradition and treat incomplete information only implicitly in Chapter 11. That is, it is assumed that players just do not use certain pieces of information; but the act of information gathering and processing is not itself modeled.

Sequence of Moves

If there is no obvious sequence suggested by the problem itself, then a *simultaneous move game* should be assumed. For instance, if two governments sign an agreement to invest in a cleaner technology, and we have no further information about the unraveling of the situation, the investment decision should be modeled as a simultaneous move game. In contrast, the formation of an environmental agreement may be modeled as a *sequential move game*. After a country has decided whether to become a signatory to an agreement, and signatories have agreed on a joint abatement target, non-signatories may choose their non-cooperative emission levels in return. Such a model is described in Chapter 13.

In most games the sequence of moves affects the strategies in a game and consequently also influences the equilibrium. Often it is advantageous for a player to move first (see, for example, Sections 4.2 and 10.2).

Remark

From what has been said above it is clear that each criterion of Table 2.1 reflects a polar pair of assumptions characterizing a game. Most generally, assumption (a) within each category simplifies the analysis, whereas

assumption (b) makes a game richer and more interesting, but also more complicated. As a general rule, it seems sensible to follow the motto of 'no fat modeling' – a phrase coined by Rasmusen (1989, pp. 14ff.). This implies that simple models should be chosen to bring out the gist of the analyzed problem. Models which are too complicated may distract attention from the main focus of an analysis. For instance, if the role of time is the main focus of an analysis, then it might be helpful to consider only two players in a first step. However, if we are interested in the coalition formation of countries, then we may start with a static or simple two-stage game model and gradually extend the analysis to longer-lasting games.

Of course, there is also the danger of using too simple models which may not capture important features driving a result in reality. For instance, if the decision on whether to switch off and dismantle a nuclear power plant is modeled as a static game with discrete strategy space, this may seem an appropriate simplification. First of all, due to the high sunk costs of the decision, a possible revision of a taken action can be ruled out and this is captured by a static game. Second, a discrete strategy space also seems a good approximation because the choice will be either to dismantle the power plant or to operate it at a low, middle or high capacity. In contrast, if countries negotiate an agreement on greenhouse gases this may be better modeled as a dynamic and continuous strategy space game. Emission reductions may range from 0 to 100 percent and the treaty might be in force for a couple of years.

Unfortunately, sometimes assumptions cannot reflect what is the most appropriate in the particular case, but have to be guided by the necessity that dropping too many simplifying assumptions makes it impossible to solve a game analytically, and one has to rely on simulations. Therefore, in most parts of the book we follow a stepwise approach, altering only one assumption at a time. However, though we have already ruled out some assumptions with respect to the criteria displayed in Table 2.1, lack of space forces us to confine the subsequent analysis further, considering only some interesting combinations of assumptions in this book. The general structure of the book is briefly laid out below.

2.4 OUTLINE OF THE BOOK

In Chapter 3 simple two-player matrix games are considered in a static environment. The Nash equilibrium and the equilibrium in dominant strategies are introduced. Pure strategies and mixed (uncorrelated and correlated) strategies will be distinguished. In Chapter 4 the framework is extended to a finite dynamic setting and in Chapter 5 to an infinite dynamic setting. The

concept of a subgame-perfect equilibrium will be laid out for both settings. Chapter 4 will also give a formal description of strategies in a dynamic framework. In Chapters 6 and 7 refinements of the subgame-perfect equilibrium concept will be discussed in finitely repeated (Chapter 6) and infinitely repeated (Chapter 7) games. These refinements comprise various forms of renegotiation-proof equilibrium (and derivatives of it) and a strongly perfect Nash equilibrium.

Based on the central results of the previous chapters, Chapter 8 looks at issue linkage games in a finite and infinite time horizon.

Whereas Chapters 3–8 illustrate the results with the help of simple matrix games (discrete strategy space), Chapters 9–15 extend the analysis to a continuous strategy space by assuming a global emission game. Chapter 9 introduces the basic model and derives important benchmarks which are used in the subsequent analysis. Again, we begin with a static framework. Chapter 10 extends the framework to two-stage games and to static representations of dynamic games. In this chapter the approaches of public goods economics will be scrutinized from a game theoretical point of view with respect to their logical consistency.

Chapter 11 describes a simple bargaining model in which countries negotiate on the level of a uniform emission quota and, alternatively, on the level of a uniform effluent charge. The bargaining equilibria are compared with each other and with the Nash equilibrium and the social optimum. This model and its central results form the basis of Chapters 12 and 14.

Chapter 12 extends the global emission game to a supergame framework and applies the equilibrium concepts of Chapters 4–7. Particular focus is given to restricted and non-simple punishment profiles and their effect on the stability of a treaty.

Whereas Chapters 9–12 either restrict the analysis to two countries or at least do not consider coalition strategies, Chapters 13–15 deal with strategies in an N-country world. Utilizing the research progress in this field, Chapters 13 and 14 look at coalition models which assume that there is one group of signatories while all other countries play as singletons. Chapter 13 starts by discussing coalition models which may be classified as static representations of dynamic games. Chapter 14 analyzes the coalition formation process in a supergame framework. In particular the bargaining equilibria derived in Chapter 11 are considered with respect to their stability in an N-country world. The model endogenously explains the choice of the policy instrument employed in an IEA, the abatement target and the size of the signatories' coalition. Finally, Chapter 15 presents recent developments in the literature on coalition formation. These concepts allow for the coexistence of several coalitions. Since these new concepts have hardly been applied to the problem of international pollution control yet, the

emphasis in this chapter lies on an evaluation concerning the question whether these concepts can be usefully employed in future research on this topic.

NOTES

1. In order to characterize an equilibrium, it is necessary to list the equilibrium strategies of all players. Thus, the notion of an equilibrium in game theory is slightly different from that in economics. For instance, whereas in economics the price resulting from the competition between firms in a Cournot oligopoly is called an 'equilibrium price', this is an outcome in game theory. From a game theoretical perspective, the strategies, which are the chosen quantities by each firm in the simple static version of this game, would also have to be listed for a full characterization of the equilibrium.
2. For details, see Section 2.3.
3. Generally, the set of players may also be infinite, see, for example, Fudenberg and Tirole (1996) for examples. However, in this book no such instances will be encountered.
4. The terms *action space* and *action set* can be used synonymously. However, the former term is used mostly in games with an infinite number of actions.
5. Many *bargaining solutions* belong to cooperative game theory, as for instance the Nash bargaining solution or the Shapley value. In Chapter 11 we argue that those solutions are based on some assumptions which are often violated in reality and therefore are not treated in this book.
6. For a discussion of the problems associated with the aggregation of welfare functions, see for instance Boadway and Bruce (1993); Just *et al.* (1982); and Sen (1984).
7. See, for instance, the Rio Declaration of 1992 and its successor protocols by which the Global Environmental Facility (GEF) was set up. Industrial countries are supposed to contribute to the GEF from which developing countries can receive financial support for environmental projects (see, for example, Bergesen and Parmann 1997, pp. 90ff.; Hanley *et al.* 1997, pp. 171ff.; Jordan and Werksman 1996, pp. 247ff.; Kummer 1994, p. 260; and Sand 1994, pp. 98ff.). Contributions by industrial countries to a fund are also required by the Convention on Biological Diversity (Beyerlin 1996, p. 617; and Gündling 1996, pp. 806ff.) and the Montreal Protocol on the Protection of the Ozone Layer (DeSombre and Kauffman 1996, pp. 89ff.; Ladenburger 1996, pp. 72ff.; and Sand 1996, p. 56). A model analyzing the stability of such transfers is provided by Barrett (1994a).
8. For instance, statistical reports on historical emissions are issued by the Norwegian Meteorological Institute in Oslo which provides emission data for the monitoring program EMEP (Emission Monitoring and Evaluation Program) which is part of the Convention on the Reduction of Long-range Transboundary Emissions in Europe (Geneva 1979). In these reports data on SO_2, NO_x, and VOCs (volatile organic compounds) have been reported since 1980.

   Note that the fact of biased reporting by governments, which frequently occurs in reality, is an issue of incomplete information but not of bounded recall.
9. The assumption of bounded recall is sometimes introduced into differential games (the term is explained below) in order to simplify the analysis. A typical open loop strategy assumed in this literature is a Markov strategy, where only the immediate previous history is considered. See, for example, Dutta and Sundaram (1993); Hoel (1992b); Mäler (1991, 1992); and Wirl (1994) in the environmental economics context.
10. Most game theorists follow this definition, implying that the *finite* repetition of a game is *not* called a supergame. This is also how we shall use this term. For a different interpretation, see, for instance, Friedman (1986, pp. 94ff.). He also calls finitely repeated games supergames.
11. General references for dynamic optimization problems are Kamien and Schwartz (1991) and Seierstad and Sydsater (1987). Differential games are treated in Basar and Olsder

(1982); Clemhout and Wan (1994); and Friedman (1994). In the environmental and resource economic context they have been applied by Dockner and van Long (1993); Hoel (1992b); Martin *et al.* (1993); and Tahvonen (1994).

12. This is not true for finite games, as we shall show in Chapters 4 and 6.
13. Exceptions include Chillemi (1996); Laffont (1993); and Steiner (1997).
14. A particular form of incomplete information is *uncertainty of information* (see, for example, Rasmusen 1989, pp. 52ff.). Uncertainty refers to the particular feature whereby after a player has moved, 'nature' makes a move which is either not completely observable by players and/or cannot be influenced by them (for example, Holler and Illing 1993, pp. 36ff.). An example is the global warming game where there is great uncertainty as to how the climate will change in the future. To solve for such a game, assumptions with respect to the risk attitude of agents (risk averse, risk neutral or risk loving) have to be made. Though most books on game theory cover incomplete information in general, they do not treat uncertainty of information since this warrants an analytical approach in its own right. Exceptions include Chichilnisky *et al.* (1998) and Machina (1989). In the context of international pollution control, see, for example, Endres and Ohl (1998a, b); Mohr and Thomas (1998); and Na and Shin (1998).

3. Static games with discrete strategy space

3.1 INTRODUCTION

The aim of this chapter is threefold:

1. To analyze the effect of the cost–benefit structure on the outcome of a game. Here we shall deal with the prisoners' dilemma (Section 3.2), the chicken game (Section 3.3), the assurance game and the no-conflict game (Section 3.4) in the two-country context. An extension to cover the general case of N countries is provided in Section 3.5.
2. To introduce some basic game theoretical concepts such as an equilibrium in dominant strategies, a Nash equilibrium in pure and uncorrelated and correlated mixed strategies.
3. To demonstrate that by playing uncorrelated or correlated mixed strategies the payoff space in a game can be convexified (Section 3.6). This is some preparatory work needed for dynamic games in subsequent chapters. An application of correlated strategies is provided in Section 3.7.

In this chapter we focus exclusively on simple static games with a discrete strategy space. The examples assume that governments can choose between two policy options; however, an extension to cover the case of larger action sets is straightforward. All games are non-cooperative and non-constant sum games. Trivially, by the definition of static games the sequence of moves is simultaneous and the time dimension and time structure are irrelevant. With respect to Table 2.1, the games in this chapter can be categorized as: 1b, 2b, 3a (b), 4a, 5a, 8a, 9a. Some aspects discussed in this and the two subsequent chapters can also be found in the political science literature, for example, Aronson (1993); Hamburger (1973); McLean (1981); Oye (1986); Raiffa (1982); Snyder (1971); Taylor (1987); Taylor and Ward (1982); Snidal (1985, 1988); Stein (1982); van der Lecq (1996); and Ward (1987, 1993). Most of this literature gives a non-technical introduction to the problem of cooperation in international policy coordination.

3.2 PRISONERS' DILEMMA

The prisoners' dilemma game, henceforth abbreviated to PD game, is the most frequently cited game in the literature to explain in a simple way the difficulties of reaching a stable IEA. Suppose two countries suffer from transboundary emissions stemming from energy production. If both countries switch from the old (for example, coal burning) to the new (for example, hydro power) technology, net benefits in each country would be higher than in the present situation. However, though such a joint energy policy is attractive to both countries, it is even more beneficial to a country not to invest in the new technology if the neighboring country does. The free-rider enjoys a cleaner environment (though not of the same quality as if both countries switched to the new technology), but does not have to carry any investment costs. For the country which unilaterally conducts the investment this is the worst case: the costs exceed the benefits of the investment. An example reflecting such a payoff constellation is provided in Matrix 3.1. Here a_i denotes the action 'invest' (or 'abate') and na_i stands for 'not invest' (or 'not abate'). The upper entry in each cell is the payoff to country 1 (row player) and the lower entry the payoff to country 2 (column player). For instance, if both countries decide not to invest in the new technology and the status quo remains, each country receives net benefits of 2 units.[1]

Matrix 3.1 Prisoners' dilemma

	a_2	na_2
a_1	3.2 3.2	1.4 4.4
na_1	4.4 1.4	**2** **2**

A payoff matrix is a convenient device to summarize all relevant information in this game. It is one possibility of the *normal form* representation of this game, where this term is defined as follows (Gibbons 1992, p. 4):

Definition 3.1: Normal form of a game
The normal form representation of a game (also called the strategic form of a game) specifies: (a) the players in the game, (b) the strategy combi-

nations, and (c) the payoffs received by each player for each possible strategy combination in this game. For short, the normal form of a game is given by $\Gamma = (I, S, \Pi)$.

Note that, due to the assumption of a one-shot game, actions and strategies coincide in this game. It can easily be checked that in the PD game each country has the *dominant strategy* 'no abatement'; that is, regardless what the neighboring country does, strategy na_i delivers the highest payoff. If country j plays a_j, na_i delivers a payoff of 4.4 whereas a_i leads to a payoff of only 3.2 to country i. If country j plays na_j, country i nets 2 by playing na_i instead of 1.4 when playing a_i. Hence, the unique equilibrium in this game is $S^{D*} = \{(na_1, na_2)\}$ where the superscript D stands for 'dominant strategy equilibrium'. The associated payoffs are printed in bold in Matrix 3.1. The equilibrium of this game implies that cooperation fails and both countries get stuck in the Pareto-inferior status quo. Hence, delegates of both countries either do not sign an agreement in the first place because they anticipate the instability of such a deal, or they sign an agreement but neither country will comply.[2]

The predictive power of this kind of equilibrium may be regarded as quite high. Each player has one strategy which leads under every contingency to a higher payoff than any other strategy. That is, a player can determine his/her 'best strategy' without having to rely on any speculation about the behavior of his/her opponents. Hence, this equilibrium is also immune to any kind of strategic considerations of players and can therefore be regarded as very robust.

Definition 3.2: Equilibrium in dominant strategies[3]
An equilibrium in dominant strategies is a strategy combination $s^* = (s_i^*, s_{-i}^*)$ for which $\pi_i(s_i^*, s_{-i}) \geq \pi_i(s_i, s_{-i}) \forall s_i \in S_i, s_{-i} \in S_{-i}, s_i \neq s_i^*$ and $i \in I$ holds, where s_{-i} denotes all strategies except that of player i, and s_i some arbitrary strategy of player i. That is, in equilibrium every strategy is a best response irrespective of the strategies of other players.

Note that for the incentive structure in the PD game (and in all the games discussed below) it is neither necessary for utility to be measured *cardinally* (for example, a payoff of 3.2 gives 3.2/2 times more utility compared to the status quo) nor that utility can be compared across countries (for example, a payoff of 4.4 generates higher utility to country 1 than a payoff of 3.2 to country 2). Also the assumption of symmetric payoffs is not necessary. With reference to Matrix 3.2, all that is required for a game to qualify as a (basic) PD game is $c_i > a_i > d_i > b_i \; \forall \; i \in I$.

However, to make the following analysis interesting from a policy point

of view, we assume that (1) utility can be aggregated, and that (2) $a_1 + a_2 > b_1 + c_2$ and $a_1 + a_2 > c_1 + b_2$ hold. The first assumption automatically implies that $a_1 + a_2 > d_1 + d_2$ (since $a_i > d_i \forall\, i \in \mathrm{I}$) and hence together with the second assumption 'mutual cooperation' is globally optimal.

Matrix 3.2 General payoff matrix

		a_2	na_2
a_1	a_1	b_1	
	a_2	c_2	
na_1	c_1	d_1	
	b_2	d_2	

To set the stage for a comparison of the outcome in the static PD game with outcomes in other models and settings, we briefly review three main assumptions which are responsible for the pessimistic result obtained above.[4]

First, the *discrete strategy space* restricts the question of whether to cooperate on environmental protection to a binary choice. In contrast, in a continuous strategy space where a country can tune its decision more finely, more optimistic results can be obtained (see Chapter 9). However, if the policy options discussed on an international platform are clearly distinct from each other, a discrete strategy space is the appropriate setting.

Second, in a *static setting* no threats and rewards can be used to prevent a country from seeking its immediate interests. Below, we shall show that cooperation might be possible if a game lasts longer. However, as pointed out in Section 2.3, a static setting is appropriate if an action is associated with high sunk cost and is not reviewed at short intervals.

Third, the *cost–benefit structure* generates this Pareto-inferior result. Apart from the free-rider incentive it does not pay a country to contribute unilaterally to the international public good 'environmental quality'. In fact, a unilateral contribution is the worst outcome from the contributor's perspective.

Looking at the record of international environmental protection it is evident that there has been successful cooperation in a few areas, though for most international environmental problems either less has been

achieved than would be advisable from a global point of view or no actions at all have been taken.[5] Therefore to be able to capture the broad spectrum of IEAs, some of the previous assumptions have to be modified and/or the model has to be extended. In this chapter we start by modifying the assumptions of the cost–benefit structure.

3.3 CHICKEN GAMES

3.3.1 Pure Strategies

In chicken games the prospects for a cleaner environment are higher (see, for example, Holler and Illing 1996, pp. 89ff.; Rasmusen 1995, pp. 72ff.). Still, the free-rider incentive exists. However, the cost–benefit structure is such that it pays a country to invest unilaterally, though it prefers its neighbor also to contribute to a joint environmental policy. An example is shown in Matrix 3.3.

Matrix 3.3 Chicken game

		a_2	na_2
a_1		4.6 4.6	**2.2** **5.2**
na_1		**5.2** **2.2**	2 2

From a casual inspection of Matrix 3.3 it is apparent that country 1 prefers to invest, a_1, if country 2 does not invest, na_2. However, if country 2 invests, a_2, country 1 prefers 'no investment', na_1. Obviously, country 1 has no dominant strategy and, by symmetry, this also applies to country 2. Now, a best reply depends on the strategy of the fellow player. The notion of a *Nash equilibrium*, henceforth abbreviated NE, takes this into consideration (Nash 1950a; see also Moulin 1986, p. 104).

Definition 3.3: Nash equilibrium
A Nash equilibrium is a strategy combination $s^* = (s_i^*, s_{-i}^*)$ for which $\pi_i(s_i^*, s_{-i}^*) \geq \pi_i(s_i, s_{-i}^*) \forall s_i \neq s_i^*, s_{-i} \neq s_{-i}^*$ and $i \in \mathrm{I}$ hold and where $s_i^*, s_i \in \mathrm{S}_i$,

and $s^*_{-i} \in S_{-i}$. That is, in equilibrium every strategy is a best response to the best strategies of the other players.

In the chicken game the strategy combinations (a_1, na_2) and (na_1, a_2) constitute such mutual best responses to each other. That is, each player has no incentive to deviate from his/her strategy, given that the other player does not deviate. Hence, $S^{N*} = \{(a_1, na_2), (na_1, a_2)\}$. The payoffs in the two NE are printed in bold in Matrix 3.3.[6]

There are basically two interpretations of a Nash equilibrium:

1. In equilibrium each player's conjecture about the other players' behavior is confirmed. Hence, the NE is supported by *self-consistent beliefs* and might therefore be viewed as a *rational expectation outcome* (Eichberger 1993, pp. 107ff.).
2. Before play, some communication takes place in which a third party or a player proposes to play an NE. The proposal will be accepted if it is self-enforcing. This is indeed the case: no player likes to change strategy, *given* the other players play their equilibrium strategies.

Of course, in the chicken game each country favors that equilibrium where it can take a free-ride and the other country conducts the investment.[7] Thus, without any further information it is not clear which of the two equilibria will be played. Though this indeterminacy is disturbing in predicting the outcome of this game, if we consider the complexity of real world situations it is hardly surprising.

The payoff relations necessary to generate a chicken incentive structure are $c_i > a_i > b_i > d_i \; \forall \; i \in I$. Assuming that mutual cooperation (investment by both countries) is globally optimal, then $a_1 + a_2 > b_1 + c_2$ and $a_1 + a_2 > c_1 + b_2$ must hold additionally. These two inequalities do not follow from the basic chicken game but are often and henceforth assumed. For the incentive structure (and therefore for the equilibrium) of the game itself, however, they are not relevant.

3.3.2 Mixed Strategies

Apart from the indeterminacy of the solution in the symmetric chicken game, there is another reason why we may feel unhappy with the two NE (in pure strategies). Though the payoff structure is symmetric in Matrix 3.3, either of the two NE generates very asymmetric payoffs. In contrast, by allowing for *mixed strategies* to be played in the chicken game a more even distribution of payoffs can be generated. Whereas playing a *pure strategy* implies that a player chooses one particular strategy out of his/her strategy

set, mixing involves playing several strategies with some probability (Rasmusen 1995, pp. 67ff.).

With respect to the chicken game in Matrix 3.3, a mixed strategy implies that a country plays a_i with probability p_i and na_i with probability $(1-p_i)$. The *expected payoff* to country i from playing a_i is $4.6 \cdot p_j + 2.2 \cdot (1-p_j)$. Alternatively, if country i chooses na_i its expected payoff is $5.2 \cdot p_j + 2 \cdot (1-p_j)$. Country i will invest if its expected payoff is higher than that from not investing:

$$4.6 \cdot p_j + 2.2 \cdot (1-p_j) \geq 5.2 \cdot p_j + 2 \cdot (1-p_j) \Leftrightarrow 1/4 \geq p_j. \tag{3.1}$$

That is, country i invests if the probability that country j invests is less than 1/4. It does not invest if this probability is greater than 1/4, and it is indifferent between both strategies if this probability is exactly 1/4. Hence, if both countries invest with probability 1/4, both countries have no incentive to deviate from their strategy and beliefs are mutually confirmed. Thus $p^* = (p_1^* = 1/4,\ p_2^* = 1/4)$ is the *Nash equilibrium in mixed strategies*. The resulting payoff to each country is 2.8 which is more in line with the symmetric structure of the chicken game in Matrix 3.3 than those payoffs derived from the pure Nash equilibria.[8]

Let us now define more formally what has been derived above:

Definition 3.4: Mixed strategy
Suppose a normal form game $\Gamma = (I, S, \Pi)$ in which player i's strategy set, $S_i = \{s_{i1}, ..., s_{iR_i}\}$, consist of R_i pure strategies s_{iRi}, $R = \{1, ..., R_i\}$. Then, a mixed strategy of player i is a probability distribution $p_i = (p_{i1}, ..., p_{iR_i})$ on S where $0 \leq p_{ir} \leq 1$ and $\sum_{r=1}^{R_i} p_{ir} = 1$.

Definition 3.5: Nash equilibrium in mixed strategies
A Nash equilibrium in mixed strategies is a probability distribution $p^{N*} = (p_1^*, ..., p_N^*)$ for which $\pi_i(p_i^*, p_{-i}^*) \geq \pi_i(p_i, p_{-i}^*) \forall p_i \neq p_i^*$ and $i \in I$ holds. That is, each player's mixed strategy is a best response to the other players' mixed strategies.

Note that now in the general case p_i denotes some probability distribution of player i and not a probability that a particular strategy s_{ir} is played. From the definition it is obvious that an NE in mixed strategy is a straightforward extension of the definition of an NE in pure strategies. In fact, a pure strategy is a special case of a mixed strategy which is played with probability 1. Therefore, pure strategies are a subset of mixed strategies (Eichberger 1993, pp. 20ff.).

Before dealing with some other games and other technical features of mixed strategies in the subsequent sections, we shall pause here and briefly discuss the pros and cons of mixed strategies.

Cons

1. The derivation of a mixed strategy equilibrium requires that utility is measured cardinally. Recall, this assumption is not necessary when deriving pure strategy equilibria.
2. For most economic problems the motivation of mixed strategies in the context of one-shot games is obscure (see, for example, Rasmusen 1989, pp. 72ff.). Why should players randomize between strategies? This question becomes particularly momentous when recalling that one argument in favor of a static setting was that the investment is associated with high sunk costs and not reviewed regularly. Thus, one should expect a government to take a clear-cut decision and not to randomize.
3. A mixed strategy equilibrium is very *sensitive* to a *change of payoffs*. Suppose the free-rider payoff in Matrix 3.3 is changed from 5.2 to 5.6. Then the new equilibrium is $p^* = (p_1^* = 1/6, p_2^* = 1/6)$ instead of $p^* = (p_1^* = 1/4, p_2^* = 1/4)$. In contrast, such a small perturbation of payoffs does not upset the pure strategy equilibrium. For most economic problems it seems sensible to expect that small alterations in payoffs, such as those caused by small variations in economic fundamentals like prices and so on, will not change the behavior of players dramatically, and therefore such variations should be buffered to some extent.
4. A mixed strategy equilibrium is *not* very *sensitive* to a *change of strategies*. If player i plays the equilibrium strategy p_i^*, then player j can choose any probability $0 \leq p_{jr} \leq 1$ and player i receives the same payoff (see, for example, Holler and Illing 1993, pp. 70ff.). Clearly, by the nature of an equilibrium in mixed strategies players are indifferent between their pure strategies. However, it may be asked why should player j play the equilibrium strategy p_j^* and not any other strategy $p_j \neq p_j^*$ if this does not affect his/her payoff? This is a weak point from a game theoretical point of view because strategies should be important for an equilibrium.

Pros

1. The equilibrium probabilities may be interpreted as the likelihood of an outcome. For the example in Matrix 3.3 we compute the probability that mutual investment takes place to be $p_1^* \cdot p_2^* = 1/16$, the likelihood

of unilateral investment $p_1^* \cdot (1 - p_2^*) = (1 - p_1^*) \cdot p_2^* = 3/16$, and that of no investment to be $(1 - p_1^*) \cdot (1 - p_2^*) = 9/16$. In a wider context of more than two countries these figures could be interpreted as 50 percent of the countries invest. This kind of information may judged more useful than knowing that any of the two pure strategy NE could be played.

2. Though the next remark should be reserved for the chapters on dynamic games, its intuition is obvious. If a game is played over several rounds, then the probabilities discussed under point 1 of the cons may be interpreted as the frequency with which some strategies are played on average.
3. There are also examples in which randomizing is more convincing than playing pure strategies. Think of an international body controlling emissions within some IEA. If control costs cannot be neglected, then inspectors may decide randomly to audit the emission record of countries. The random decision may concern which country to audit, the conciseness and the frequency of audits. In this case, mixing between strategies seems plausible.[9]

In the following we shall always assume pure strategies as long as no explicit reference is made to mixed strategies. Some technical aspects of mixed strategies will be reconsidered in Sections 3.6 and 3.7. However, since the main focus of this chapter is on the cost–benefit structure of games, we first discuss assurance and no-conflict games.

3.4 ASSURANCE AND NO-CONFLICT GAMES

In an assurance and no-conflict game the degree of cooperation is higher than in the previous games. Whereas in the PD and in the chicken game, no or only one country invests in the new technology (if only pure strategies are considered), in the games we discuss in this section both countries (most likely) invest. Typical examples of both games are given in Matrices 3.4 and 3.5. Again, symmetry is assumed for convenience but is not essential for the basic incentive structure.

3.4.1 Assurance Games

In *assurance games* the payoff structure is such that it pays neither country to invest unilaterally. A country prefers either the status quo or joint investment. The assurance payoff structure could be generated by economies of scale in the development and production of new and cleaner power plants. If the costs of R&D are high, an investment in the new technology might

Matrix 3.4 Assurance game

	a_2	na_2
a_1	**9.5** **9.5**	0 1
na_1	1 0	*2* *2*

Matrix 3.5 No-conflict game

	a_2	na_2
a_1	**9.5** **9.5**	5 8
na_1	8 5	2 2

only pay if both countries cooperate on this issue. No investment is preferred by both countries since unilateral investment is so costly.

Another example generating such a payoff structure is the invention of catalytic converters in the European Union several years ago (Heal 1994). Suppose that only some governments had required their industry to meet the stricter EU regulations. Then, on the one hand, the car industry in the 'environmentally concerned' countries would have faced the dilemma of either higher production costs, caused by production for two separate markets, or losing market shares if they only produced the more costly 'cat cars'. Moreover, unleaded petrol had to be provided all over Europe in order not to restrict travel. Without such a common market, travel would have become inconvenient and costly.

On the other hand, in countries with less environmentally concerned governments, the car industry faced the same kind of a dilemma – either producing for two markets or losing market share. Moreover, if a country opted not to produce cat cars it would risk falling behind in acquiring technical know-how. This can turn out to be a disadvantage in the future if, due to higher environmental awareness among its citizens, tougher regulations become binding in an increasing number of countries.

Taken together, neither to go ahead nor to free-ride would pay. Only a joint coordinated environmental policy could improve upon the status quo.[10]

The NE strategy pairs in the assurance game are $s^{N*(1)} = (a_1, a_2)$ and $s^{N*(2)} = (na_1, na_2)$. In Matrix 3.4 the payoffs of the Pareto-superior equilibrium are printed in bold and the payoffs of the Pareto-inferior equilibrium are printed in italic. Due to the Pareto-superiority of $s^{N*(1)}$, this equilibrium is the *focal point* in this game and we expect both countries to invest.[11, 12, 13] All that is required is some basic form of communication in order to coordinate the strategies of the players.[14]

With respect to the general payoff matrix (Matrix 3.2), the inequalities $a_i > d_i$, $d_i > b_i$ and $d_i > c_i \; \forall \; i \in \mathrm{I}$ define the basic payoff structure of an assurance game. Though $b_i < c_i$ has been assumed in Matrix 3.4, $b_i \geq c_i$ also preserves the incentive structure of an assurance game. Obviously, mutual investment is globally optimal if utility can be aggregated.

3.4.2 No-conflict Games

No-conflict games have only one equilibrium, which is an equilibrium in dominant strategies, that is, $\mathrm{S}^{\mathrm{D}*} = \{(a_1, a_2)\}$. Generally, in a no-conflict game $a_i > b_i$, $c_i > d_i \; \forall \; i \in \mathrm{I}$ holds. The relation of b_i and c_i remains unspecified, but $b_i < c_i$ seems plausible if we follow the assumption that free-riding generates a higher payoff than unilateral investment. In this game environmental benefits from the investment are perceived to be so high that unilateral and joint investment are rational. Hence, cooperation on joint environmental policy should not prove difficult.

3.5 AN EXTENSION TO *N* COUNTRIES

In this section we briefly discuss the extension of the previous models from two countries to the general case of *N* countries. Analytically, this extension is straightforward. Conceptually, this is a little more difficult in the case of a chicken and an assurance game. A simple way to display the incentive structure in the symmetric *N*-country case is suggested by Matrix 3.6.

Matrix 3.6 N-*country environmental game*

	0	1	...	$N^* - 1$	$N - 1$
a_i	$\mathrm{MB}_i^1 - \mathrm{MC}_i$	$\mathrm{MB}_i^1 + \mathrm{MB}_i^2 - \mathrm{MC}_i$	...	$\sum_j^{N^*-1} \mathrm{MB}_i^{j+1} - \mathrm{MC}_i$	$\sum_j^{N-1} \mathrm{MB}_i^{j+1} - \mathrm{MC}_i$
na_i	0	MB_i^1	...	$\sum_j^{N^*-1} \mathrm{MB}_i^1$	$\sum_j^{N-1} \mathrm{MB}_i^j$

MB_i^j stands for marginal benefits and MC_i for marginal costs which accrue to country *i* from the investment in the new technology and the status quo payoff has been normalized to zero. The term 'marginal' is used to stress that these benefits and costs occur additionally compared to the status quo. For instance, MB_i^4 is the marginal benefit country *i* receives if

four instead of three countries contribute to abatement. The number of cooperating countries is denoted by N^*.

In Matrix 3.6 country i is assumed to be the row player, all other countries to be the column player. The first column represents the case if none of the $N-1$ countries cooperates. The second column depicts the case if one of the $N-1$ other countries cooperates and so on. The payoffs in each cell are those of country i. The incentive to free-ride is simply computed by subtracting the first row from the second in each column.

3.5.1 PD Games

An important feature of the incentive structure of PD games is that it never pays a country to invest unilaterally and that defection from mutual cooperation pays. If this incentive structure is to be preserved in the N-country case, then $MB_i^{j+1} < MC_i$ or all $i \in I$ and all $j = \{0, ..., N-1\}$ must hold. If we assumed that marginal benefits decrease in the number of contributors N^* (as economic wisdom would suggest), that is, $\partial MB_i^j / \partial N^* < 0$, then the free-rider incentive would be particularly strong in columns to the far right.

Moreover, a basic prerequisite that cooperation is attractive at all is $\Sigma MB_i^{j+1} > MC_i$ if country i cooperates too, and $\Sigma MB_i^j > 0 \; \forall \; i \in I$ and $1 \leq j \leq N-1$ if country i does not cooperate.

3.5.2 No-conflict Games

The incentive structure of no-conflict games implies that the only NE, which is at the same time an equilibrium in dominant strategies, has all countries investing. Hence, we have to require $MB_i^{j+1} > MC_i \; \forall \; i$ and j and $MB_i^j > 0 \; \forall \; i \in I$ and $1 \leq j \leq N-1$. These conditions also ensure that cooperation is attractive and full cooperation globally optimal.

3.5.3 Assurance Games

In its 'pure' version the incentive structure of assurance games implies that either all countries or no country invest. Moreover, joint investment is preferred among these two options. For such an incentive structure $\Sigma MB_i^{j+1} < MC_i$ for all $i \in I$ and $j < N-1$, and $\Sigma MB_i^{j+1} > MC_i$ and $MB_i^j > MC_i$ for $j = N-1$ and all $i \in I$ is required. Since in an assurance game cooperation among some countries never pays a country as long as not all countries cooperate – even though it does not contribute anything to joint abatement – $\Sigma MB_i^j < 0$ for all $1 \leq j < N-1$ must hold. That is, only if all countries invest will marginal benefits exceed a threshold so that the investment is beneficial to country i and globally optimal.

In the context of the economies of scale argument we presented in Section 3.4, $\partial MB_i^j/\partial N^* > 0$ seems plausible,[15] however, it is not necessary to generate this result (though $MB_i^N > MC_i > 0 > MB_i^{j+1} \forall\ j < N-1$ is necessary).

In a modified version of the assurance game the following scenario could also be constructed. If a certain number of countries invest, say N^*, payoffs to these countries are higher than in the status quo and defection does not pay. However, if the number of countries contributing to abatement falls short of N^*, countries prefer not to invest. This implies that if $\partial MB_i^{j+1}/\partial N^* > 0$ holds, then $\Sigma MB_i^{j+1} < MC_i$ for all $i \in I$ and $j < N^* - 1$, and $\Sigma MB_i^{j+1} > MC_i$ for all $j \geq N^* - 1$ and $i \in I$. Then, of course, it is also an NE if $N^* + 1, N^* + 2, \ldots, N$ countries contribute to abatement. However, a focal point argument should put us in the position to select among those multiple equilibria.

3.5.4 Chicken Games

An important feature of two-country chicken games is that a unilateral contribution to abatement is preferred to no abatement at all but a country prefers not to invest in the new abatement technology if the other country does. In an N-country game this implies that country i does not invest as long as it expects that at least N^* countries will invest, but it invests if it expects the number of contributors to fall short of N^*. Thus, the incentive structure for all i would be $MB_i^{j+1} > MC_i\ \forall\ j \leq N^* - 1$ and $MB_i^{j+1} < MC_i\ \forall\ j > N^* - 1$. $\Sigma MB_i^{j+1} > MC_i$ for all i and j and $MB_i^j > 0$ for all i and $1 \leq j \leq N-1$ ensures that cooperation is always beneficial compared to the status quo.

The assumption $\partial MB_i^j/\partial N^* < 0$ is in line with the requirements (though not necessary) and the number of countries contributing to abatement in equilibrium, N^*, depends on the relation of marginal benefits to marginal costs. As in the two-country chicken game, there will be multiple equilibria. For instance, if $N = 3$ and $N^* = 1$, there are three equilibria in pure strategies, each with one of the three countries investing unilaterally. In the case where $N^* = 2$ each equilibrium has a combination of two countries ({1, 2},{1, 3} or {2, 3}) investing. Departing from the symmetry assumption may reduce the number of multiple equilibria. For instance, assume a mix of a chicken and a PD game. The 'chicken' group of countries might value environmental benefits from the investment higher than the PD group and are therefore the signatories to an IEA.

3.5.5 Discussion

From the discussion it is evident that with the help of the simple models presented in this chapter an environmental agreement ranging from zero to N participating countries can be depicted by varying the relation of marginal benefits to marginal costs. For example, observing that a pollution problem is not regulated within an IEA, one may suspect that the benefit–cost structure is that of a PD game. Many authors argue that for most international problems, in particular global ones, marginal costs exceed marginal benefits on a country basis. This could explain why so far not much success can be detected regarding the problem of global warming. In contrast, if an effective IEA is in operation with some but not all countries participating, a chicken or modified assurance game is a possible incentive pattern underlying the problem. Following the main opinion in the literature, judging the Montreal Protocol as a relatively successful treaty, the ozone problem could be such an instance.

However, though the models in Chapters 13–15 on coalition formation confirm that the cost–benefit structure is a crucial variable for the success of a treaty, for a final evaluation we first have to analyze whether modifications of the assumptions made so far can also explain partial or full cooperation. Since free-riding is a problem, particularly in PD and chicken games, the analysis in Chapters 4–8 is confined to these two types of games.

3.6 CONVEXIFICATION OF PAYOFF SPACE

Since most of the results of dynamic games presented in subsequent chapters are formulated in terms of equilibrium payoffs, we must deal in this section with the convexification of payoff space to provide a general platform for this discussion. This term refers to the fact that by playing mixed strategies any linear combination of pure strategy payoff vectors can be generated in a game. To show that this is possible in *all* games, we have to extend the definition of mixed strategies to comprise *correlated strategies* as well (Rasmusen 1995, pp. 75ff.). An application of correlated strategies, showing how a consultant or an arbitrator can improve upon the outcome in a game, though he/she is *not* equipped with enforcement power, will be provided in Section 3.7.

3.6.1 Uncorrelated Strategies

The convexification of payoff space is illustrated for the chicken game in Matrix 3.3. In Figure 3.1 the two-dimensional convex payoff space of this game has been drawn.

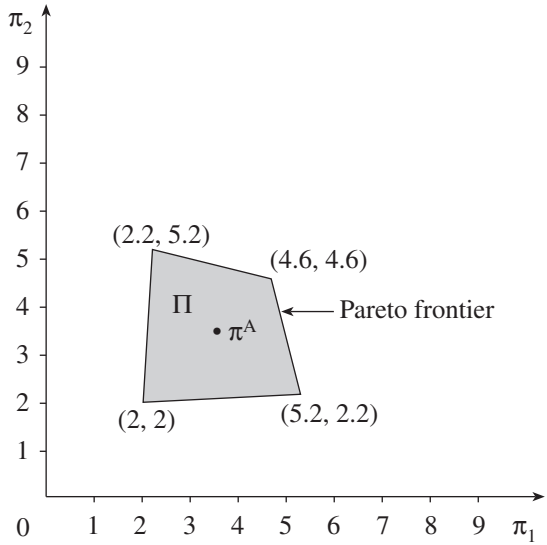


Figure 3.1 Payoff space of the chicken game in Matrix 3.3

All possible payoff combinations of this symmetric one-shot chicken game are represented by the shaded area. A point at each corner of this area is a payoff tuple resulting from a pure strategy combination. In order to generate other payoff tuples, mixed strategies have to be played. By varying the probability p_i and p_j of playing the first strategy (and accordingly $1-p_i$ and $1-p_j$ of playing the second strategy), the discrete strategy spaces of countries i and j become continuous. Thus, mixing strategy a_i and na_i is like increasing the number of strategies from two to an infinite number of strategies and the payoff space, $\Pi = \Pi_1 \times \Pi_2$, becomes a convex set as drawn in Figure 3.1. For example, if $p_1 = p_2 = 0.5$, then the payoff tuple π^A can be generated.

3.6.2 Correlated Strategies

In some games the convexification of payoff space is less straightforward than in the chicken game. In such games it does not suffice to play 'only' mixed strategies, that is, to play *uncorrelated mixed strategies* as assumed so far; instead, in these games *correlated strategies* have to be played to convexify the payoff space. Whereas uncorrelated strategies imply that the probability distributions of players are independently chosen, correlated strategies imply that the probability distributions are related to each other

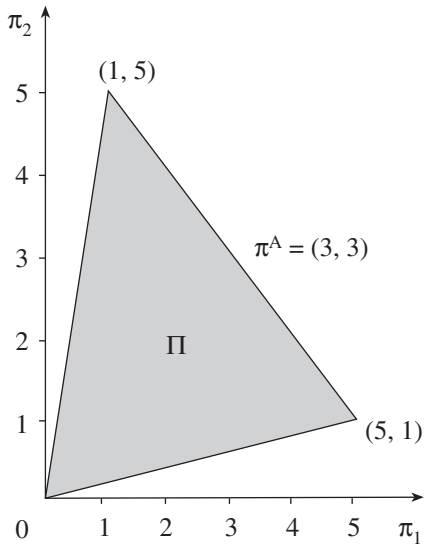


Figure 3.2 Convex payoff space of the matrix game in Matrix 3.7

in some way. Consider the simple example in Matrix 3.7 which is illustrated in Figure 3.2. The example is chosen for illustrative purpose only and is not economically motivated.[16]

This game has two NE in pure strategies, that is, $S^{N^*} = \{(a_{11}, a_{21})\ (a_{12}, a_{22})\}$, where the associated payoff tuples, that is, $\pi^{N^*(1)} = (5, 1)$ and $\pi^{N^*(2)} = (1, 5)$, are printed in bold in Matrix 3.7. There is also NE in mixed strategies, that is, $p^{N^*} = (p_i^* = 1/6, p_j^* = 5/6)$, with associated payoff tuple $\pi^{N^*(3)} = (5/6, 5/6)$.

Suppose we want to find an *uncorrelated* mixed strategy to generate the payoff tuple $\pi^A = (3, 3)$ which is an element of the convex payoff space Π. This implies that the following two equations must be satisfied:

$$5p_1p_2 + (1-p_1)(1-p_2) = 3 \tag{3.2}$$

$$p_1p_2 + 5(1-p_1)(1-p_2) = 3. \tag{3.3}$$

However, solving this inequality system by adding (3.2) to (3.3), dividing through by 6, and rearranging terms, shows:

$$(p_1p_2 - p_1) + (p_1p_2 - p_2) < 0 \tag{3.4}$$

Matrix 3.7

	a_{21}	a_{22}
a_{11}	**5** **1**	0 0
a_{12}	0 0	**1** **5**

Matrix 3.8

	a_{21}	a_{22}
a_{11}	$\frac{1}{2}$	0
a_{12}	0	$\frac{1}{2}$

Matrix 3.9

	a_{21}	a_{22}
a_{11}	z_1	z_2
a_{12}	z_3	z_4

due to $0 \leq p_1 \leq 1$ and $0 \leq p_2 \leq 1$. Hence, the payoff tuple π^A cannot be generated using uncorrelated mixed strategies.

However, suppose that both parties agree to use some random device which picks the probabilities of the various strategy combinations. For instance, both parties agree to flip a coin and play their first strategy if heads shows up and play their second strategy if tails shows up. In other words, both parties agree to correlate their strategies. Because of an equal winning chance, the two players play the strategy combination (a_{11}, a_{21}) and (a_{12}, a_{22}) with a probability of 50 percent. Based on these probabilities of outcomes (see Matrix 3.8), the expected payoff to each party is 3.

More generally, by choosing the probabilities of correlated strategies accordingly, any payoff tuple in the convex payoff space Π (shaded area in Figure 3.2) can be generated. For reference reasons, we denote the correlated probabilities z_u, $u \in \{1, ..., R\}$, where $R = R_i \times R_j$ is the number of strategy combinations and $\Sigma z_u = 1$. Thus, z_u is an element of a probability distribution $z = (z_1, ..., z_R)$. The notation is visualized in Matrix 3.9 above.

3.7 COORDINATION THROUGH CORRELATED STRATEGIES

Whereas it was the aim of the last section to demonstrate that in some games correlated strategies are necessary to convexify the entire payoff space, we are now concerned to show how correlated strategies can be used to improve upon a non-cooperative status quo. By constructing an appropriate random device, global welfare can be raised in some games which would not be possible by using pure strategy or an uncorrelated mixed NE only. Before studying the details, we wish to clarify two issues first:

1. How can the play of correlated strategies be motivated?
2. What are the necessary conditions to ensure that both parties comply with playing correlated strategies?

3.7.1 Motivation

Suppose that playing a particular set of correlated strategies is stable and that both parties benefit from the agreement. Then, one could think of the randomizing device as a third party; say, a coordinator who is in charge of making recommendations. In the environmental context the coordinator could be an international agency. Thus, instead of agreeing to play a particular strategy combination, parties agree to hand over the decision power to the coordinator. For instance, in the chicken game in Matrix 3.3 we have two NE in pure strategies – each of them favored by one party. Consequently, both parties will find it difficult to agree on one of those equilibria. If each player insists on playing its preferred equilibrium, it is not unlikely that they end up playing (na_1, na_2), each receiving a payoff of only 2. Moreover, also in the mixed uncorrelated strategy NE each party receives a payoff of only 2.8. Therefore, coordination may be attractive to both parties if they can find an appropriate 'coordination device'.

3.7.2 Compliance

A basic prerequisite that players agree to make a third party responsible for coordination is that the rules according to which these recommendations are given are common knowledge. Moreover, due to our fundamental assumption of a non-cooperative game, this coordinator has *no* enforcement power. Hence, the recommendations must be *self-enforcing* and the correlated strategies *must* constitute a *Nash equilibrium in correlated strategies*. We define:[17]

Definition 3.6: Nash equilibrium in correlated strategies
Let z be a probability distribution over all strategy combinations. Then, a Nash equilibrium in correlated strategies is a strategy tuple s^* and a recommendation of the coordinator z^* for which $\sum_{u=1}^{R} z^*_u \pi_i(s^*_i, s^*_{-i}) \geq \sum_{u=1}^{R} z^*_u \pi_i(s_i, s^*_{-i})$ $\forall$ $s_i \neq s^*_i$, $s_{-i} \neq s^*_{-i}$ and $i \in$ I hold.

From the definition it is evident that an NE in correlated strategies is a straightforward extension of our previous definition of an NE in pure or uncorrelated mixed strategies. This equilibrium is characterized by the strategy tuples s^* and z^*. However, since the coordinator recommends

playing a particular pure strategy combination with probability z^*_u, z^* contains all relevant equilibrium information.

3.7.3 Application

Now we want to find out which stable correlated probability distribution generates the highest aggregate payoff in a chicken game (Holler and Illing 1993, pp. 90ff.). For instance, in the chicken game in Matrix 3.3 aggregate payoffs in either of the two *pure* NE are 7.4 and in the mixed uncorrelated strategy NE the payoff is 5.6. Thus we look for a correlated equilibrium which brings about an aggregate payoff above 7.4.

It turns out that in the example the correlated equilibrium which generates the highest aggregate payoff is $z^* = (z^*_1 = 1/7,\ z^*_2 = z^*_3 = 3/7)$ with aggregate payoff $\Sigma\pi_i(z^*) = 7.656$. More generally we have:

Proposition 3.1
In a two-player chicken game there is a Nash equilibrium in correlated strategies with aggregate payoffs higher than in the pure and uncorrelated mixed strategy equilibria.

Proof: See Appendix I. QED

3.7.4 Discussion

The example shows that in games where no enforcement mechanism is available to achieve a cooperative solution, coordination can increase aggregate welfare. This result is particularly interesting because coordination takes place in a purely non-cooperative setting.

Whether it is possible to use a coordination device to improve upon a non-cooperative situation depends crucially on the particular game. For instance, it is easy to check that in a prisoners' dilemma game no correlated Nash strategy exists which could improve upon the Pareto-inferior outcome. In contrast, in an assurance game $z^*_1 = 1$ is an equilibrium which confirms that coordination should prove easy in this game. Generally, it would be interesting to find out whether it is possible to classify games into those in which correlated strategies can improve upon the pure and mixed NE and those in which this is not possible. Moreover, it seems promising to investigate whether and how it is possible to transform a game so that a coordination device can be successfully established. Finally, future research has to shed light on the question of which correlation devices are the most appropriate with respect to certain environmental conflicts. It seems that the environmental economics literature has not recognized this issue so far.

NOTES

1. The game derives its original name from the following situation (see, for example, Luce and Raiffa 1957, pp. 95ff.): Two suspects are arrested for some serious crime. The police lack sufficient evidence to convict the suspects unless one confesses. Therefore, the police confront the suspects with the following alternatives. If neither confesses, then both will be convicted for some minor offenses and sentenced to a short period in jail, say one year. If both confess, both will be sentenced to jail for five years. If one confesses and the other does not, the one who confesses will be released and the other will be sentenced to jail for ten years. If both prisoners were to agree not to confess, both would be released after one year. However, such an agreement is not stable; both prisoners will confess and will be sentenced to five years in jail.
2. If there existed an authority (in the sense of cooperative game theory) which was able to enforce international treaties, then the problem could be solved trivially. Holler and Illing (1993, pp. 23ff.) call this the Mafia solution to the PD game. In the original game (see the previous note) this implies that both prisoners do not confess because of fear of being killed by the Mafia members.
3. The following definitions assume the general case of possibly more than two players.
4. The degree of cooperation in experimental PD-type situations is reported in Axelrod (1984); Feeny *et al*. (1990); Gardner and Ostrom (1991); and Ostrom *et al.* (1990, 1994).
5. A report of the US government identifies 170 multilateral IEAs (USITC 1991). Up until now only political scientists have launched and conducted comprehensive empirical projects on the evaluation of the effectiveness of those IEAs. Due to different evaluation methods the results are rather ambiguous; see, for example, Benedick and Pronove (1992); Haas *et al.* (1993); Sand (1992); Victor *et al.* (1998); Werksman (1997); and Young and von Moltke (1994).
6. From the definition of a Nash equilibrium it should be clear that a dominant strategy equilibrium is always a Nash equilibrium too, that is, $S^{D*} \subseteq S^{N*}$. (However, the opposite does not hold.)
7. The chicken game derives its name from the following situation which has become famous in the movie *Rebel without a Cause* (see Rasmusen 1995, p. 87). Two young rebels race in their cars towards a cliff. The player who jumps out first is the chicken and the other the winner. The game is constructed such that each wants to be the winner, but prefers to be the chicken instead of being killed. Also, both players prefer to be embarrassed when they either both back out before the race or jump out of the car simultaneously. In the original game the strategies are continuous and therefore more complicated. However, the discrete representation as given in Matrix 3.3 preserves the basic incentive structure (a_i would be the strategy 'take no part in the race' or 'jump out first'; na_i would represent the strategy 'jump out last').
8. The expected payoff is computed from $0.25 \cdot 0.25 \cdot 4.6 + 0.25 \cdot 0.75 \cdot 2.2 + 0.75 \cdot 0.25 \cdot 0.25 \cdot 0.25 + 0.75 \cdot 0.75 \cdot 2 = 2.8$.
9. In the context of incomplete information there are more interpretations of mixed strategies (see Gibbons 1997; and Harsanyi 1973).
10. Of course, the example requires expanding the analysis to N countries. However, this extension is straightforward, as we argue in Section 3.5.
11. The term was coined by Schelling (1960). He used this term in games with several equilibria, where an equilibrium (or some equilibria) is (are) more convincing than others. The criteria which define a focal point are rather vague and, strictly speaking, do not follow from the game itself. They require some additional story, for example, psychological or sociological motives and so on. In the present context it seems obvious that the players play the Pareto-superior equilibrium because it is in both countries' interest. For a discussion of this point see, for example, Rasmusen (1989, pp. 36ff., pp. 228ff.).
12. There is a third Nash equilibrium in mixed strategies in which each country invests with probability $p_1^* = p_2^* = 2/10.5$. However, this would generate a lower payoff to each player

($\pi_i = 1.8$) than in either of the two pure strategy equilibria. Therefore, using the focal point argument, this equilibrium can be discarded.

13. In Chapter 6 and the subsequent chapters we discuss the equilibrium concept of a strong Nash equilibrium (Aumann 1959) which would pick $s^{N*(1)}$ as the only equilibrium in the assurance game. Roughly speaking, in a static context this concept requires a Nash equilibrium to be Pareto-efficient. In many games, however, this requirement implies that no equilibrium exists at all, as would be the case in the PD game. In the chicken game all pure uncorrelated strategy Nash equilibria are also strong Nash equilibria; this is not so, however, with the mixed strategy equilibrium.
14 The term 'assurance game' is probably due to the fact that players like to 'assure' each other of their interest in cooperation, so that they can coordinate on $s^{N*(1)}$ instead of getting stuck in $s^{N*(2)}$.
15. Of course, alternatively, economies of scale could be expressed by $\partial MC_i / \partial N^* < 0$ instead of $\partial MB_i^{j+1} / \partial N^* > 0$. This also applies to the subsequently discussed cases.
16. This game is basically a version of what is known in the literature as the 'battle of the sexes'; see, for example, Rasmusen 1989, p. 34.
17. As a matter of terminology, henceforth, when talking about a Nash equilibrium, we are referring to a pure strategy Nash equilibrium. With mixed Nash equilibrium we shall refer to an uncorrelated mixed Nash equilibrium and reserve the term correlated strategy Nash equilibrium for a Nash equilibrium in which correlated mixed strategies are played.

4. Finite dynamic games with discrete strategy space: a first approach

4.1 INTRODUCTION

In Chapter 3 it became apparent that in a static PD or a chicken game a full cooperative outcome cannot be achieved due to the free-rider incentive. Now, in a dynamic context, we have to investigate whether *contingent co-operation* can be established by using threats and punishments. The term 'contingent' emphasizes that it can never be an equilibrium strategy to cooperate *unconditionally* as long as there is a free-rider incentive. In order to establish contingent cooperation two requirements are necessary: first, it must be possible to check compliance; second, in case of a deviation from an agreed strategy, an appropriate punishment must be available to players.

Due to the assumption of complete information, the first requirement is satisfied by definition, though in reality it may only partially be fulfilled. Whether the second requirement can be satisfied depends basically on two questions:

1. How severe and credible is the punishment?
2. Does it pay to forgo an immediate gain from free-riding in order to be rewarded by cooperation?

For the first question the punishment options in a game are important. Obviously, the harsher the punishment, the higher is the potential of deterrence from cheating. However, if the player conducting the punishment also suffers some loss because of the punishment, credibility becomes an important issue. In the game theoretical literature the problem of credibility has attracted great attention and we shall deal with this issue throughout this book, gradually strengthening the definition of credibility.

The second question has to be answered by comparing the sums of periodically accruing payoffs. This is done by means of discounting with the following implications: the lower the discount rate, also called time preference rate, the lower are short-term gains from free-riding weighted against the long-term gains from cooperation; also, the lower the discount rate, the

more weight potential punishments receive. Consequently, a low discount rate is conducive to cooperation.

Though non-cooperative game theory only requires equilibrium strategies to be self-enforcing and abstracts from the question why a particular strategy combination is played, in the following we have implicitly in mind that countries coordinate their strategies. This coordination could be done in a (not modeled) preliminary stage in which countries' representatives negotiate and agree to pursue jointly a particular strategy which could be formalized by a treaty. This interpretation seems suggestive since, as it turns out below, there are often many equilibrium strategies in repeated games and some coordination device should therefore be expected.

For the interpretation of all results obtained in this chapter it is important to keep in mind that the time horizon is finite; that is, though the game may last a very long time, the definite end is known to all participants. The example given at the beginning of Chapter 3 in which countries decide whether to invest in a new energy production technology, is a possible scenario applying to the two-stage games discussed in Section 4.2. The only difference is that countries move sequentially instead of simultaneously. For the longer-lasting games (repeated games) the example has to be modified. Now, for instance, cooperation implies that filters in the power plants have to be renewed at regular intervals so as to meet certain emission standards. The general feature of examples described by dynamic games is that it is possible to alter a decision after some time.

In the following we start with an informal introduction to finite dynamic games, considering a two-stage sequential PD and chicken game in Section 4.2. Subsequently, in Section 4.3, we give a more formal definition of important terms and concepts which are needed in this and subsequent chapters. Sections 4.4 and 4.5 derive some theorems which generalize the preliminary results obtained in Section 4.2.

4.2 SOME EXAMPLES AND FIRST RESULTS

Consider a *two-stage sequential* chicken game with no discounting and only pure strategies. In a first step country 1 has the move and in a second step country 2. Thus, decisions have to be taken at two points in time. This is why this game is classified as a dynamic game, though the *real duration* of the game might be rather short. The normal form of this game is given in Matrix 4.1, where payoffs of Matrix 3.3 are assumed.

In Matrix 4.1 country 1 is the row player and country 2 the column player. Country 1 has only two strategies: either to invest, s_{11}, or not to invest, s_{12}; whereas country 2 has four strategies: that is, $S_2 = \{s_{21}, ..., s_{24}\}$.

| stands for 'provided that ...' or 'conditional on ...'. For instance, $s_{21} = (a_2|a_1; na_2|na_1)$ reads as 'invest provided country 1 invests and do not invest if country 1 does not invest'. For country 1 strategies and actions coincide (as in the one-shot game) since it moves first; however, for country 2 actions and strategies differ since it can condition its decision on country 1's behavior.

A routine check reveals that there are two (pure strategy) Nash equilibria in this game: $S^N = \{s^{N(1)} = (s_{12}, s_{22}), s^{N(2)} = (s_{11}, s_{24})\}$ of which the payoffs are printed in bold in Matrix 4.1. However, the question arises whether the second equilibrium involves credible strategies. It implies that country 2 plays a strategy of 'unconditional no investment' and country 1 gives in to this threat. That is, country 2 threatens to play na_2 if country 1 chooses na_1 even though its best reply is a_2.

Matrix 4.1 Two-stage sequential move chicken game[a]

	s_{21} $a_2\|a_1; na_2\|na_1$	s_{22} $na_2\|a_1; a_2\|na_1$	s_{23} $a_2\|a_1; a_2\|na_1$	s_{24} $na_2\|a_1; na_2\|a_1$
$s_{11} = a_1$	4.6 4.6	2.2 5.2	4.6 4.6	**2.2** **5.2**
$s_{12} = na_1$	2 2	**5.2** **2.2**	5.2 2.2	2 2

Note: [a] Assumption: Country 1 moves first.

From the *game tree* in Figure 4.1, which emphasizes the *sequence of moves*, the argument is even more transparent. Whereas a matrix is one way to represent a game in normal form, a game tree is frequently used to represent a game in extensive form which is defined as follows (see, for example, Eichberger 1993, pp. 14ff.; Gibbons 1992, pp. 115ff.):[1]

Definition 4.1: Extensive form of a game
The extensive form representation of a game specifies: (a) the players in the game; (b) when each player has the move; (c) the actions available to a player at each decision node; (d) the information each player has if s/he has the opportunity to move; and (e) the payoffs to each player at each possible end node.

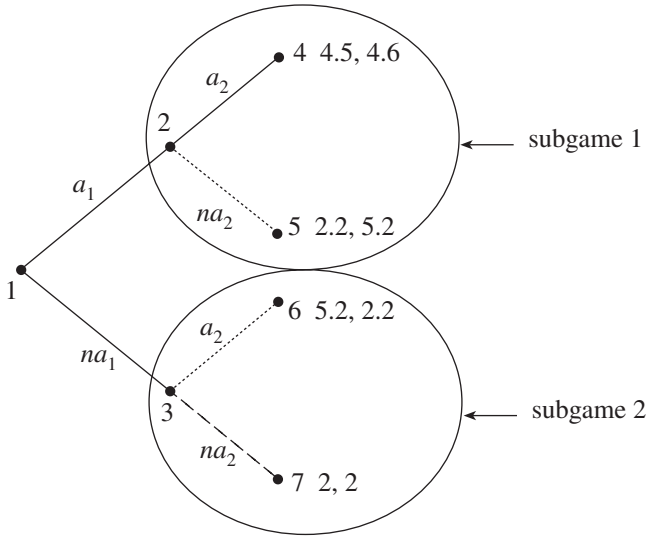


Figure 4.1 Two-stage sequential move chicken game

In Figure 4.1 the threat of country 2 entailed in the second Nash equilibrium is indicated by the broken line whereas country 2's best replies in the second stage are represented by the dotted lines branching off from nodes 2 and 3. Since (due to the assumption of complete information) country 1 knows country 2's strategies, it can solve the game for country 2 (in particular it knows that country 2 cannot do any better than to play a_2 if it plays na_1) and can use its *first-mover advantage* to enforce its preferred equilibrium $s^{N(1)}$. In other words, the threat contained in $s^{N(2)}$ is disclosed as an empty threat. Thus in the sequential chicken game, though there are *two Nash equilibria* (NE), only the first equilibrium $s^{N(1)}$ is credible, that is, it is a *subgame-perfect equilibrium* (henceforth abbreviated SPE), that is, $s^{\text{SPE}} = s^{N(1)}$. Hence, in a two-stage sequential move chicken game, the outcome (unilateral investment) does not differ from the one-shot chicken game, though equilibrium strategies differ.

The concept of subgame-perfect equilibrium is due to Selten (1965) and derives its name from the concept's requirement that in equilibrium a strategy must be a best response in *every subgame of the game*. Simply speaking, a subgame is a part of a game which can be viewed as a game in its own right. For an extensive form representation a subgame can be defined as follows:

Definition 4.2: Subgame in an extensive form game
A subgame begins at a decision node which is a singleton in every player's information partition and includes all subsequent nodes following this node including the terminal nodes.

For the understanding of this definition two more definitions are necessary (Rasmusen 1989, pp. 48ff.):

Definition 4.3: Information set
An information set of a player comprises all decision nodes at a particular point in the game where the player knows that s/he has to move, but does not know which node has been reached.

Definition 4.4: Information partition
A player's information partition at a particular point in the game comprises all his/her information sets such that (a) each path is represented by one node in a single information set in the partition, and (b) the predecessors of all nodes in a single information set are in one information set.

Thus, the information partition refers to the information set of a player at a particular stage of the game. For example in Figure 4.1, nodes 2 and 3 are members of the same information partition but nodes 1 and 2 are not. The information partition after the first move contains two information sets with one single node. Hence, there are two subgames beginning at the second stage of the game (see Figure 4.1). Definition 4.2 implies that the entire game, starting at node 1, is also a subgame (not encircled) besides the two subgames shown in Figure 4.1.[2]

In contrast, in a simultaneous move chicken game there is only one subgame comprising the entire game. Its extensive form representation is shown in Figure 4.2, where the broken elipse around nodes 2 and 3 indicates that country 2 does not know at which node it is.[3, 4] That is, its information partition contains only one information set with two nodes. Only at node 1 player 1's information partition contains a single node, player 2's information partition contains nodes 2 and 3 and therefore only the entire game is a subgame.

With these elaborations an SPE may now be defined as follows (Selten 1975):

Definition 4.5: Subgame-perfect equilibrium (SPE)
A subgame-perfect equilibrium is a strategy combination constituting a Nash equilibrium in every subgame of the entire game.

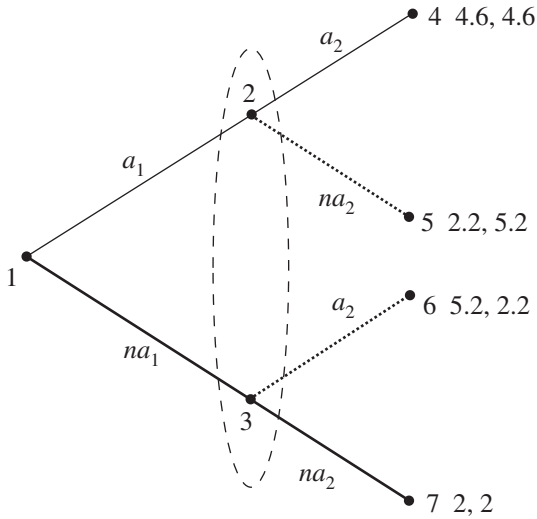


Figure 4.2 Simultaneous move chicken game

The concept implies that a strategy is only subgame-perfect provided it is not only a best response *on the equilibrium path*, that is, the path which is actually played (for example, (na_1, a_2) in Figure 4.2), but also *a best response off the equilibrium path* (for example, na_2 if a_1 were played). A threat to enforce a particular path is only credible provided it is really carried out if tested, though the challenge is off the equilibrium path. The definition above may be contrasted with an NE in a dynamic game:

Definition 4.6: Nash equilibrium (NE) in a dynamic game
A Nash equilibrium is a strategy combination including strategies which are best mutual responses to each other with respect to the entire game.

The definition implies that it is necessary to distinguish between an *NE of the entire game* (Definition 4.6) and a *stage game NE*. The distinction is important since, as will become apparent below, there are repeated games in which an NE (and an SPE) does not necessarily require the playing of a Nash strategy in each round.[5]

Comparing both definitions, it is evident that the set of SPE is a subset of the set of NE in a game. Hence, subgame-perfection is a *refinement* of the concept of Nash equilibrium which makes predictions about the outcome of a dynamic game more reliable.

The example highlights the usefulness of the extensive form representation. In particular, in sequential move games, the strategic interaction between the players is more transparent than in a normal form representation. However, once the number of stages is large (and/or the number of players and actions is large), a game tree may become intractable and one is led back to the normal form representation.

Let us now turn to the question of whether it is possible to obtain more positive results if the chicken game is played repeatedly; that is, the constituent game as described by Figures 4.1 and 4.2 is played several times. Assume first simultaneous moves, three rounds and that players agree on the following strategy: cooperate in the first round, that is, (a_1, a_2), and play the two pure (stage game) strategy equilibria, that is, (a_i, na_j), in turn in the last two rounds, which implies a 'good' equilibrium for one player and a 'bad' equilibrium for the other player (LHS payoffs in the inequalities in (4.1)). If a player deviates (which, by the definition of a stage game Nash equilibrium, will only happen, if at all, in the first round), the punishment involves playing that player's 'bad' equilibrium for the rest of the game (RHS payoffs in the inequalities in (4.1)). Since:

$$a_i + b_i + c_i > c_i + b_i + b_i \Leftrightarrow a_i > b_i \text{ and } a_j + c_j + b_j > c_j + b_j + b_j \Leftrightarrow a_j > b_j \tag{4.1}$$

hold – using the general payoffs of Matrix 3.2 – this is an equilibrium strategy. In fact, it is an SPE since the threat implies playing an NE in every remaining subgame. The play of a stage game NE in the terminal rounds is necessary for two reasons:

1. In the last round a stage game NE must be played in any case since deviations cannot be punished.
2. The alternation between the two Pareto-undominated NE in at least two terminal rounds is necessary because if countries played the same equilibrium twice one (potential) deviator could not be punished.

Obviously, in longer-lasting games cooperation can be established for the first T-2 rounds, and only in the last two rounds must an NE be played.

Now consider a sequential move version of the chicken game. Again, we approach this problem by the method of *backwards induction*; that is, we consider first what happens in the last round and then work back to the first round. From the discussion of the two-stage game version of this game it is clear that in the last round the only (stage game) SPE is $s^{SPE} = s^{N(1)} = (s_{12}, s_{22})$ – country 1's preferred equilibrium. Hence, it is not possible to punish country 1 in the penultimate round subgame-perfectly if it deviates from

the agreement on (a_1, a_2). In other words, there is no reason for country 1 to follow a contingent cooperative strategy and to sacrifice a free-rider gain since such behavior is not rewarded. Consequently, the stage game equilibrium $s^{N(1)}$ will also be played in the second last round and, due to the symmetric structure of the game, in any previous round. Thus, the only SPE in this game involves the repeated play of $s^{N(1)}$.

However, since $s^{N(2)} = (s_{11}, s_{24})$ is a stage game NE, cooperation can be established as a Nash equilibrium by using strategy $s^{N(1)}$ against country 1 and strategy $s^{N(2)}$ against country 2 as a punishment in case it deviates, as in the simultaneous move version.

The results may be summarized as follows:

Proposition 4.1
In a finitely repeated simultaneous move two-player chicken game which is played for T periods, assuming payoff relations $c_i > a_i > b_i > d_i \; \forall \; i \in I$ (see Matrix 3.2) and no discounting, cooperation can be supported as a subgame-perfect equilibrium (in pure strategies) for the first T-1 rounds. In a sequential move chicken game, cooperation can only be established as a Nash equilibrium.

From the previous information it is easy to solve for the equilibrium in a PD game. From Chapter 3 it is known that in a one-shot game $s^D = s^N = (na_1, na_2)$ is the only equilibrium in this game. In contrast to the chicken game, in a two-stage sequential move PD game the first mover *cannot* enforce its preferred outcome since the second mover has a dominant strategy. That is, the first mover, say, country 1, will play na_1 since it knows that country 2 will play na_2 in any case. With respect to Matrix 4.1, adjusting payoffs for the PD game accordingly, $s^N = (s_{11}, s_{24})$ is the only NE, implying that both countries will not invest.

This result carries over if the PD game is played repeatedly. Regardless of whether countries move simultaneously or sequentially, in the last round a stage game NE (which is also an SPE in the sequential move version) will be played. Hence, cooperation does not pay in any previous round and the only equilibrium is the repeated play of the stage game NE (Holler and Illing 1993, p. 141).

Proposition 4.2
In a finitely repeated PD game the only equilibrium of the whole game is the T-fold repetition of the Nash equilibrium of the stage game.

Taken together, playing a chicken game for more than one period can resolve the deadlock caused by the free-rider incentive. If the number of

stages is large enough (and time is not discounted), the fully cooperative outcome can approximately be obtained. If moves occur sequentially cooperation can only be established as an NE; if they are taken simultaneously cooperation can be supported as an SPE. In a PD game the only equilibrium (SPE and NE) involves the repeated play of the stage game NE and cooperation fails.

4.3 THE CONCEPTUAL FRAMEWORK

4.3.1 Preliminaries

From the chicken game example it appeared that results may depend crucially on whether moves occur *simultaneously* or *sequentially*. This is true whether the constituent game is played once or repeatedly. However, the literature on repeated games either implicitly or explicitly restricts attention to simultaneous moves only. This restriction may be defended on two grounds: first, from a conceptual point of view, including sequential moves renders definitions more complicated; second, from a motivational point of view, one may argue that in longer-lasting games players move simultaneously 'on average'. In particular, if the problem itself does not suggest a certain order of moves, simultaneous moves seem a 'natural' assumption (see Chapter 2). Acknowledging these arguments, we shall continue to stress the difference between simultaneous and sequential moves only in the subsequent parts of this chapter, implying that the subsequent definitions and theorems are more general than those typically found in the literature. From Chapter 5 onwards, however, the term repeated games is exclusively reserved for simultaneous move games.

From the previous section it appeared that a distinction can be made between strategies played *within* a stage game and strategies which are played in the *overall* game. From Chapter 3 it is known that by playing mixed strategies any payoff tuple in the convex payoff space can be obtained. This is also possible in dynamic games. However, in order to avoid dealing with incomplete information games, the mixing device must be assumed to be publicly observable. Moreover, now, there is an additional possibility of generating convex payoff tuples if different stage game strategies are played in various periods. This amounts to mixing strategies *between* stages and not only *within* a stage. However, departing from the assumption of *stationary strategies* implies that payoffs differ between stages and *simple strategy profiles* in the sense of Abreu cannot be constructed when testing for an equilibrium (Abreu 1986, 1988). Therefore, non-stationary strategies are not treated in the following.

For simple strategies it suffices to construct a two- or at most three-phase strategy to check whether a payoff tuple can be backed by equilibrium strategies. This implies that a game can typically be divided into a *cooperative phase* (also called a normal phase), in which the agreed strategy is played, and a *punishment phase*. The latter phase is a theoretical device to test that, provided a player deviated from the equilibrium play (cooperative phase), a credible punishment would be possible. Depending on the equilibrium concept and the number of players, the punishment phase may be further divided into subphases.

Moreover, simple strategies often use strategy profiles which punish all deviators in the same way and/or specify the same degree of punishment independently of the kind of deviation. On the one hand, apart from simplifying the analysis, this may be defended by noting that too complicated strategy profiles will hardly be realized by players in reality anyway. They require high computational effort and the transparency of threats may suffer. On the other hand, intuition and international law suggest that punishment should be conducted according to the severity of the crime.

In the following we restrict our attention to simple punishment profiles since many proofs of the subsequent theorems would get rather messy by introducing such complications. However, in Chapter 12 we shall analyze the effect of relative and restricted punishment profiles on the stability of an IEA in the context of renegotiation-proof strategies. Moreover, for notational simplicity, the following definitions are restricted to pure strategies, though an extension to the more general case of mixed stationary strategies is straightforward.

4.3.2 Important Terms and Definitions

History and subgame in a normal form game

Suppose a *constituent game* of $\Gamma(\mathrm{I}, \mathrm{S}, \Pi)$ and complete information is repeated T times. Let the number of stages within a constituent game be v. Hence the total number of stages is $\mathrm{V} = \mathrm{T} \cdot v$.[6] The game starts at time $t = 0$. After each *stage* actions are observed by all players; after each *period* payoffs are received, which occurs for the first time at $t = 1$.[7] The history of a game at stage V, h^{V}, comprises the sequence of all action combinations played up to stage V. For instance, in a two-period simultaneous move PD game ($\mathrm{V} = \mathrm{T} = 2$; $v = 1$) in which the stage game NE is played twice, $h^1 = (na_1, na_2)$ and $h^2 = ((na_1, na_2), (na_1, na_2))$. h^0 is empty by definition. The set of possible histories up to period t is therefore the t-fold Cartesian product of the set of possible action combinations of the stage games, that is, $t \cdot \mathrm{A}_1 \times \ldots \times \mathrm{A}_N = t \cdot \mathrm{A}$. In a repeated sequential move stage game, where t periods have been played and in the $t+1$ stage, say, two moves have been made, the

set of histories would be given by $t \cdot A \times A_1 \times A_2$. A history-dependent strategy of player i, σ_i, is an instruction for each possible history in the game, that is, $\sigma_i = (s_{ir}(h_1^1), s_{ir}(h_2^1), ..., s_{ir}(h_1^2), ..., s_{ir}(h_1^{\bar{V}_i}), ..., s_{ir}(h_\Omega^{\bar{V}_i}))$.[8] The superscript $\bar{V}_i$ of h_m^V refers to the last stage where player i has a move in the game, the subscript m to the history at stage $\bar{V}_i$, $m \in \{1, ..., \Omega\}$ where $\Omega = A \cdot (T-1) \times A_1 \times ... \times A_{i-1}$. If each player chooses a strategy σ_i, this strategy combination $\sigma = (\sigma_i, \sigma_{-i})$ leads to the *path of play* from which payoffs $\pi_{it}(\sigma)$ at time t are received by player i.

With this notation we can also define a *continuation history* of h_m^V, denoted $h_m^{V'}(h_m^V)$, $V' > V$, as any history that follows h_m^V and where the action combinations of h_m^V have been played previously to V' (see Eichberger 1993, pp. 220ff.). Thus, a subgame may alternatively be defined as follows:

Definition 4.7: Subgame in a normal form game
A subgame in a normal form game at stage V is a continuation history $h_m^{V'}(h_m^V)$ following a history h_m^V. For each possible history up to stage V a subgame starts.

Accordingly, a continuation strategy of player i after history h^V is defined as $\sigma_i(h_m^V) = (s_{ir}(h_m^{V+1}(h_m^V)), s_{ir}(h_m^{V+2}(h^V{}_m)), ..., s_{ir}(h_m^{\bar{V}_i}(h_m^V)))$.

Discounting[9]

There are three reason for assuming that agents value a payoff received today higher than one at some future date: (a) agents are impatient; (b) future payoffs are uncertain by their nature: there is always a risk that these payoffs will never be realized for some reason; (c) payoffs received today can be invested to yield interest: hence, delayed payoffs involve opportunity costs.

Discounting makes payoffs received at different stages of a game comparable. Since information about optimal strategies is required right from the beginning of a game, it is advisable to discount all payoff streams to the initial stage of a game. The net present value of a payoff stream received over T periods at time $t = 0$ is computed by using the *discount factor* δ_i, $\delta_i = 1/(1+\rho_i)$ and $0 \leq \delta_i \leq 1$, where the index refers to the player. ρ_i denotes the *discount rate* or *time preference rate*. For example, a discount rate of 10 percent implies $\rho_i = 0.1$ and $\delta_i = 0.909$.

Omitting the subscript i (denoting the player) for convenience, we have the following relations (see, for example, Eichberger 1993, p. 211; Holler and Illing 1993, p. 140):

$$\pi^\delta = \pi_1 + \delta\pi_2 + ... + \delta^{T-1}\pi_T = \sum_{t=0}^{T-1} \delta^t \pi_t; \sum_{t=0}^{T-1} \delta^t = \frac{\pi}{1-\delta}(1-\delta^T) \text{ if } \pi_1 = ... = \pi_T \tag{4.2}$$

$$\pi^{\delta} = \delta\pi_1 + \delta^2\pi_2 + \ldots + \delta^{T}\pi_T = \sum_{t=1}^{T} \delta^t\pi_t; \sum_{t=1}^{T}\delta^t = \frac{\delta\pi}{1-\delta}(1-\delta^{T}) \text{ if } \pi_1 = \ldots = \pi_T \tag{4.3}$$

$$\bar{\pi}^{\delta} = \frac{1}{\sum_{t=0}^{T-1}\delta^t} \cdot \sum_{t=0}^{T-1}\delta^t\pi_t; \frac{1}{T}\cdot\sum_{t=0}^{T-1}\pi_t \text{ if } \delta = 1; \pi \text{ if } \pi_1 = \ldots = \pi_T. \tag{4.4}$$

In (4.2) the net present value at time $t = 0$ is computed when payoffs are received at the beginning of each period, starting at $t = 0$. Alternatively, it is the net present value at time $t = 1$ if payoffs are received at the end of each period and for the first time at $t = 1$. ($T - 1$ has to be replaced by T if players receive payoffs in the last round.) In (4.3) the net present value of a payoff stream at time $t = 0$ is computed, when payoffs are received at $t = 1$ for the first time, and then after each stage.

Sometimes it is convenient to express payoffs as average payoffs $\bar{\pi}^{\delta}$ because this makes the payoffs received in a dynamic game comparable to those derived in a one-shot game (see (4.4)). For large T, with $T \to \infty$ at the limit, the average payoff is defined as:

$$\bar{\pi}^{\delta} = (1-\delta)\cdot\sum_{t=0}^{\infty}\delta^t\pi_t \text{ if } \delta < 1; \pi \text{ if } \delta < 1 \text{ and } \pi_1 = \ldots = \pi_T. \tag{4.5}$$

Since average payoffs may not be defined for $\delta = 1$ (see, for example, Eichberger 1993, p. 211; Holler and Illing 1993, p. 140), we usually assume δ to be strictly smaller than 1, though δ might be close to 1.

Normal form

The normal form of a repeated game, in which the constituent game $\Gamma(I, \Pi, S)$ is repeated T times, may be written as $\Gamma^{T}(I, \Pi, \Sigma, \Delta, T)$, with I the set of players, Π the set of payoffs, $\Sigma = \{\Sigma_1, \ldots, \Sigma_N\}$ the set of history-dependent strategies ($\sigma_i \in \Sigma_i$), $\Delta = \{\delta_1, \ldots, \delta_N\}$ the set of discount factors, and T the number of stages which may be finite, that is, $T = T$, or infinite, that is, $T = \infty$. Alternatively, the normal form of a dynamic game may be written as $\Gamma^{T}(I, \Pi^{\delta}, \Sigma)$ or $\Gamma^{T}(I, \overline{\Pi}^{\delta}, \Sigma)$, where Π^{δ} is the set of discounted payoffs and $\overline{\Pi}^{\delta}$ the set of average payoffs.

Nash and subgame-perfect equilibria in repeated games

The definitions above allow us to define an NE and an SPE in a straightforward and compact way:

Definition 4.8: Nash equilibrium (NE) in a repeated game
A Nash equilibrium in a repeated game $\Gamma^{T}(I, \Pi^{\delta}, \Sigma)$, $T = T$ or $T = \infty$, is a strategy combination $\sigma^* = (\sigma_i^*, \sigma_{-i}^*)$ for which $\pi_i^{\delta}(\sigma_i^*, \sigma_{-i}^*) \geq \pi_i^{\delta}(\sigma_i, \sigma_{-i}^*) \forall$ $i \in I$ and $\sigma_i \neq \sigma_i^*$ holds.

Definition 4.9: Subgame-perfect equilibrium (SPE) in a repeated game
A subgame-perfect equilibrium is a continuation strategy combination $\sigma^*(h_m^V) = (\sigma_1^*(h_m^V), \ ..., \ \sigma_N^*(h_m^V))$ for which $\sum_{t=g}^{T} \delta_i^t \pi_{it}(\sigma_i^*(h_m^V), \sigma_{-i}^*(h_m^V)) \geq \sum_{t=g}^{T} \delta_i^t \pi_{it}(\sigma_i(h_m^V), \sigma_{-i}^*(h_m^V)) \forall i \in I, m \in \{1, ..., \Omega\}; g \in \{1, ..., T\}$ and $\sigma_i(h_m^V) \neq \sigma_i^*(h_m^V)$ holds for each possible history h_m^V.

It is important to note that the continuation strategy at time t (stage V) must be a best reply with respect to *all* possible histories m and not only with respect to the path played up to period t (stage V). That is, if player j deviated from the equilibrium path, player i's strategy must still be optimal.

To prepare the following discussion, we have to define four terms: *minimax*, *maximin*, *maximax* and *defection payoff*. They are defined with respect to the payoffs in a stage game.

Minimax payoff
If player j punishes player i, the *minimax payoff* to player i is defined as follows:[10]

$$\pi_i^M = \min_{s_j} \max_{s_i} \pi_i(s_i, s_j). \tag{4.6}$$

That is, given player i chooses an optimal response to the punishment of player j, s/he can guarantee him- or herself a payoff of π_i^M. We denote the minimax strategy of player j when punishing player i by m_j^i and the best reply of player i to the punishment by m_i^i. Thus the superscript denotes the player who is punished and the subscript a player's strategy. We write $\pi_i^M = \pi_i^M(m_i^i, m_j^i)$ for short. However, sometimes it is necessary to distinguish between $\pi_i^{M(i)}$ and $\pi_j^{M(i)}$ where $\pi_i^{M(i)} = \pi_i^M$ and the latter payoff refers to the punisher j's payoff when minimaxing player i.

Maximin payoff
In contrast, the *maximin payoff* of player i, which is also called his/her *security level*, is the lowest payoff a player can guarantee him- or herself in a stage game. It is defined as:

$$\pi_i^{SC} = \max_{s_i} \min_{s_j} \pi_i(s_i, s_j) \tag{4.7}$$

where $\pi_i^M \geq \pi_i^{SC}$ holds. As an auxiliary device, minimax and maximin payoff may be distinguished with respect to the sequence of moves (Rasmusen 1989, pp. 103ff.). According to maximin, player *i* moves first, whereas this would be player *j* under minimax. Under maximin player *i* chooses his/her best reply first, assuming that player *j* wants to get him/her. Under minimax player *j* chooses a strategy to punish player *i*, taking into account player *i*'s reaction, but moves first.

In the PD and the chicken game, minimax and maximax payoff are identical. This is also true in the extended PD game I in Matrix 4.2, where the basic PD game has been extended to include a third 'punishment action' p_i. In Matrix 4.2 it is assumed that if player *j* punishes player *i*, the best reply of *i* is a_i, resulting in a minimax payoff of 1. Moreover, by playing a_i, player *i* can guarantee a payoff of 1 if player *j* wants to punish him/her and therefore security level and minimax payoff coincide. In contrast, in the extended PD game II in Matrix 4.3 the minimax payoff is 1, whereas the maximin payoff is 0.

Matrix 4.2 Extended PD game I

	a_2	na_2	p_2
a_1	3.2 3.2	1.4 4.4	1 0
na_1	4.4 1.4	**2** **2**	0 0
p_1	0 1	0 0	0 0

Matrix 4.3 Extended PD game II

	a_2	na_2	p_2
a_1	3.2 3.2	1.4 4.4	0 0
na_1	4.4 1.4	**2** **2**	0 0
p_1	0 0	0 0	**1** **1**

Thus, in a *simultaneous move* repeated game, playing the minimax strategy at every stage game is the harshest possible punishment player *j* can inflict on player *i*. That is, player *j* can only guarantee to put player *i* down to π_i^M in the long run, though other strategy combinations may lead to a lower payoff to player *i*. This implies that in such games each player must receive at least the minimax payoff as an average payoff to take part in the game. All payoffs for which $\pi_i \geq \pi_i^M$ hold are therefore called *individually*

rational. The convex payoff space defined by $\pi_i \geq \pi_i^{\mathrm{M}} \; \forall \; i \in \mathrm{I}$ is called *individually rational payoff space* and denoted Π^{IR}.

In a *sequential move* repeated game, player *j* may have an even stronger punishment available if s/he moves *after* player *i*. In this case, player *j* can put player *i* down to his/her maximin payoff. Then $\Pi^{\mathrm{IR}} = \{\pi \mid \pi_i \geq \pi_i^{\mathrm{SC}}$ and $\pi_j \geq \pi_j^{\mathrm{M}}\}$. Since $\pi_i^{\mathrm{M}} \geq \pi_i^{\mathrm{SC}}$ this implies that the first mover may have a disadvantage in a repeated game, though s/he has typically a first-mover advantage in a two-stage game as demonstrated for the chicken game. We return to this point later.

Note that in the extended PD game I country *j*'s payoff when minimaxing country *i* is lower than its own minimax payoff, that is, $\pi_j^{\mathrm{M}(j)} > \pi_j^{\mathrm{M}(i)}$. This is not what one would normally expect when thinking of a punishment; however, this possibility cannot be ruled out *a priori*. In a broader context, one may think of trade sanctions used to enforce an environmental agreement, which, depending on trade flows and terms of trade, can hurt the punisher more than the free-rider.

In the extended PD game II the payoffs in the 'bad' NE are equivalent to the minimax payoffs, and punisher and punished player receive the same payoff.

Maximax payoff

The maximax payoff of a player *i* is defined as:

$$\pi_i^{\mathrm{U}} = \max_{s_j} \max_{s_i} \; \pi_i(s_i, s_j) \tag{4.8}$$

where the superscript U stands for upper boundary. Basically, this payoff is derived if all players make an effort to maximize player *i*'s payoff. In other words, it is the highest payoff to player *i* obtainable in a particular game and therefore lies on the Pareto frontier. For instance, in both extended PD games the maximax payoff is $\pi_i^{\mathrm{U}} = 4.4$.

Deviation payoff

The deviation payoff, denoted π_i^{D}, which may also be called the free-rider payoff, is the maximum payoff a player can achieve in a stage game if s/he deviates from an agreed strategy and the other player complies with the agreement. Suppose s_i and s_j denote some agreed strategy combination, then we have:

$$\pi_i^{\mathrm{D}} = \max_{s_i} \; \pi_i(s_i, s_j) \tag{4.9}$$

which may alternatively be written as $\pi_i^{\mathrm{D}} = \pi_i(s_i(s_j), s_j)$ where $s_i(s_j)$ denotes the best reply (deviation) of player *i* if player *j* plays strategy s_j. Obviously,

$\pi_i^{\mathrm{D}} \leq \pi_i^{\mathrm{U}} \; \forall \; i \in \mathrm{I}$, that is, π_i^{U} is an upper bound for π_i^{D}. For instance, in a one-shot version of the extended PD games I and II, if players agree on playing action combination (a_1, a_2), the deviation payoff is $\pi_i^{\mathrm{D}} = 4.4$.

4.4 SOME GENERAL RESULTS

With the definitions of the previous sections we are now prepared to derive some theorems, some of which are a generalization of the results we obtained in Section 4.2. Theorem 4.1 is useful in that it forms the basis for many strategy profiles below:

Theorem 4.1
If a simultaneous move constituent game Γ has a Nash equilibrium s^N, then the repeated play of s^N in every stage game is a subgame-perfect equilibrium of the repeated game Γ^{T}. That is, the strategy combination $\sigma(h^t) = (s_i^N(h^{t+1}(h^t)),\ s_{-i}^N(h^{t+1}\ (h^t)),\ \ldots,\ s_{-i}^N\ (h^{\mathrm{T}}(h^t)))$ forms a subgame-perfect equilibrium for any $\delta_i \in [0, 1]$ and $t \in \{1, \ldots, \mathrm{T}\}$. In a sequential move game the repeated play of a stage game subgame-perfect equilibrium constitutes a subgame-perfect equilibrium of Γ^{T}.

Proof: Theorem 4.1 is rather simple and a proof obvious.[11] Assume simultaneous moves and that player i plays some other stage game strategy s_i' instead of s_i^N, but all other players choose continuously s_{-i}^N. $s_i' \neq \mathrm{s}_i^N$ in at least one stage game implies $\pi_i(\mathrm{s}_i^N, s_{-i}^N) \geq \pi_i(\mathrm{s}_i', s_{-i}^N)$. Consequently we have:

$$\sum_{t=g}^{\mathrm{T}} \delta_i^t \pi_i(s^N) \geq \pi_i(s_i', s_{-i}^N) + \sum_{t=g+1}^{\mathrm{T}} \delta_i^t \pi_i(s^N) \geq \sum_{t=g}^{\mathrm{T}-1} \delta_i^t \pi_i(s_i', s_{-i}^N) \forall\ g \in \{1, \ldots, \mathrm{T}\}. \tag{4.10}$$

Replacing the superscript N by SPE above, the case of sequential move games is covered as well. QED

Theorem 4.2
In a finitely repeated simultaneous move game Γ^{T} with a unique stage game Nash equilibrium s^N, the only subgame-perfect equilibrium of the whole game is the strategy combination $\sigma^{\mathrm{SPE}}(h^t) = (\sigma_1^{\mathrm{SPE}}(h^t)(s_1^N),\ \ldots,\ \sigma_N^{\mathrm{SPE}}(h^t)(s_N^N))$; that is, the stage game Nash equilibrium strategy is played by all players in every period. In a sequential move game with a unique subgame-perfect stage game equilibrium the only subgame-perfect equilibrium s^{SPE} is the repeated play of s^{SPE}.

Proof: The proof is obvious (see, for example, Binmore 1990, pp. 353ff.; Eichberger 1993, p. 224; Myerson 1991, pp. 309ff.). According to backward induction in the last period T, subgame-perfection requires that a player plays his/her best stage game response strategy. This is true whether no deviation occurred in previous rounds or whether some punishment strategy is played as a response to defection. Thus, in period T each player must play $\sigma_i^{SPE}(h^T) = s_i^N$ in a simultaneous move game or $\sigma_i^{SPE}(h^T) = s_i^{SPE}$ in a sequential move game.[12] Hence, in period $T-1$, no other strategy than $\sigma_i^N(h^{T-1}) = s_i^N$ ($\sigma_i^N(h^{T-1}) = s_i^{SPE}$) can be played because no adequate threat is available. Obviously, this holds also for any period previous to $T-1$. QED

With respect to the concept of NE, a similar theorem to Theorem 4.2 can be established; however, the statement has to be slightly altered (see, for example, Eichberger 1993, pp. 224ff.; Friedman 1986, pp. 95ff.):

Theorem 4.3
In a finitely repeated simultaneous (sequentially) move game Γ^T with unique Nash equilibrium s^N of the stage game Γ and where $\pi_i^N(s^N) = \pi_i^M$ holds for all $i \in I$ ($\pi_i^N(s^N) = \pi_i^{SC}$ for the first mover and $\pi_i^N(s^N) = \pi_i^M$ for the other players), the only Nash equilibrium of the entire game is the strategy combination $\sigma^N(\sigma_1^N(s_1^{N*}), ..., \sigma_N^N(s_N^{N*}))$, that is, the stage game Nash equilibrium strategy is played in every period $t = 1, ..., T$.

In games in which $\pi_i^N(s^N) > \pi_i^M$ holds for all $i \in I$ ($\pi_i^N(s^N) > \pi_i^{SC}$ for the first mover and $\pi_j^N(s^N) > \pi_j^M$ for the other players), any payoff vector in Π^{IR} can be obtained as an average payoff vector for large T and discount factors close to one playing Nash equilibrium strategies.

Proof: The first part of Theorem 4.3 simply follows from the fact that in the terminal round s^N has to be played and if no harsher punishment is available than the Nash payoff no other strategy combination can be enforced in previous rounds. The proof of the second part of Theorem 4.3 is only briefly sketched since it is similar in spirit to the proof of Theorem 4.4 below (see van Damme 1991, pp. 195ff.; Sabourian 1989, pp. 75–6). First, one constructs a trigger strategy[13] which calls for minimaxing a deviator until the end of the game. Second, one shows that for discount factors close to 1 the payoff stream from cooperating for some time t^*, and then playing the stage game NE for the rest of the game ($t^* + 1$ until T) is higher than when deviating at any time $t = t^0$, $t^0 \leq t^*$, and then being minimaxed for the rest of the game. Third, one determines the limit of the average payoff for large T if players follow the equilibrium

path which approaches the stage game payoff sustained in the finitely repeated game for t^* periods. QED

Theorem 4.2 and the first part of Theorem 4.3 cover the PD game in its simultaneous and sequential move version. Since $\pi_i^N = \pi_i^M = \pi_i^{SC}$ in this game no other strategy tuple than the stage game NE will be played. Theorem 4.2 also covers the sequential move chicken game where there is only one stage game SPE which has to be played throughout the game if one requires strategies to be subgame-perfect. A typical example which illustrates the second part of Theorem 4.3 is the extended PD game I in its simultaneous move version where $\pi_i^N = 2 > \pi_i^M = 1$. To see this, assume that the game is played three times and assume no discounting. In the first round the parties agree to play (a_1, a_2), and in the second and third rounds to play the stage game NE (na_1, na_2). The payoff stream of such an agreement is $3.2 + 2 + 2 = 7.2$. If a player deviates s/he will be minimaxed afterwards which delivers the payoff stream $4.4 + 1 + 1 = 6.4$. Since $7.2 > 6.4$, free-riding can be deterred by such a strategy. Accordingly, by increasing the number of repetitions, only in the last two rounds must the stage game NE be played and hence for large T ($T \to \infty$) the average equilibrium payoff converges to the stage game payoff from mutual cooperation, that is, 3.2. With discounting the same can be shown provided the discount factors are sufficiently close to 1. Of course, the above strategy is only an NE of the entire game and not subgame-perfect as follows from Theorem 4.2.

Though we cannot refer to an example to which the second part of Theorem 4.3 applies in the case of sequential moves, it should be evident that in repeated games payoff tuples which imply a payoff above the minimax payoff to all players except the first mover, who may receive only a payoff above his/her security level, can be sustained in the long run. Moving first may in this case be a disadvantage since the other players have a harsher punishment (in the sense of NE strategies) available to them.

The next result is an extended version of a theorem derived by Friedman (1985):

Theorem 4.4

Let Γ^T (I, Π^δ, Σ) be a finitely repeated simultaneous move game with at least two strictly Pareto-undominated stage game Nash equilibria, that is, $\pi_i^{N(1)}(s^{N(1)}) \neq \pi_i^{N(2)}(s^{N(2)})$ $\forall$ $i \in$ I. Then for large T and δ_i close to 1 any average payoff vector $\pi^* = (\pi_i^*, \pi_{-i}^*)$ for which $\pi_i^*(s) \geq \pi_i^N(s^i)$ $\forall$ $i \in$ I holds, where $\pi_i^N(s^i)$ denotes the lowest Nash payoff to player i, can be obtained in the T-fold repeated game by playing subgame-perfect strategies. In a sequential move game at least two Pareto-undominated subgame-perfect

stage game equilibria must exist so that any payoff vector $\pi_i^*(s) \geq \pi_i^{\text{SPE}}(s^i)$ can be obtained in the finitely repeated game.

Proof: See Appendix II. QED

Though the proof of Theorem 4.4 is straightforward, it is rather messy. Therefore, we consider only two examples to illustrate the idea of the proof, the second of which has been discussed before. For instance, consider the extended PD game II in Matrix 4.3 and assume simultaneous moves. This game has two stage game NE, that is, $s^{N(1)} = (na_1, na_2)$ and $s^{N(2)} = (p_1, p_2)$, where $\pi_i^{N(1)}(s^{N(1)}) > \pi_i^{N(2)}(s^{N(2)})\ \forall\ i \in \text{I}$ holds. Assume three stages and that it has been proposed to play (a_1, a_2) in the first stage and (na_1, na_2) in the second and third stages. If a country deviates, it will do so in the first stage and is punished afterwards by the stage game NE resulting (which is SPE strategy by Theorem 4.1) from the action combination (p_1, p_2). Thus, compliance pays, provided:

$$3.2 + 2\delta_i + 2\delta_i^2 \geq 4.4 + \delta_i + \delta_i^2 \Rightarrow \delta_i \geq \delta_i^{\min} = 0.704\ \forall\ i \in \text{I} \qquad (4.11)$$

holds. That is, the proposal is an equilibrium if players are not too impatient. We denote $\delta_i^{\min}$ the 'minimum discount factor requirement' to establish cooperation. That is, $\delta_i \geq \delta_i^{\min}\ \forall\ i \in \text{I}$ is a necessary and sufficient condition to establish cooperation as SPE.

Increasing the number of stages at the end of the game in which a player can be punished, $\delta_i^{\min}$ decreases and cooperation becomes easier.[14] Alternatively, looking at it from a different angle, if discount factors are close to 1 and the number of stages T is large, then the number of cooperative stages t^* can be chosen rather large. At the limit for $\text{T} \to \infty$, $\mu = t^*/\text{T}$ approaches 1 from below. That is, the fraction of cooperative stages compared to the total number of stages, μ, is large. Consequently, the average payoff along the equilibrium path approaches the stage game cooperative payoff $\pi_i^* = 3.2$.

The simultaneous move chicken game is another example covered by Theorem 4.4. Here $\pi_i^{N(1)}(s^{N(1)}) > \pi_i^{N(2)}(s^{N(2)})$ for player i and $\pi_j^{N(1)}(s^{N(1)}) < \pi_j^{N(2)}(s^{N(2)})$ for player j. Though in Section 4.2 no discounting was assumed, it should be obvious that the same result could be established with discounting. The only problem in games in which a sequence of alternating stage game NE in the terminal rounds is required is that punishment may be delayed. This leads to high discount factor requirements, though for discount factors close to 1 Theorem 4.4 still holds at the limit, as demonstrated in Appendix II.

Of course, as discovered before, the sequential move version of the

chicken game is not covered by Theorem 4.4 since there is only one stage game SPE. To cover the NE strategy of alternating stage game NE in the terminal rounds as laid out in Section 4.2, Theorem 4.4 must be weakened in that Nash equilibrium strategies are also allowed to be played in the entire game.

For completeness, note finally that there is an even more general result than Theorem 4.4. Benoît and Krishna (1985) show that Theorem 4.4 does not only hold for payoff vectors which give each player more than in his/her worst stage NE but extends to all individually rational payoff vectors. Since the proof requires far more complicated strategy profiles than those used above, it is omitted here.

NOTES

1. According to Rasmusen (1989, p. 46), the only difference between an extensive form representation and a game tree is that in (e) the word 'payoffs' is replaced by 'outcomes'. Thus, for our purposes, we can use both terms more or less interchangeably.
2. Some authors exclude the starting node in the definition above (see, for example, Gibbons 1992, pp. 122ff.). However, this is only a matter of terminology. Definition 4.2 only implies that in a one-shot game an NE could also be called an SPE. Nevertheless, we reserve the term SPE for an equilibrium in a dynamic game.
3. Of course, by the symmetry of a simultaneous move game, also country 2 could move first.
4. If players are not completely informed about the history of the game this is referred to as imperfect information. Consequently, a simultaneous move game is always a game of imperfect information. Moreover, a game of incomplete information is automatically a game of imperfect information, though the opposite is not implied.
5. In a simultaneous move repeated game, the terms 'stages' and 'rounds' are used synonymously. A similar term is 'period'. In sequential move games this is different. Each move is one stage of the game. If after the last move in a *constituent game* (also called a *stage game*), the same game is played anew, we refer to this as the second round or second period in the game. Thus, in a sequential move game there are several stages within a constituent game.
6. In a simultaneous move game $\overline{V} = T$ and $v = 1$ by definition.
7 In the general case of mixed strategies, observability of past moves requires that the random device of each player is common knowledge. This is the standard assumption in the literature, though it lacks an intuitive explanation for why this should be the case. Departing from this assumption, however, complicates the analysis tremendously.
8. For notational convenience, we refrain from indexing σ_i. Of course, a player has as many history-dependent strategies as there are possible histories in a game.
9. For an extensive discussion on discounting see, for example, Mohr (1995), Pearce *et al.* (1989, ch. 6).
10. 'Player *j*' may comprise one but also more players.
11. Of course, a similar theorem can be established for Nash equilibrium strategies in a dynamic game (see, for example, Eichberger 1993, p. 212, Theorem 8.1). However, since every subgame-perfect strategy is also a Nash equilibrium, Theorem 4.1 is a stronger result.
12. If this were not true, then for any history h^T such that $\sigma_i(h^T) \neq s_i^N(\sigma_i(h^T) \neq s_i^{SPE})$, there would be at least one player who could improve his/her payoff in period T without any effect on the payoffs in previous periods. Therefore, we must have $\sigma_i(h^T) = s_i^N(\sigma_i(h^T) =$

s_i^{SPE}) for all players *i*, implying that the stage game strategy in period T will be chosen *history independent*.

13. The trigger strategy is sometimes also called *grim strategy* (see, for example, Dasgupta 1990; Rasmusen 1989, p. 91). It refers to a strategy which calls for punishment until the end of the game once a deviation occurs, regardless of what the deviator plays afterwards.
14. For example, adding a fourth stage to the game, it is easily checked that $\delta_i^{min} \approx 0.61$.

5. Infinite dynamic games with discrete strategy space: a first approach

5.1 INTRODUCTION

In Chapter 4 it was demonstrated that in finitely repeated simultaneous (sequential) move games with a single stage game Nash equilibrium (NE) (subgame-perfect equilibrium (SPE)), only this equilibrium will be played throughout the game. In negative externality games, like the PD game, this implies that cooperation cannot be established. In this section, we start again by considering a PD game and generalize the results subsequently. As pointed out in Section 4.3, we shall assume from now onwards simultaneous moves for simplicity, though in most cases an extension to cover the case of sequential moves is straightforward. In contrast to a finite setting, we cannot use backwards induction to solve the game. Though this may seem a disadvantage from an analytical point of view, it is this very fact which is responsible for obtaining more 'optimistic' results in an infinite time horizon. Since the end of the game is not known with certainty, a player must always reckon with a costly and long-lasting punishment if s/he takes a free-ride. This might enforce compliance (Holler and Illing 1996, p. 22; Rasmusen 1995, p. 123; Taylor 1987).

Generally, there are two interpretations of infinite games: either, as the name suggests, the games last until perpetuity, or the end of the game is not known with certainty. The latter interpretation implies that there is a probability – though this might be very small – that the game continues (Gibbons 1992, p. 90; Osborne and Rubinstein 1994, p. 135). Of course, in the context of IEAs players have finite lives and therefore the first interpretation of supergames does not apply (Mäler 1994, pp. 360ff.). However, most IEAs do not specify the termination of the contract. Moreover, it is common practice for successor governments to recognize obligations signed by their predecessors and therefore the second interpretation (approximately) applies. Thus, a *supergame* may be a good approximation of the *continuing relationship between countries* (Barrett 1990).

To accommodate the interpretations above, we have to change the definition of the discount factor slightly. It now comprises, apart from the

discount rate ρ_i, also the *probability that the game continues*, p, $0 \leq p \leq 1$, so that $\delta_i = p/(1+\rho_i)$. Thus, a low probability that the game continues shows up in a small discount factor. Since for $\delta_i = 1$ the average payoff may not be defined, we assume in what follows discount factors strictly smaller than 1, though they may nevertheless be very close to 1.[1] To compute the net present value of an infinite payoff stream, we can use formulae (4.2)–(4.4), replacing T by ∞ (Gibbons 1992, pp. 90, 97). For convenience, the formulae are reproduced below (subscript *i*, denoting the player, has been omitted):

$$\pi^{\delta} = \pi_1 + \delta\pi_2 + \ldots + \delta^{\infty}\pi_{\infty} = \sum_{t=0}^{\infty} \delta^t \pi_t; \frac{\pi}{1-\delta} \text{ if } \pi_1 = \pi_2 = \ldots \tag{5.1}$$

$$\pi^{\delta} = \delta\pi_1 + \delta^2\pi_2 + \ldots + \delta^{\infty}\pi_T = \sum_{t=1}^{\infty} \delta^t \pi_t; \frac{\delta\pi}{1-\delta} \text{ if } \pi_1 = \pi_2 = \ldots \tag{5.2}$$

$$\bar{\pi}^{\delta} = (1-\delta) \cdot \sum_{t=0}^{\infty} \delta^t \pi_t \text{ if } \delta < 1; \bar{\pi}^{\delta} = \pi \text{ if } \pi_1 = \pi_2 = \ldots. \tag{5.3}$$

Now, if we want to check whether cooperation is possible in a PD supergame, we have to think about an adequate strategy profile which could do the job. Since we have already encountered the simple trigger strategy in Chapter 4, we may try whether this strategy can be used in the present context as well (for example, Osborne and Rubinstein 1994, pp. 143ff.). The only difference in comparison to the trigger strategy used to prove Theorem 4.4 is that now there is only one stage game NE and that defection triggers the infinite play of the stage game NE. Formally, we have:

$$\sigma_i = \begin{cases} s_i^j \text{ if in any } h^t(0) \ldots h^t(t-1) = (s_j', s_{-j}); h^t(0) \ldots h^t(t-1) = (s_j^j, s_{-j}^j) \\ \quad h^t(0) \ldots h^t(t-1) = (s_j'', s_{-j}^j) \text{ or } h^t(0) \ldots h^t(t-1) = (s_k'', s_j^j, s_{-k}^j) \; \forall \; t = 0 \ldots \infty \\ s_i \text{ in } t = 0 \\ s_i \text{ if in any } h^t(0) \ldots h^t(t-1) = (s_i, s_{-i}) \; \forall \; t = 0 \ldots t-1 \end{cases} \tag{5.4}$$

where s_i is the agreed 'cooperative' stage game strategy, s_i^j is the punishment strategy against player j where in the PD game $s_i^j = s_i^N$. Moreover, $s_j' \neq s_j$, $s_j'' \neq s_j^j$ and $s_k'' \neq s_k^j$ are some arbitrary strategies of player j or k respectively. $h^t(0) \ldots h^t(t-1) = (s_j'', s_{-j}^j)$ and $h^t(0) \ldots h^t(t-1) = (s_k'', s_j^j, s_{-k}^j)$ are only mentioned for completeness but are unlikely events because the best reply to $s_{-j}^j = s_{-j}^N$ is $s_j^j = s_j^N$ by definition (see Appendix II). Hence, deviation during the punishment pays neither for the punishers nor for the punished player. Moreover, from Theorem 4.1 it is known that to play a stage game NE in every period throughout a game is an SPE. Hence, it remains to be shown that deviation from the cooperative strategy does not pay either and therefore $\sigma = (\sigma_i, \sigma_{-i})$ is an SPE, that is, $\sigma = \sigma^{SPE}$. Hence:

$$\frac{\pi_i^*(s)}{1-\delta_i} \geq \pi_i^D(s_i', s_{-i}) + \frac{\delta_i \pi_i^N(s^N)}{1-\delta_i} \Leftrightarrow \pi_i^* \geq (1-\delta_i)\pi_i^D + \delta_i \pi_i^N$$

$$\Leftrightarrow \delta_i \geq \delta_i^{\min} = \frac{\pi_i^D - \pi_i^*}{\pi_i^D - \pi_i^N} \; \forall \; i \in I \qquad (5.5)$$

must hold. The first inequality is expressed as net present values, the second inequality in terms of average payoffs. This is to stress the equivalence of both approaches. The first inequality implies that the net present value from complying (LHS term of the inequality sign) must be at least as high as when taking a free-ride (first term on the RHS of the inequality sign) and then being punished for the rest of the game (second term on the RHS of the inequality sign).

Thus, for given discount factors δ_i, and $\delta_i^{\min} \leq \delta_i < 1$, $\pi_i^* \geq \pi_i^N$ must hold for all i, so that stability can be guaranteed using the trigger strategy above. Since in the basic PD game playing the stage game NE for the rest of the game is the harshest punishment players can inflict on a free-rider, that is, $\pi_i^N = \pi_i^M$, there is no other strategy which could deliver a lower $\delta_i^{\min}$. Therefore, $\delta_i \geq \delta_i^{\min} \; \forall \; i \in I$ may be regarded as a *necessary and sufficient* condition for stability.

For instance, if we assume payoffs of the PD game as given in Matrix 3.1, and that mutual investment shall be sustained, then $\pi_i^D = 4.4$, $\pi_i^* = 3.2$, $\pi_i^N = 2$ and, consequently, $\delta_i^{\min} = 0.5$. Of course, other payoff vectors (for example, resulting from playing mixed strategies within a stage or from alternating between pure strategies in different stages) could also be sustained as long as they give each player at least an average payoff of 2.

From (5.5) it is evident that the larger the free-rider incentive $\pi_i^D - \pi_i^*$, the more difficult it is to ensure stability. In contrast, a harsh punishment in the form of a low value of π_i^N is conducive to cooperation ($\delta_i^{\min}$ will be lower).

The 'actual' discount factor of players itself has also a great influence on the stability of an agreement. In a politico-economic context, one would expect that, if politicians seek short-term success, the value of ρ_i would be high and therefore stability may be jeopardized. Though it is common practice to assume that the discount factor remains constant over time, an assumption we shall follow throughout this book, it is nevertheless conceivable that short-term success is particularly important before elections (Hahn 1989).[2] Hence, long-term commitments may be particularly jeopardized during election campaigns.

Moreover, though one may think of p as an objective probability, it may also be a subjective estimation by agents about the uncertainty of future events. In the latter case, p could be indexed as well and p_i may reflect a bundle of factors, for example, the general risk attitude of politicians; the

evaluation of political stability in neighboring countries; the expectations of politicians about the chances that the partners with whom they strike a deal will be re-elected, and so on. Thus, for instance, civil war, social unrest or any kind of political instability in country j would imply the representatives in country i putting a low value on p_i. Consequently, δ_i will be rather low, and this is a source of instability. In contrast, regular meetings of politicians of different states may increase the mutual confidence of partners that the game will continue and therefore will improve upon the stability of treaties.

A further interpretation of the discount factor arises by noting that a payoff π received at time $t=1$ is worth $\delta\pi$ at time $t=0$ in discrete time, whereas if it is continuously discounted it is worth $\pi e^{-\rho\Delta}$ where Δ is the length of the time span (Bac 1996). Thus δ in discrete time is equivalent to $e^{-\rho\Delta}$ in continuous time. By reducing Δ, $e^{-\rho\Delta}$ becomes bigger. Thus, the shorter the time span of a stage, the higher the (equivalent) value δ, and hence the higher the chances of cooperation. This suggests that short discrete time intervals are particularly important for the stability of an IEA if time is discounted very much. A short time span has two advantages: first, punishment has a higher threat potential since action can be taken immediately: in a wider context this implies that a short time span Δ reflects the possibility of immediate discovery of free-riding; second, transitory gains from free-riding are smaller.

5.2 FOLK THEOREMS

In this section, we want to put the preliminary results of Section 5.1 on a more general footing.[3] Though we interpreted the strategy profile in (5.4) in terms of the basic PD game, we have already used a general notation in the previous section. Therefore, the first folk theorem[4] which is due to Friedman (1971) can be stated immediately:

Theorem 5.1: Folk Theorem I
Let Γ^∞ be an infinitely repeated stage game Γ with a stage game Nash equilibrium s^N and an associated payoff vector π^N. Then each payoff vector π^* of the one-shot game Γ for which $\pi_i^* \geq \pi_i^N \; \forall \; i \in \mathrm{I}$ is true can be sustained in the infinitely repeated game Γ^∞ by playing subgame-perfect strategies provided discount factors of the players are sufficiently close to 1 (that is, $\delta_i \geq \delta_i^{\min} \; \forall \; i$, where $\delta_i^{\min} \in [0, 1]$, is a critical value) so that each player receives an average payoff of π_i^*.

Proof: From (5.5) it is evident that if π_i^* approaches π_i^N from above, $\delta_i^{\min}$ approaches 1 from below. Thus, provided δ_i is sufficiently close to 1,

$\delta_i \geq \delta_i^{\min}$ is possible. Since the cooperative strategy vector s delivers π_i^* in each period, the average payoff is π_i^* (see (5.3)). QED

It should be evident that Theorem 5.1 holds irrespective of the number of stage game NE in a game. The minimum requirement is only that each player receives at least his/her worst NE payoff when playing the cooperative strategy. For example, in the extended PD game II in Matrix 4.3, each player must receive at least a payoff of 1.

The following Folk Theorem II has been known in the literature as the classical 'folk theorem' since the 1950s. The theorem derives its name from the fact that it was well-known among game theorists for a long time before there was published a reference. In contrast to Folk Theorem I, there are two modifications: first, with respect to payoffs, it comprises a larger set of payoffs by extending the payoff space to include all individually rational payoff vectors; second, with respect to the equilibrium concept, the theorem relies only on Nash equilibrium strategies (see, for example, Fudenberg and Tirole 1996, pp. 154ff.).

Theorem 5.2: Folk Theorem II
Any individually rational payoff vector of the stage game Γ, $\pi^* \in \Pi^{IR}$, can be sustained in the infinitely repeated game Γ^∞ by playing Nash equilibrium strategies where each player receives an average payoff of π_i^* provided that discount factors of all players are sufficiently high.

Proof: The proof of Folk Theorem II is very much in the spirit of the proof of Folk Theorem I. It also uses a trigger strategy. Along the equilibrium path each player plays the stage game strategy s_i which generates the payoff vector $\pi_i^*(s_i, s_{-i}) \geq \pi_i^M(m_i^i, m_{-i}^i)\ \forall\ i \in I$. Once a player i deviates – either in the cooperative or the punishment phase – that player is minimaxed by all players for ever. Thus, by contrast with Theorem I where $s_i^j = s_i^N$, now $s_i^j = m_i^j\ \forall\ i, j, k \in I$ (for example, $s_{-k}^j = m_{-k}^j$ and so on) in (5.4), and in (5.5) π_i^N has to be replaced by π_i^M and the same arguments to prove Theorem 5.1 apply. QED[5]

Since Folk Theorem II comprises a larger payoff set than Folk Theorem I but uses a weaker equilibrium concept, it comes as no surprise that scholars in game theory tried to establish a folk theorem which combines elements of both theorems. As a first step, Aumann and Shapley (1976) and Rubinstein (1976) showed successfully that each individually rational payoff vector can be sustained by playing subgame-perfect strategies. However, their proof is only valid for games without discounting.

Theorem 5.3: Folk Theorem III
Any rational payoff vector $\pi^* \in \Pi^{IR}$ of the stage game Γ can be obtained in the infinitely repeated game Γ^∞ – giving each player an average payoff of π_i^* – using subgame-perfect strategies if there is no discounting.

Proof: Two basic types of game have to be distinguished. In the first type each player minimaxing player i derives a higher payoff than when s/he is minimaxed him- or herself, that is, $\pi_j^{M(i)} > \pi_j^{M(j)}$. Once defection occurs, either during cooperation or punishment, then the punishment involves the infinite punishment. Since $\pi_i^* \geq \pi_j^{M(i)}$ deviation does not pay during cooperation. Because of the prospect of being punished forever, no player deviates when punishing a player i, and by the definition of a minimax payoff the punished player i does not deviate either. Thus the punishment is an NE of the subgame beginning at time $t = m + 1$ when the deviation occurred at time $t = m$.

In the second type of game a punisher receives less than when s/he is minimaxed him- or herself, that is, $\pi_j^{M(i)} \leq \pi_j^{M(j)}$, and an infinite punishment is *not* possible according to the notion of subgame-perfection. Now, if player i deviates, the other players minimax him/her only for some time until the gains from the deviation are wiped out. Then, all players return to the 'cooperative phase'. If player j among the players minimaxing player i deviates from the punishment, then s/he is minimaxed by all the others for an even longer time span. This hierarchy of successively higher-order punishments is used for any further deviation (which is not *not* simple, in the sense of Abreu). The existence of the cooperative phase at the end of the punishment makes the strategy credible in the sense of subgame-perfection. If a player j deviates from conducting the punishment of player i s/he receives a deviation payoff π_i^D but is then punished by the others for t_j^P periods. Since there is always a $t_j^P > t_i^P$ in the infinite time horizon for which:

$$t_i^P \pi_j^{M(i)} + \sum_{t=t_i^P+1}^{\infty} \pi_{jt}^* \geq \pi_j^D + t_j^P \pi_j^{M(j)} + \sum_{t=t_j^P+1}^{\infty} \pi_{jt}^* \tag{5.6}$$

is true, the punisher j will comply with his/her punishment obligations. QED

An example of the first type of game is the extended PD game I in Matrix 4.2. The extended PD game II in Matrix 4.3 belongs to the second category of game, though it is a 'knife-edge' case because the minimax payoff and the payoff when minimaxing other players coincide for all players.

Though these hierarchical punishment strategies work in the case of $\delta_i = 1$, one may not be able to use them in a game with discounting. The problem

can arise in games of the second type where the punisher receives less than his/her minimax payoff. Then, if a player deviates from conducting the punishment, the threat to punish him/her for some time might not be sufficient to deter deviation. Due to discounting, punishment payoffs in the future become less of a deterrent and the punishment hierarchy may stop working at some higher order. Put differently, with discounting, arbitrarily long punishments are not arbitrarily severe since far-off punishments are relatively unimportant (see Fudenberg and Maskin 1986 for details).

It took exactly ten years for Fudenberg and Maskin (F/M) to extend Folk Theorem III in 1986 to games with discounting. The proof is particularly straightforward in the two-player case. In the general case of N players, slightly more complicated strategy profiles have to be used and some restrictions on the payoff structure have to be imposed. In the following we proceed stepwise: first, we prove F/M's folk theorem for two players (Folk Theorem IV) and then, second, we cover the general case of N players in Folk Theorem V.

Theorem 5.4: Folk Theorem IV (Fudenberg and Maskin 1986, Theorem 1)
In a two-player game any individually rational payoff vector $\pi^* \in \Pi^{IR}$ of Γ can be sustained in the infinitely repeated game Γ^∞ by playing subgame-perfect strategies provided the discount factors of all players are sufficiently large so that each player receives an average payoff of π_i^*.

Proof: Consider the following strategy:

$$\sigma_i = \begin{cases} m_i^j \text{ if in any } h^t(t-t^P)...h^t(t-1)=(s_i, s_j'), \text{ or } h^t(t-t^P)...h^t(t-1)=(m_i^j, s_j'') \\ s_i \text{ if in any } h^t(t-t^P)...h^t(t-1)=(s_i, s_j), \text{ or } h^t(t-t^P)...h^t(t-1)=(m_i^j, m_j^i) \end{cases} \tag{5.7}$$

where $h^t(t-t^P) \ldots h^t(t-1)$ means all t^P periods previous to period t, $t_1^P = t_2^P = t^P$. The strategies imply that the game may be divided into a cooperative and a punishment phase. In the case of compliance, the cooperative phase continues and players choose strategy s_i. The minimax strategy m_i^j is played if there is either a deviation in the cooperative phase of the game or a deviation in the punishment phase. Thus, the strategy implies that players 'agree' to *minimax each other mutually* for t^P periods in the case of a deviation where $\pi_i^G(m_i^j, m_j^i) \leq \pi_i^M(m_i^j, m_j^i) \forall\, i \in I$ holds. Thus, the severity and length of the punishment is independent of the degree of deviation and of the player who deviates. That is, the strategies are simple in the sense of Abreu. For an SPE three basic conditions have to be satisfied which are reflected in (5.8)–(5.10).

Subgame-perfect equilibrium conditions in a two-player game:

$$\pi_i^* \geq (1-\delta_i)\pi_i^{\mathrm{D}} + \delta_i \pi_i^{\mathrm{P}} \tag{5.8}$$

$$\pi_i^{\mathrm{P}} \geq \pi_i^{\mathrm{M}} \tag{5.9}$$

$\forall\, i \in \mathrm{I}$ and $\pi^* \in \Pi^{\mathrm{IR}}$ where:

$$\pi_i^{\mathrm{P}} = (1-\delta_i^{t^{\mathrm{P}}})\pi_i^{\mathrm{G}} + \delta_i^{t^{\mathrm{P}}}\pi_i^*. \tag{5.10}$$

1. Deviation in the cooperative phase does not pay, due to subsequent punishment. Therefore, in (5.8) the average payoff in the cooperative phase, π_i^*, must exceed the average payoff received in the deviation period, $(1-\delta_i)\pi_i^{\mathrm{D}}$, and the payoffs in the subsequent periods of the punishment phase, $\delta_i \pi_i^{\mathrm{P}}$.
2. If a player gets punishment s/he must have an incentive to go along with the punishment. That is, this player 'cooperates' during the punishment in order to return to the cooperative phase after some time (because otherwise punishment is continued). Since each player can secure a minimax payoff, the *average continuation payoff* in the punishment phase, π_i^{P}, must exceed π_i^{M} (see (5.9)).
3. The punisher him- or herself must have an incentive to conduct the punishment, and therefore the punishment strategy must constitute a best reply for him or her. Due to the simple punishment strategy of mutual minimaxing, this condition is also expressed in (5.9). The punisher conducts the punishment, otherwise the punishment phase will be prolonged and s/he receives π_i^{M} instead of π_i^{P}.[6]

Now we have to show that for sufficiently high values of the discount factor, that is, $1 > \delta_i \geq \delta_i^{\min}$, the strategy σ_i *always* satisfies conditions (5.8) and (5.9), as long as $\pi_i^* \geq \pi_i^{\mathrm{M}}\ \forall\, i \in \mathrm{I}$. This is done as follows.

First, substitute (5.10) into (5.8) and rearrange terms to have:

$$\delta_i^{t^{\mathrm{P}}} \geq \frac{\pi_i^{\mathrm{M}} - \pi_i^{\mathrm{G}}}{\pi_i^* - \pi_i^{\mathrm{G}}} \Rightarrow \mathrm{C}_1 : \delta_i \geq \frac{\pi_i^{\mathrm{M}} - \pi_i^{\mathrm{G}}}{\pi_i^* - \pi_i^{\mathrm{G}}} = \delta_i^{1} \tag{5.11}$$

where, due to $\pi_i^* > \pi_i^{\mathrm{M}}$, the RHS term of each inequality is smaller than 1 and C_1 is a necessary condition for an SPE since $\delta_i < 1$, $t^{\mathrm{P}} \geq 1$, and therefore $\delta_i \geq \delta_i^t$.

From (5.9) it follows that π_i^{P} should be as small as possible which, considering (5.10), is equivalent to minimizing $\delta_i^{t^{\mathrm{P}}}$. Thus, if we choose the punishment time t^{P} such that:

$$t^{\mathrm{P}} = t^{*} = \frac{\log\left(\dfrac{\pi_i^{\mathrm{M}} - \pi_i^{\mathrm{G}}}{\pi_i^{*} - \pi_i^{\mathrm{G}}}\right)}{\log \delta_i}$$

then $\pi_i^{\mathrm{P}} = \pi_i^{\mathrm{M}}$ and (5.9) becomes binding.[7] Upon substitution, (5.8) reads:

$$\pi_i^{*} \geq (1-\delta_i)\pi_i^{\mathrm{D}} + \delta_i \pi_i^{\mathrm{M}} \Rightarrow \mathrm{C}_2\colon \delta_i \geq \frac{\pi_i^{\mathrm{D}} - \pi_i^{*}}{\pi_i^{\mathrm{D}} - \pi_i^{\mathrm{M}}} = \delta_i^2 \tag{5.12}$$

where $\delta_i^2 \leq 1$ due to $\pi_i^{*} > \pi_i^{\mathrm{M}}$. Taken together, C_1 and C_2 are *necessary conditions* for the discount factor so that the strategy in (5.7) is an SPE. Hence, these conditions provide the starting values of $\delta_i^{\min}$ since $\delta_i^{\min} \geq \max(\delta_i^1, \delta_i^2)$ holds. The actual $\delta_i^{\min}$ is determined as follows.

Since $t^{\mathrm{P}} \geq 1$, we may start by assuming $t^{\mathrm{P}} = 1$, choose $\delta_i = \max(\delta_i^1, \delta_i^2)$ (see C_1 and C_2 above) and insert this in (5.8). If (5.8) is satisfied, then we have finished and $\delta_i^{\min} = \max(\delta_i^1, \delta_i^2)$. However, if (5.8) fails to hold, that is, π_i^{P} is too big, consider raising t^{P}. Thus, eventually, (5.8) will be satisfied. However, this may imply that $\pi_i^{\mathrm{P}} < \pi_i^{\mathrm{M}}$ and hence (5.9) is no longer satisfied. Thus, we have to raise δ_i so that (5.9) can still be satisfied. Eventually that δ_i which makes (5.9) binding is $\delta_i^{\min}$. Since for large δ_i, π_i^{P} declines almost continuously with increasing t^{P}, the minimum discount factor we arrive at by this procedure comes close to the 'true' $\delta_i^{\min}$. Thus if $\delta_i \geq \delta_i^{\min}$, then (5.8) and (5.9) can be satisfied. Consequently, since it is always possible to find such a $\delta_i^{\min}$ (probably very close to 1), all payoffs in Π^{IR} can be sustained as SPE provided the discount factors of all players are close to 1. QED

An example will clarify the determination of $\delta_i^{\min}$. Suppose payoffs as in the extended PD game I in Matrix 4.3 and that both countries agree to play the cooperative strategy a_1/a_2. Hence, $\pi_i^{*} = 3.2$, $\pi_i^{\mathrm{D}} = 4.4$, $\pi_i^{\mathrm{G}} = 0$ and $\pi_i^{\mathrm{M}} = 1$. Inserting this in (5.11) and (5.12) gives $\delta_i^1 = 0.3125$ and $\delta_i^2 = 0.2$ and therefore $\delta_i^{\min} = \max(\delta_i^1, \delta_i^2) = 0.3215$ is the starting value. Inserting this value in (5.8) gives $3.2 \geq 3.3375$, which is obviously not true. Therefore, we choose $\delta_i = \delta_i^{\min} = 0.375$ so that (5.8) is satisfied. A routine check reveals that choosing alternatively $t^{\mathrm{P}} = 2$ leads to $\delta_i^{\min} = 0.56$ and this alternative can therefore be discarded.

From (5.8) the following corollary is readily proved:

Corollary 5.1

For $\delta_i \to 0$ the payoff space which can be sustained by subgame-perfect strategies converges to the Nash equilibrium payoff space of the stage game Γ, that is, $\Pi^{\mathrm{SPE}} = (\pi^{N(1)}, \ldots)$.

Proof: For $\delta_i \to 0$, (5.8) becomes $\pi_i^*(s_i, s_j) \geq \pi_i^D(s_i, (s_j), s_j)$, which can only be true if $\pi_i^*(s_i, s_j) = \pi_i^N(s_i^N, s_j^N)$. $\pi^{N(1)}$, ... refers to the general case that there might be more than one NE. QED

Let us now consider the case of more than two players. The proof of Folk Theorem IV was based on the punishment strategy of mutual minimaxing. With this strategy we killed two birds with one stone: on the one hand, the deviator was punished; on the other hand, the punisher had an incentive to conduct the punishment. For both players the incentive to go along with the punishment is the prospect of returning to cooperation as soon as possible. Now, in the case of more than two players, mutual minimaxing might not be possible any more. That is, some punishers might receive more than their minimax payoff during the punishment phase if they deviate from punishing. To render this a non-beneficial option, a three-phase strategy is needed which may be summarized as follows.

Players cooperate in phase 1 if no player deviates. If a player deviates, phase 2 starts in which the deviator is punished by being minimaxed for some time. The deviator must be punished long enough for the gains from the deviation to be wiped out. If a player deviates in phase 2, this phase is started anew. If there is no deviation in phase 2, phase 3 starts in which the punishers are rewarded who went through with the punishment. In this phase the punishers of phase 2 receive a payoff which by construction exceeds their minimax payoff.

This three-phase strategy allows us to state the following Folk Theorem V, which is proved in Appendix III:

Theorem 5.5: Folk Theorem V (Fudenberg and Maskin 1986, Theorem 2)
In an N-player game any individually rational payoff vector π^* of the stage game Γ can be obtained as an average payoff vector in the infinitely repeated game Γ^∞ using subgame-perfect strategies, provided the discount factor of each player is sufficiently close to 1 and the payoff space is of full dimensionality (that is, of dimensionality N).

Proof: See Appendix III. QED[8]

5.3 DISCUSSION

In contrast to finite games, cooperation is less of a problem in supergames. In fact, the set of SPE is very large and one could even talk of an abundance of equilibria, which makes predictions about an outcome difficult.

Basically, there are two approaches to tackling this problem. First, one thinks of whether it is possible to develop the definition of an equilibrium concept further, for example, by defining more restrictively what is meant by a credible threat and punishment strategy. This is the route we follow in Chapters 6 and 7. Second, a bargaining stage is added to the game in which the negotiation of players to settle for an equilibrium is modeled. Moreover, in richer games the problem itself may suggest some restrictions on the set of possible equilibria. For instance, in a second-best world institutional restriction may exclude some equilibria. We shall encounter such examples in Chapters 11–14.

We wish to end this chapter with two remarks. First, as noted already in Chapter 4, when mixed strategies are played, it has to be assumed that the mixing device is known to all players, otherwise incomplete information must be modeled explicitly, which is more complicated. Second, in the case of two players the mutual minimaxing strategy, though it is a feasible subgame-perfect strategy from a game theoretical point of view, seems a rather curious kind of punishment when thinking of applications. Basically, it implies that if one player free-rides and the other player does not conduct the agreed punishment, s/he in turn is punished by the free-rider. It seems rather paradoxical to expect a player to punish someone if s/he does not get punished him- or herself. Therefore, the three-phase strategy in the context of more than two players appears to be more convincing. Players agree on a penalty code according to which each party is required to participate in the punishment. Since enforcement itself is a public good (since punishment is costly), potential punishers have an incentive to ensure that all parties fulfill their punishment obligations. This is done by establishing a two-stage hierarchy of punishments.

However, the difficulties in international politics of reaching stable agreements suggests that it is not always that easy to establish an incentive-compatible punishment code. The failure of many trade embargoes may serve as an example. The available second-order punishments are obviously often not severe enough to overcome the incentive of some potential punishers to trade with the country on which a boycott is imposed and therefore a successful punishment code (including first-order punishments) cannot be established. In particular, receiving only slightly more than the minimax payoff may not be a strong enough incentive for players to penalize a free-rider. This concern is taken up in the next two chapters.

NOTES

1. This assumption does not affect the generality of the subsequent proofs. As will become apparent, if a payoff can be backed by equilibrium strategies for discount factors close to 1, this will be even more true for a discount factor equal to 1.
2. This idea is elaborated in various models of political business cycles. For a survey see, for example, Willet (1988) and the literature cited therein.
3. A survey of the issues treated in this section can be found in Pearce (1992) and Sabourian (1989).
4. For an explanation of the term 'folk theorem', see below.
5. Of course, though m_i^i is a best reply to m^i_{-i}, and hence a punished player will 'comply with his/her punishment obligations', the punishers could do better than playing m^i_{-i} and hence this strategy is not an SPE.
6. (5.10) states that the average continuation punishment payoff comprises the mutual minimaxing payoff received for t^P periods and the cooperative phase payoff which is received afterwards if there are no deviations during the punishment phase. Note that π_i^P decreases in the punishment time t^P because π_i^G receives a higher weight compared to π_i^*.
7. Of course, $t^P = t^*$ is only a first approximation since t^P must be an integer value.
8. The assumption of full dimensionality requires that all payoff vectors in a game are independent so that any feasible payoff combination in the N-dimensional payoff space can be generated. If this does not hold, Π^{SPE} may only be a subset of Π^{IR}. This assumption is needed in phase 3, so that all punishers receive at least slightly more than their minimax payoffs, but at the same time the punished player is held down to his/her minimax payoff.

6. Finite dynamic games with discrete strategy space: a second approach

6.1 INTRODUCTION

In Section 4.4 we demonstrated that in games with two or more stage game Nash equilibria (NE) all payoff vectors can be sustained by subgame-perfect equilibrium strategies which give each player more than in his/her worst NE provided discount factors are close to 1. Such an abundance of equilibria was also found in supergames where even weaker conditions must be satisfied to derive folk theorem type of results. Thus, although we strengthened the equilibrium concept for dynamic games by requiring strategies not only to be an NE but also to be a subgame-perfect equilibrium (SPE), the set of equilibrium payoffs remains large.

A concept which is capable of reducing (though not eliminating) this *lack of predictability* in repeated games is the concept of *renegotiation-proofness*. Though there emerged many versions of this concept, in the context of finite games the interpretation seems not very controversial (Benoît and Krishna 1993; Bergin and MacLeod 1993; Bernheim *et al.* 1987; Fudenberg and Tirole 1996, pp. 174ff.). The subsequent discussion is based on Benoît and Krishna's definition which is restricted to two-player games. In this case their concept coincides with Bernheim *et al.*'s definition of coalition-proof equilibria, which we discuss in Chapter 15.

In the above-cited literature it is argued that threats which imply a lower payoff to deviators and punishers *alike* will be subject to renegotiations and therefore lose their credibility. If defection occurs, it is in the interest of all players to treat bygones as bygones, and punishment will not be carried out. Three examples will illustrate the idea. In each of them no discounting is assumed and only pure stage game strategies are considered.

6.1.1 Extended PD Game II

Suppose that the extended PD game II in Matrix 4.3 is played three times and assume no discounting. Both countries agree to cooperate in the first round by playing (a_1, a_2) and subsequently playing the good equilibrium

(na_1, na_2) twice. Defection in the first round will be punished by playing the bad equilibrium (p_1, p_2) until the end of the game. Since $3.2+2+2>4.4+1+1$ this strategy is subgame-perfect. However, suppose a country deviates in the first round; then, before period 2, the deviator will suggest that the punisher forget about the past and play the good NE in the last two rounds. Obviously, since there is no reward phase, it is in the punisher's interest to accept the proposal. Since a 'potential' free-rider knows this before round 1, s/he will not be afraid of the punishment and will take a free-ride. Therefore, the only renegotiation-proof equilibrium, henceforth abbreviated RPE, in this game will be the thrice-repeated play of the good stage game NE (na_1, na_2). (The bad NE (p_1, p_2) is no RPE because it is Pareto-inferior to (na_1, na_2).)

Since this result holds also if the game is repeated more than three periods, it also holds for the limited case of $T\rightarrow\infty$. Hence, the set of average equilibrium payoffs shrinks from $\Pi^{IR}=\overline{\Pi}^{SPE}$ (which includes all payoff tuples which give each player at least a payoff of 1 in the example) to the single payoff tuple $\overline{\Pi}^{RPE}=\bar{\pi}^{RPE}=(2, 2)$ if we require equilibrium strategies to be renegotiation-proof and not only subgame-perfect. Hence, $\overline{\Pi}^{RPE}\subseteq\overline{\Pi}^{SPE}$ or in the more general case $\overline{\Pi}^{\delta,RPE}\subseteq\overline{\Pi}^{\delta,SPE}$ where $\delta_i<1$ and T is not necessarily large. Though in the following we concentrate mainly on equilibrium payoffs, it should always be kept in mind that the above relation has its analogy with respect to strategies, that is, $\Sigma^{RPE}\subseteq\Sigma^{SPE}$.

From the example it is evident that though renegotiation-proofness requires equilibrium strategies to be *subgame-perfect and* to be *Pareto-efficient*, renegotiation-proof equilibria, also known as *Pareto-perfect equilibria* may *not* be efficient with respect to the set of all SPE (Fudenberg and Tirole 1996, p. 177).

6.1.2 Extended PD Game III

Next consider the extended PD game III in Matrix 6.1 where the associated payoff space is displayed in Figure 6.1. This game has three-stage game NE. Here p_i^j denotes a punishment strategy of player i to punish player j. The best reply to this has player j choose p_j^j. For this game it is easily checked that the minimax payoffs in this game are $\pi_i^M=2\ \forall\ i\in I$ which correspond to the worst NE payoff to each player. Hence, $\overline{\Pi}^{SPE}=\{\pi_i\in\overline{\Pi}^{SPE}\mid\pi_i\geq 2\ \forall\ i\in I\}$. What about the renegotiation-proof payoff space $\overline{\Pi}^{RPE}$?

In order to determine the equilibrium path in this game, we start from $T=1$ and successively increase the number of stages. If $T=1$, only an NE can be played. Since none of the three NE Pareto-dominates the other, any of them may be played. Now if $T=2$, the following SPE payoff tuples can be obtained:

Matrix 6.1 Extended PD game III

	a_2	na_2	p_2^1	p_2^2
a_1	6 6	1 7	0 0	0 0
na_1	7 1	**3** **3**	0 0	0 0
p_1^2	0 0	0 0	0 0	**4** **2**
p_1^1	0 0	0 0	**2** **4**	0 0

Matrix 6.2 Extended PD game IV

	a_2	na_2	p_2^1	p_2^2
a_1	6 6	1 7	0 0	0 0
na_1	7 1	0.5 0.5	0 3	0 0
p_1^2	0 0	3 0	0 0	**4** **2**
p_1^1	0 0	0 0	**2** **4**	0 0

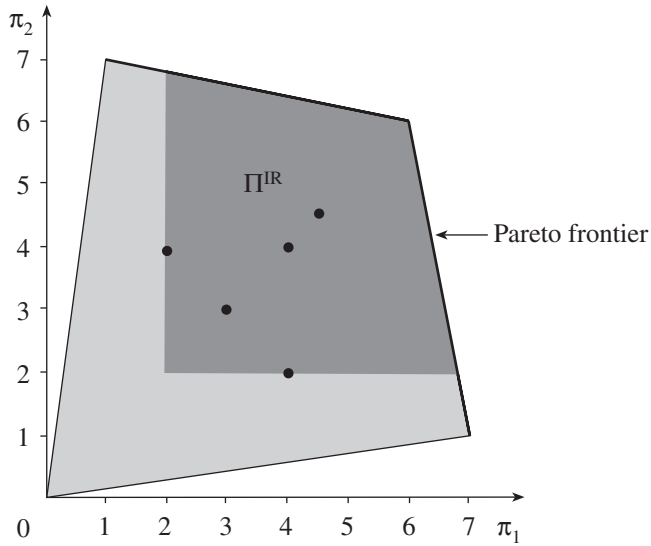


Figure 6.1 Payoff space of the extended PD games III and IV

$$\Pi^{SPE}(T=2)\text{: } (4, 8), (8, 4), (5, 7), (7, 5), (6, 6), (9, 9)$$

where payoffs are summed over the two stages and Π^{SPE} stands for the set of (total) SPE payoffs. The first two tuples result from payoffs (2, 4), (2, 4) or (4, 2), (4, 2); the second two from (3, 3), (2, 4) and (3, 3), (4, 2); the fifth from (3, 3), (3, 3); and the sixth from (6, 6), (3, 3). Whereas the first five payoffs result from the play of stage game NE, and therefore the underlying strategies are automatically subgame-perfect, the last payoff vectors can be obtained by using the threat strategy to play the NE (p_i^i, p_j^i) against deviator i. Since a deviator can net a gain of 1 and loses 1 through punishment, deviation does not pay (assuming no discounting).

Requiring the SPE to be efficient, leads to:

$$\text{Eff}(\Pi^{SPE}(T=2))\text{: } (9, 9)$$

where Eff stands for *weakly* efficient.

Next consider $T=3$. Since it is known that in the last two periods (a_1, a_2) and (na_1, na_2) is the only efficient SPE path, free-riding cannot be punished in the first period and, hence, one of the stage game NE must be played. Hence, for $T=3$, at $t=1$ a strategy combination leading to payoff vectors in $\text{Eff}(\Pi^{SPE}(T=1))$ must be played and $\text{Eff}(\Pi^{SPE}(T=3)) = \text{Eff}(\Pi^{SPE}(T=1)) + (9, 9)$. Thus, we have:

$$\text{Eff}(\Pi^{\text{SPE}}(\text{T}=3)): (11, 13), (12, 12), (13, 11).$$

Continuing with this kind of reasoning, it is evident that in longer-lasting games, the same sequence $((a_1, a_2), (na_1, na_2))$ is played repeatedly. If T is odd, then this sequence is played T/2 times; if T is uneven, then this sequence is repeated $(\text{T}-1)/2$ times and in the first round one of the three NE is played. Consequently, for large T the average payoff to each player approaches $\bar{\pi}_i^{\text{RPE}} = 4.5$. Thus, as in the extended PD game II, the renegotiation-proof space shrinks to a single point, though the subgame-perfect payoff space is large. Also note that, again, the renegotiation-proof payoff vector is *not efficient* and lies below the Pareto frontier.

6.1.3 Extended PD Game IV

Finally, consider the extended PD game IV in Matrix 6.2. Compared to the extended PD game III, it has only two (asymmetric) Pareto-undominated Nash stage equilibria – everything else remains unchanged. Now the sequence length must be increased from two to three periods. In the last two periods of this sequence the players must alternate between the two Nash equilibria. Thus, for $\text{T}=3$, we would have the sequence (6, 6), (2, 4) and (4, 2). If a player deviates, s/he nets a gain of 1 but is punished by a loss of 2 which, at best, s/he suffers two rounds later (receiving only a payoff of 2 instead of 4). It is straightforward to check that for larger T (and again $\delta_i = 1 \ \forall \ i \in \text{I}$), the sequence in the case of $\text{T}=3$ is repeated. Denoting the sequence length by m, that is, $m=3$ in the example, and the 'remaining stages' by R, then:

$$\text{R}(m, \text{T}) = \text{T} - m[\text{T}/m] \tag{6.1}$$

is the number of rounds at the beginning of the game in which one of the stage game NE is played, followed by $[\text{T}/m]$ times in which the sequence of length m is played.[1] Therefore, for large T, the average payoff to a player in the extended PD game IV approaches 4. Compared to the extended PD game III, the lack of a symmetric Nash equilibrium causes the equilibrium average payoff to be lower in the extended PD game IV because the sequence length has to be extended to three periods.

6.2 SOME GENERAL REMARKS AND RESULTS

From the introductory examples it is evident that the concept of renegotiation-proofness is quite powerful in reducing the number of equilibria (and

equilibrium average payoff tuples). Though it was rather simple to determine the set of RPE in these examples, in other games this task can become a tricky one. Therefore, the restriction to pure strategies is commonly used in this literature. However, since some games have mixed-strategy equilibria that Pareto-dominate pure-strategy equilibria, this restriction is not innocuous.[2]

A central feature of RPE is that in repeated games the future is given more weight (and, consequently, the past less weight) than in SPE and NE (Farrell and Maskin 1989a, p. 331). Of course, if the past were treated as sunk and assumed to be irrelevant to future behavior, then threats could not be used and (conditional) cooperation must fail. Thus, history must be given some weight. The Nash equilibrium concept gives the past much weight. In fact, it goes as far as to assume that the history of the game can overcome players' individual incentives to optimize in the future. The concept of SPE is an intermediate assumption: history cannot overcome players' incentive to optimize but affects the choice of the equilibrium continuation payoff, even to the extent that this can lead to Pareto-dominated continuation payoffs. The RPE concept is in between 'history does not matter' and the SPE concept.

Whereas in an SPE a deviation may trigger a punishment for the rest of the game, in an RPE the possible negative future implications for the punisher are also considered. This seems plausible as long as we do not assume that the punisher receives some additional utility from taking revenge.[3] In a wider context this may be taken as a reason why, for instance, trade sanctions or boycotts have hardly been successful in the past. As long as those countries which boycott the supposed 'culprit' for some wrong-doing also suffer from these restrictions, announcements of boycotts are non-credible threats.

Since renegotiation-proofness combines subgame-perfection with the idea of Pareto-optimality, it can be (in the case of finite games) defined recursively like subgame-perfection. Let $\mathrm{Eff}(\Pi)$ denote the set of weakly efficient points in Π, and let $\Pi^{\delta,\mathrm{SPE}}(t=\mathrm{T})$ be the set of discounted payoffs of pure-strategy SPE of Γ^{T} with typical element $\pi_i^{\delta}=\Sigma\delta_i^t\pi_{it}(\sigma_i^{\mathrm{SPE}}(h^t))$. Moreover, let $\Pi^{\delta,\mathrm{RPE}}(t=\mathrm{T})$ denote the set of discounted payoffs of RPE and T^* the length of the game.

Renegotiation-proof equilibria in finite games:

Set:

$$\mathrm{T}:=1$$

$$\mathrm{X}(\mathrm{T})=\Pi^{\delta,\mathrm{SPE}}(\mathrm{T}),$$

$$\Pi^{\delta,\mathrm{RPE}}(\mathrm{T})=\mathrm{Eff}(\Pi^{\delta,\mathrm{SPE}}(\mathrm{T})).$$

For $T > 1$:

$T := T + 1$

$X(T) = \{\Pi^{\delta,SPE}(T) \mid$ all continuation payoffs $\Pi^{\delta,SPE}(\sigma(h^1))$ specified for Γ^{T-1} lie in $\Pi^{\delta,RPE}(T-1)\}$.

Continue until $T = T^*$.

In words, the process works as follows. First, determine the set of SPE payoffs in the last round. Second, select the efficient ones among them. Third, increase the number of rounds by 1 as long as $T < T^*$ and find for this game those SPE payoffs, *given* the continuation RPE payoffs in the rounds succeeding the first round. Choose the efficient SPE payoffs determined for Γ^T. From then onwards the procedure is repeated. The definition also covers the general case of discounting, examples of which will be provided below.

We now summarize and generalize the results of the PD game examples in the introductory section. PD game II is covered by Theorem 6.1, PD games III and IV by Corollary 6.1 below.

Theorem 6.1
In a finitely repeated game Γ^T, where there is either a strictly Pareto-dominant stage game Nash equilibrium or only a single stage game Nash equilibrium, the only renegotiation-proof equilibrium is the T-fold repetition of the Pareto-dominant Nash equilibrium or the single Nash equilibrium respectively.

Proof: Consider, first, games with a strictly Pareto-dominant stage game NE. In the last stage the Pareto-efficient Nash equilibrium must be played in any case. Denote the payoff to a player derived from this equilibrium $\pi_i^{N^*(1)}$. If in any previous round a stage game strategy different from the Nash equilibrium were to be played, then there would be an incentive for at least one player to deviate, that is, $\pi_{it}^{D}(s_i(s_j), s_j) - \pi_{it}^{*}(s) > 0$. In order to deter free-riding, it must be possible to punish deviation. Since the punishment path must be Pareto-efficient, only the Pareto-efficient Nash equilibrium can be used. Since $\pi_{it}^{*}(s) - \pi_{it}^{N^*(1)}(s^{N^*}) \geq 0 \ \forall \ i$ must hold – otherwise the alternative strategy combination s is not Pareto-efficient – deviation cannot be deterred efficiently.

Since the case of a single NE is a special case of games with a Pareto-dominant NE, the same kind of reasoning applies. Since the unique NE will be played in the last round, there is no punishment

strategy which can deter free-riding. Any other punishment strategy different from the NE is either not subgame-perfect or implies a lower payoff to the punisher. QED

An immediate implication of Theorem 6.1 is summarized in the following corollary:

Corollary 6.1
In a finitely repeated game Γ^T at least two weakly undominated stage game Nash equilibria must exist to establish a stage game strategy combination different from the Nash equilibrium.

Proof: Follows from Theorem 6.1. QED

Since this corollary is just a negation of Theorem 6.1, it suffices to stress its validity by giving a 'knife-edge example'. For instance, consider the example $T = 2$, $\pi^{N^*(1)} = (4, 2)$, $\pi^{N^*(2)} = (2, 2)$, $\pi^* = (5, 2)$, $\pi_1^D = 6$ and $\pi_2^D = (\pi_2 \leq 2)$ where only player 1 can gain from deviating. In this case player 1 gains 1 in the first period, but if in the second period the second equilibrium is played s/he receives only 2 instead of 4. Since player 2 is indifferent between both NE, free-riding can be deterred efficiently as long as $\delta_1 \geq 1/2$ holds.

In the PD game examples discussed above, the set of average RPE payoffs contained only a *single* element (assuming large T and no discounting) which *did not* lie on the Pareto frontier. This is a first category of games identified by Benoît and Krishna (1993). However, there are also games of a second category in which the (closed) set of RPE payoffs is a subset of the points lying on the Pareto frontier. Benoît and Krishna show that all games belong either to the first or to the second category.

Theorem 6.2
In finitely repeated games Γ^T with no discounting and large T, the average equilibrium payoff tuple is either unique and lies in the interior of the payoff space or the set of average payoff tuples is a closed and connected set lying on the Pareto frontier.

Proof: Omitted. See Benoît and Krishna (1993).

Since the proof is quite involved and rather lengthy, we skip it here, but provide an example of the second category of games below. Assume that cooperation will be established in a chicken game where both countries invest in abatement technology (action a_i) and receive a payoff of a_i (see

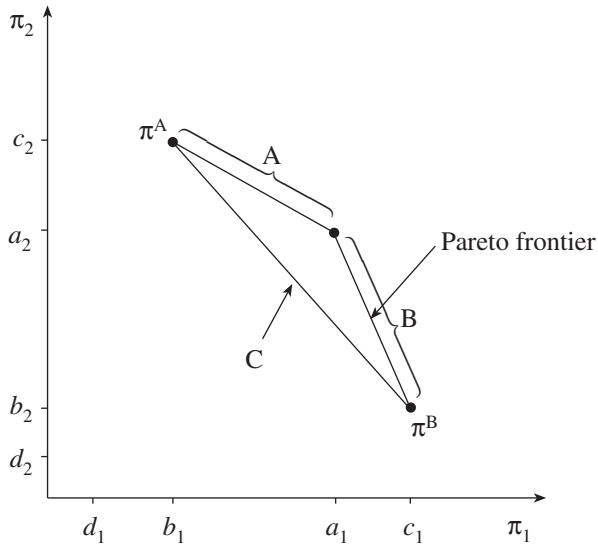


Figure 6.2 Renegotiation-proof equilibrium payoffs in a finitely repeated chicken game

Matrix 3.2). Recall that there are two strictly undominated (pure strategy) Nash equilibria in this game. Like in the extended PD game IV, these equilibria have to be played in turn in the last two periods. However, in the chicken game, the Nash equilibria are Pareto-efficient and can therefore be played at any time and for any duration. Moreover, the payoff in the Nash equilibria is for one player *superior* to the cooperative payoff ($c_i > a_i$) and may therefore be used as punishment. Hence, in contrast to extended PD games III and IV, *no* sequence has to be played involving playing a cooperative strategy in one period followed by playing Nash equilibria subsequently. Consequently, total equilibrium payoffs are given by $(\mathrm{T}-2)\cdot a_i + c_i + b_i$ and $(\mathrm{T}-2)\cdot a_j + b_j + c_j$ with average payoffs $\bar{\pi}_i^{\mathrm{RPE}} = a_i \; \forall \; i \in \mathrm{I}$ if T is large.

An immediate implication of the arguments presented above is that to play: (a) always one of the two NE; (b) a combination of (a_1, a_2) and one NE, or (c) a combination of NE, are also an RPE as long as average payoffs lie on the Pareto frontier. Playing only one Nash equilibrium throughout the game leads to an average payoff tuple at the boundary of the Pareto frontier (either π^{A} or π^{B} in Figure 6.2). A combination of (a_1, a_2) and one Nash equilibrium implies a payoff tuple on the segment A or B of the Pareto frontier. A combination of the two NE may, depending on the specific payoffs, lead to average payoff tuples lying on the Pareto frontier.

However, due to our (restrictive) assumption in Sub-section 3.3.1 that in the chicken game (a_1, a_2) is socially optimal, that is, $a_i + a_j > b_i + c_j$ and $a_i + a_j > c_i + b_j$, the Pareto frontier is strictly convex. Hence, a combination of NE will lead to an efficient payoff tuple lying on the line segment C which is (strictly) below the Pareto frontier. Hence, given this restriction, the alternation between both NE is not an RPE.

However, the RPE path described in (a) and (b) above is sufficient to generate any payoff tuple on the Pareto frontier (line segments A and B). Thus the chicken game is a member of the second category of games in which the set of renegotiation-proof average payoffs is a closed subset of the Pareto frontier.[4] The result may be summarized as follows:

Theorem 6.3
In finitely repeated chicken games the set of renegotiation-proof average payoffs comprises all points on the Pareto frontier provided discount factors are close to 1.

Proof: The proof basically follows from the discussion above. It remains to stress the validity in the case of discounting. First, deviation does not pay if an NE is played in some stage. Second, if (a_i, a_j) is played, then deviation may occur. The weakest threat is if the *equilibrium play* involves playing a Nash equilibrium for t^* periods in which the potential deviator receives only a payoff of b_i subsequent to the deviation. This implies that punishment is delayed by at least t^* periods. After these t^* periods either the last two terminal rounds have been reached in which an alternation between the two NE would be played in equilibrium, or the game does not terminate after the t^* periods and equilibrium play involves resuming the play of a_i/a_j. In the former case the 'worst' case implies that the deviator suffers a loss of $c_i - b_i$ from the punishment only in the last round. In the latter case, the *lower bound* of the loss due to punishment is $a_i - b_i$ in period $t^* + 1$. Thus, deviation does not occur if either $\delta_i \geq [(c_i - a_i)/(c_i - b_i)]^{(1/(t^*+2))} = \delta_i^{\min}$ or $\delta_i \geq [(c_i - a_i)/(a_i - b_i)]^{(1/(t^*+1))} = \delta_i^{\min}$ hold. Both inequalities can generally be satisfied provided discount factors are sufficiently large. QED

It should be evident that, in the case of the extended PD games III and IV, RPE can be derived in the presence of discounting. Generally, however, it should be noted that, though discounting does not change the main results for considerations at the limit (sufficiently high discount factors), the discount factor may very well affect equilibrium play in a game and hence the average payoffs obtainable in a game. This will be briefly demonstrated for the extended PD game III.

In the extended PD game III discounting implies that the sequence length $m=2$ where (a_1, a_2) and (na_1, na_2) is played has to be increased. Since the gain from a deviation of 1 can be punished by a loss of 1 per period in the subsequent rounds, deviation can be deterred provided:

$$\sum_{t=1}^{m-1} \delta_i^t \geq 1 \Leftrightarrow m \geq \frac{\ln(2\delta_i - 1)}{\ln(\delta_i)} := m^{\min}. \tag{6.2}$$

(6.2) can only be satisfied if $\delta_i > 1/2$. Suppose this to be the case, then it is straightforward to show that $\partial m^{\min}/\partial \delta_i < 0$. That is, the lower the discount factor, the longer will be the sequence m in which the symmetric NE must be played $m-1$ times. Since in this game the cooperative payoff strictly Pareto-dominates the Nash equilibrium payoffs, this implies that with decreasing δ_i the average payoff to each player decreases. For $\delta_i \to 1/2$ and large T, we have $\bar{\pi}_i^{\delta,\mathrm{RPE}} \to 3$ whereas for $\delta_i = 1$ we found $\bar{\pi}_i^{\delta,\mathrm{RPE}} = 4.5$ above.

Summarizing, the concept of renegotiation-proofness reduces the number of equilibria and average payoffs sustainable in repeated games compared to subgame-perfection. Renegotiation-proofness defines credible threat strategies more narrowly. It requires that during the punishment (that is, off the equilibrium path) the punisher receives at least the same continuation payoff as along the equilibrium path. To sustain a cooperative outcome for some time, a game must have at least two undominated stage game NE. Whereas this is true in the chicken game and the extended PD games III and IV, this does not hold in the ordinary PD game and the extended PD game II. In the case of no discounting, the equilibrium average payoff either reduces to a single point below the Pareto frontier or comprises a closed set of average payoffs on the Pareto frontier.

6.3 EXTENSION: STRONGLY PERFECT EQUILIBRIA

In the previous sections it became apparent that, though renegotiation-proofness requires equilibrium strategies to be Pareto-efficient, equilibrium payoffs may not be Pareto-efficient with respect to the entire payoff space. The reason is that renegotiation-proofness requires that only among the subgame-perfect continuation strategies are the efficient ones played but not with respect to all possible continuation strategies. The extended PD game versions discussed above are typical examples.

A concept which rules out inefficient continuation payoffs is the concept of strongly subgame-perfect equilibrium (SSPE), or for short, *strongly*

perfect equilibrium.[5] According to this concept, only efficient strategies are allowed to be played at each point in time, implying that the average SSPE payoff must be an element of the Pareto frontier. The basic idea is that players never agree on strategies which deliver them payoffs below Pareto-efficient levels. This is true for equilibrium payoffs but also for payoffs off the equilibrium path, for example, for punishments in the case of a deviation.

Let $P(\Pi)$ denote the Pareto frontier, then, using our previous notation to define an RPE, an SSPE may be defined as follows:

Strongly subgame-perfect equilibrium

Set:

$T: = 1$

$P(\Pi) = \text{Eff}(\overline{\Pi}^{\delta}(T))$,

$\overline{\Pi}^{\delta,\text{SSPE}}(T) = \text{SPE}(P(\Pi))$.

For $T > 1$:

$T: = T + 1$

$P(\Pi) = \text{Eff}(\overline{\Pi}^{\delta}(T))$,

$\overline{\Pi}^{\delta,\text{SSPE}}(T) = \{\text{SPE}(P(\Pi)) \mid \Pi(1) \subseteq P(\Pi)$ and all average continuation payoffs $\overline{\Pi}^{\delta,\text{SSPE}}(T)(\sigma(h^1))$ prescribed on Γ^{T-1} lie in $P(\Pi)\}$.

Continue until $T = T^*$.

Thus, in contrast to the recursive definition of a renegotiation-proof equilibrium, it is convenient (though not necessary) to express all payoffs as average payoffs, so that regardless of the discount factor we can write $P(\Pi) = \text{Eff}(\overline{\Pi}^{\delta}(T))$. SPE(..) means subgame-perfect and hence $\text{SPE}(P\overline{\Pi}^{\delta}(T))$ denotes the set of Pareto-efficient subgame-perfect payoffs.

Since the above definition imposes very strong requirements on equilibrium strategies, an SSPE does not exist in many games, as in all PD game versions we considered above. In contrast, in the chicken game both stage game NE lie on the Pareto frontier (and are individually rational) and can be used as a punishment. Therefore, in this game all RPE average payoffs are also in the set of SSPE average payoffs, that is, $\overline{\Pi}^{\delta,\text{RPE}} = \overline{\Pi}^{\delta,\text{SSPE}}$. These findings may be summarized in the following two theorems:

Theorem 6.4
In the finitely repeated PD game and its extended versions II, III and IV no strongly perfect equilibrium exists.

Theorem 6.5
In the finitely repeated chicken game the set of strongly perfect equilibrium average payoffs comprises all points on the Pareto frontier provided discount factors are close to 1.

Proof: Follows from SSPE = Eff(RPE) where the RPE have been derived above. QED

6.4 DISCUSSION

Like an RPE, an SSPE requires from an equilibrium that it is not challenged by any other subgame. However, unlike an RPE, it also allows a challenge by another subgame, even though the payoffs of this alternative subgame may require playing strategies which are not self-enforcing. Therefore, one may wonder whether the SSPE requirements are not unduly restrictive. Of course, if there are efficient punishments with respect to the entire payoff space, we would expect them to be played. However, if such strategies do not exist in a game, renegotiation-proof punishment strategies seem a pragmatic solution. They ensure that the punisher will carry out the punishment and so free-riding can be credibly deterred. In particular, one may wonder why players would not partially cooperate in an extended PD version, just because of a possibly inefficient punishment.

However, the RPE concept also has shortcomings. First, the cyclical cooperation in the case of the extended PD game versions has no immediate interpretation. It is difficult to see how cyclical cooperation could be transformed in an IEA. It is hardly conceivable that governments would cooperate in one year, not cooperate in consecutive years and then resume cooperation. Of course, as in the case of mixed strategies, one may interpret the outcome of cyclical cooperation as an 'average result' over a certain period of time. But nevertheless, some doubts remain.

Second, the reason for playing such sequences or cycles involving inefficient payoffs is that it must be possible to punish a deviator incentive-compatible at any time. This requires that the punishment delivers a payoff to the punisher at least as high as when the play continued along the equilibrium path. To make this possible, and only for this reason, equilibrium play requires an inefficient strategy in some stages. Thus, one may wonder why players would forgo a possible cooperative welfare gain just so that punishment could be constructed incentive-compatible.

The advantage of both concepts is that they nicely demonstrate an important dilemma in international policy coordination. To establish partial cooperation as an equilibrium, the minimum requirement is for two undominated stage game NE. That is, it must be possible to punish non-compliance by strategies which are at the same time incentive-compatible for the punisher. This is a condition which might not be met in reality and stresses an important problem frequently encountered in international politics. We have already pointed out the problem of using trade sanctions to discipline governments, but the basic problem pertains also with respect to other measures which might be used to enforce an IEA.

NOTES

1. The bracket indicates an integer value.
2. An extension to mixed strategies can be found in Benoît and Krishna (1993, s. 6).
3. If this were the case, this would have to be modeled explicitly in the utility functions of the players.
4. In other games the RPE average payoffs are a true subset of the Pareto frontier. See Benoît and Krishna (1993).
5. For games with more than two players, SSPE requires more than subgame-perfection and Pareto-efficiency of all continuation payoffs. See Chapter 15. The basic idea of this equilibrium concept has been formulated by Aumann (1959) for static games. A definition for dynamic games may be found in Rubinstein (1980).

7. Infinite dynamic games with discrete strategy space: a second approach

7.1 WEAKLY RENEGOTIATION-PROOF EQUILIBRIA

7.1.1 The Concept[1]

In Chapter 6 it became clear that by requiring strategies to be renegotiation-proof the number of equilibria in repeated games could be substantially reduced. Moreover, requiring strategies to constitute a strongly perfect equilibrium reduced the set of equilibria even further. However, it turned out that for many games for which a renegotiation-proof equilibrium exists, no strongly perfect equilibrium can be found.

For finite games an obvious way to define a Pareto-efficient subgame-perfect strategy involved a recursive definition. Now, in an infinite time horizon, such a definition is not available, which leaves some leeway for finding an adequate formulation of what renegotiation-proofness means for supergames. Since Farrell and Maskin's (1989a, b) definition has probably found the most widespread application in the literature, we concentrate exclusively on their concept of weakly and strongly renegotiation-proof equilibria.[2] It should be mentioned that the authors exclusively restrict the validity of their concept to two-player games and we follow this assumption in this chapter too. The possibility of an extension to N-player games will be discussed in Chapter 14.

Farrell and Maskin's definition of a *weakly renegotiation-proof equilibrium* (WRPE) takes up the central idea of the previous chapter that an equilibrium strategy should have no Pareto-dominated continuation payoff in any subgame. In particular, in the 'punishment subgame' the punisher should not find it attractive to skip the punishment.

Once more, it turns out that in order to check whether a payoff vector can be backed by weakly renegotiation-proof strategies it suffices to use a simple strategy profile *à la* Abreu. Like the check for a subgame-perfect equilibrium (SPE) (see in particular the proof of Fudenberg and Maskin's Folk Theorem 5.4, pp. 69–71), the game may be divided into a *cooperative*

and a *punishment* phase. In the cooperative phase (equilibrium path) both players comply with the agreed strategy $s = (s_i, s_j)$. If player i defects (off the equilibrium path) in one period and chooses his/her best deviation strategy $s_i(s_j)$, which follows from max $(s_i)\pi_i(s_i, s_j)$, player j starts the punishment from the next period onwards with a strategy s_j^i. Player i has two possibilities for reacting: either to accept the punishment and play strategy s_i^j for the punishment duration t_i^P (*repentance phase*) or not to give in and continue with defection, playing his/her best response to the punishment $s_i(s_j^i)$ which follows from max $(s_i)\pi_i(s_i, s_j^i)$. In the first case, the players go back to the cooperative phase after the punishment. In the second case, the punishment is prolonged (*retaliation phase*).

Of course, by symmetry, the same strategy profile is played against player j. In particular, if player j deviates during the punishment, s/he will be punished. Formally, the strategy may be summarized as follows:

$$\sigma_i = \begin{cases} s_i^j \text{ if in any } h^t(t-t_j^P) \ldots h^t(t-1) = (s_i, s_j'), \text{ or } h^t(t-t_j^P) \ldots h^t(t-1) = (s_i^j, s_j'') \\ s_i \text{ if in any } h^t(t-t_j^P) \ldots h^t(t-1) = (s_i, s_j), \text{ or } h^t(t-t_j^P) \ldots h^t(t-1) = (s_i^j, s_j^i). \end{cases} \tag{7.1}$$

Strategy σ_i is a WRPE strategy, where the stage game strategy $s = (s_i, s_j)$ is played along the equilibrium path and where players receive the payoff π_i^* as an average payoff, provided the following conditions are satisfied (see Farrell and Maskin 1989a, pp. 335ff.):

Weakly renegotiation-proof equilibrium conditions in a two-player game

$$\pi_i^*(s_i, s_j) \geq (1-\delta_i)\pi_i^D(s_i(s_j), s_j) + \delta_i \pi_i^C(s_i(s_j^i), s_j^i) \tag{7.2}$$

$$\pi_i^P(s_i^i, s_j^i, s_i, s_j) \geq \pi_i^C(s_i(s_j^i), s_j^i) \tag{7.3}$$

$$\pi_i^*(s_i, s_j) \geq (1-\delta_i)\pi_i^D(s_i(s_j), s_j) + \delta_i \pi_i^P(s_i^i, s_j^i, s_i, s_j) \tag{7.4}$$

$$\pi_j^*(s_i, s_j) \leq \pi_j^R(s_i^i, s_j^i) \tag{7.5}$$

$\forall\, i \in \{1, 2\}$ and $(\pi_i^*, \pi_j^*) \in \Pi^{IR}$ where:

$$\pi_i^P = (1-\delta_i^{t_i^P})\pi_i^R(s_i^i, s_j^i) + \delta_i^{t_i^P}\pi_i^*(s_i, s_j). \tag{7.6}$$

Here payoffs are expressed as average payoffs of the infinite game (discounted payoffs are multiplied by $(1-\delta_i)$) assuming $\delta_i \neq 1$ (see Section 5.1).

Inequality (7.2) states that the discounted payoff from cheating, π_i^D (best

deviation payoff),[3] and subsequently receiving the 'retaliation phase payoff', π_i^C, must be lower than the payoff in the cooperative phase, π_i^*. This deters deviation in the first place.

Inequality (7.3) ensures that a potential cheater has an incentive to accept the punishment after his/her deviation. Therefore, the average continuation payoff if a player complies with his/her punishment, π_i^P, must be at least as high as when punishment is continued. Thus if deviation occurred and punishment followed, it would ensure that there was an incentive to resume cooperation.[4]

Inequality (7.4) guarantees that it is also not profitable for player i to deviate in the cooperative phase and then to accept the punishment afterwards.

Condition (7.5) ensures that the punisher has no incentive to renegotiate the agreed punishment. For this the payoff when conducting the punishment, $\pi_j^R(s_i^i, s_j^i)$, must be at least as high as in the cooperative phase.[5] This last condition in particular represents the central idea of the WRPE concept and distinguishes it from the SPE concept.

Finally, (7.6) represents the continuation punishment payoff π_i^P of player i at the beginning of his/her repentance phase. It is a linear combination of the repentance payoff, π_i^R, and the cooperative continuation payoff, π_i^*. Thus, the continuation punishment payoff rises as the punishment proceeds.

Looking at inequalities (7.2)–(7.4), it is evident that, because of (7.3), (7.4) is a stronger requirement than (7.2). Hence, we can drop inequality (7.2) and we are left with three inequalities.

Moreover, note that for δ_i close to 1 ($\delta_i^{ti} \to 1$) (7.3) and (7.4) reduce further to:[6]

$$\begin{aligned} &\pi_i^*(s_i, s_j) > \pi_i^C(s_i(s_j^i), s_j^i) \text{ if } s_i \neq s_i(s_j) \\ &\pi_i^*(s_i, s_j) \geq \pi_i^C(s_i(s_j^i), s_j^i) \text{ if } s_i = s_i(s_j). \end{aligned} \tag{7.7}$$

Taken together, a strategy tuple is weakly renegotiation-proof for δ_i close to 1 if it satisfies (7.5) and (7.7) simultaneously $\forall\, i \in I$ and (7.3)–(7.5) $\forall\, i \in I$ if $\delta_i << 1$. In the former case, Theorem 7.1 below is an immediate implication:

Theorem 7.1 (Farrell and Maskin 1989a, Theorem 1)
Let $\pi^* = (\pi_i^*, \pi_j^*) \in \Pi^{IR}$ be an individually rational payoff tuple from some cooperative stage game strategy combination $s = (s_i, s_j)$. If there exists a strategy tuple $s^i = (s_i^i, s_j^i)$ such that $\pi_i^C(s_i^i, s_j^i) \leq \pi_i^*(s_i, s_j)$ and $\pi_j^R(s_i^i, s_j^i) \geq \pi_j^*(s_i, s_j)$, then the payoff tuple π^* can be backed by weakly renegotiation-proof equilibrium strategies for sufficiently large discount factors.

Proof: Follows immediately from derivation of conditions (7.5) and (7.7). QED[7]

An immediate implication of Theorem 7.1 is that there is no existence problem of a WRPE in all games for which a stage game Nash equilibrium exists. For $\pi_i^N = \pi_i^*$ and $\pi_j^N = \pi_j^*$ the conditions above reduce to $\pi_i^C = \pi_i^N$ and $\pi_j^R = \pi_j^N$ and are therefore trivially satisfied. Since Nash equilibria are typically Pareto-inefficient for negative externality games (see, for instance, all the PD game versions), it is evident that a WRPE is not necessarily efficient with respect to the entire payoff space. This is similar to the result we found for renegotiation-proof equilibria in finite games and we shall therefore present some extensions of the WRPE concept below. Before doing so in Section 7.2, however, we shall first discuss some other results with respect to WRPE and also illustrate the concept with some examples.

Another immediate implication of the WRPE conditions (7.2)–(7.5) is summarized in the following corollary:

Corollary 7.1
For $\delta_i \rightarrow 0\ \forall\ i$ the set of weakly renegotiation-proof equilibria converges to the set of stage game Nash equilibria.

Proof: Since (7.2)–(7.6) are continuous in δ_i, the set of equilibria satisfying these WRPE conditions for $\delta_i \rightarrow 0\ \forall\ i \in I$ converges to the set for $\delta_i = 0\ \forall\ i \in I$ which contains only stage game Nash equilibria. For $\delta_i \rightarrow 0\ \forall\ i \in I$, (7.4) becomes $\pi_i^D = \pi_i^*\ \forall\ i \in I$. Hence, $s_i = s_i(s_j)$ and $s_j = s_j(s_i)$ from which $s^N = (s_i^N, s_j^N)$ follows. QED

Thus, Corollary 7.1 is analogous to Corollary 5.1 in Section 5.2, where this result has been shown to apply to subgame-perfect strategies.

To see what changes are needed if equilibrium strategies are required to be weakly renegotiation-proof instead of subgame-perfect, we consider some examples below. We start by assuming $\delta_i \rightarrow 1\ \forall\ i \in I$ and hence conditions (7.5) and (7.7) apply (Sub-section 7.1.2). Subsequently, we consider discount factors strictly smaller than 1 and therefore conditions (7.3), (7.4) and (7.5) must be satisfied (Sub-section 7.1.3).

7.1.2 Discount Factors Close to 1

Extended PD game V
Consider the extended PD game V in Matrix 7.1, the payoff space of which has been drawn in Figure 7.1. This PD game version is basically the same as the extended PD game I (Matrix 4.2), except that now the punishment

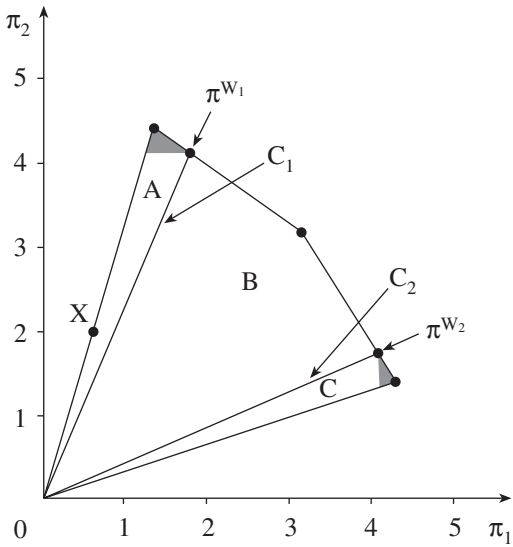


Figure 7.1 Payoff space of the extended PD game V

option delivers a payoff of 0 to both players alike instead of 1 to the player being punished and 0 to the punisher. Strategy na_i is a *dominant strategy* in this game and $\Pi = \Pi^{IR}$ which corresponds to the area A, B and C (including the hatched triangles) in Figure 7.1. Consequently, according to Folk Theorem IV, the entire payoff space can be backed by subgame-perfect strategies, that is, $\Pi = \overline{\Pi}^{SPE}$. What about the weakly renegotiation-proof payoff space $\overline{\Pi}^{WRPE}$?

To capture the entire payoff space, we allow for mixed strategies. Then WRPE conditions (7.5) and (7.7), assuming player 1 to be the potential deviator, may be written as:

$$\pi_1^* \geq \pi_1^C(p_1, p_2, q_1^1, q_2^1) \tag{7.8}$$

$$\pi_2^* \leq \pi_2^R(p_1^1, p_2^1, q_1^1, q_2^1). \tag{7.9}$$

Since we know that any payoff vector $\pi^* = (\pi_1^*, \pi_2^*)$ can be generated by mixed strategies, the probabilities on the LHS of both inequalities have been omitted as arguments. The probabilities on the RHS refer to the strategies in Matrix 7.1. Again, superscripts refer to the player who gets punished.[8]

Matrix 7.1 Extended PD game V

	a_2 (q_1)	na_2 (q_2)	p_2 ($1-q_1-q_2$)
$a_1(p_1)$	3.2 3.2	1.4 4.4	0 0
$na_1(p_2)$	4.4 1.4	**2** **2**	0 0
$p_1\,(1-p_1-p_2)$	0 0	0 0	0 0

First note that $p_2 = 1$ is a dominant defection strategy in this game and hence $p_1 = 0$ and $1 - p_1 - p_2 = 0$ in (7.8). Moreover, irrespective of player 2's strategy, (7.9) is 'most easily' satisfied if player 1 plays strategy a_1 with probability $p_1^1 = 1$. Since we are interested in the outer boundaries of the WRPE space, we set $p_1^1 = 1$ in (7.9). Thus, inserting this information into inequalities (7.8) and (7.9) and using the payoffs of Matrix 7.1, we get:

$$\pi_1^* \geq 4.4q_1^1 + 2q_2^1 \tag{7.10}$$

$$\pi_2^* \leq 3.2q_1^1 + 4.4q_2^1. \tag{7.11}$$

From (7.10) it is evident that to determine the lowest WRPE payoff to player 1, q_1^1 should be low compared to q_2^1. From (7.11) it is evident that to find the highest payoff to player 2 which can be backed by WRPE strategies a high value should be chosen for q_2^1 compared to q_1^1. Thus, we may well assume $q_1^1 = 0$ in (7.10) and (7.11). Then, assuming (7.11) to be binding, we can solve for $q_2^1 = \pi_2^*/4.4$. Inserting this into (7.10), we obtain the first WRPE condition C_1:

$$C_1: = \pi_1^* \geq \frac{2\pi_2^*}{4.4}. \tag{7.12}$$

Of course, a similar condition can be derived assuming player 2 to be the potential deviator:

$$C_2: = \pi_2^* \geq \frac{2\pi_1^*}{4.4}. \tag{7.13}$$

Both conditions are drawn in Figure 7.1. All feasible payoff tuples which lie to the right of C_1 and to the left of C_2 can therefore be backed by WRPE strategies. Therefore, $\overline{\Pi}^{WRPE}$ corresponds to the area denoted B in Figure 7.1 which is a (true) subset of $\overline{\Pi}^{SPE}$.

Ordinary PD game

Let us now turn to the ordinary PD game and check for the WRPE payoff space. Denote p_1 the probability that a_1 is played and q_1 the probability that a_2 is played. Assume payoffs as given in Matrix 3.1, the payoff space of which is drawn in Figure 7.2. The entire payoff space, Π, comprises Π^{IR} and, additionally, the areas A and B (including the hatched triangles).

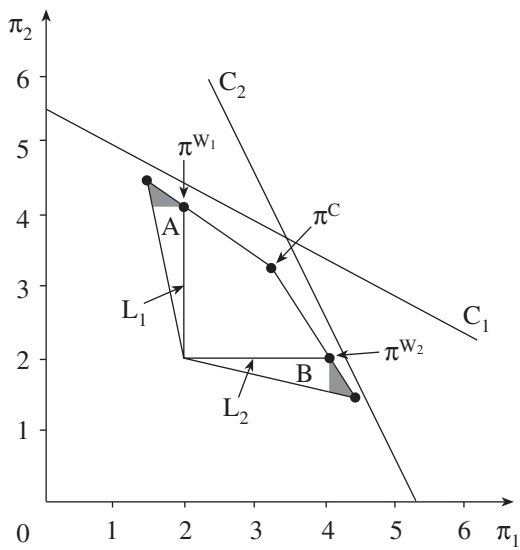


Figure 7.2 Payoff space of the PD game

Again, players have a dominant defection strategy and hence $p_1 = 0$ ($q_1 = 0$) if player 1 (player 2) deviates. Moreover, in the punisher 2 (1)'s inequality $p_1^1 = 1$ ($q_1^2 = 1$) delivers the highest possible payoff. Hence, assuming player 1 to be the potential defector, the WRPE conditions are given by:

$$\pi_1^* \geq 2 + 2.4q_1^1 \tag{7.14}$$

$$\pi_2^* \leq 4.4 - 1.2q_1^1. \tag{7.15}$$

Since the minimax payoff in this game is 2, we assume the first inequality to be binding and solve for q_1^1. We get $q_1^1 = (\pi_1^* - 2)/2.4$ and substitute this into (7.15) which delivers:

$$C_1: = \pi_2^* \leq 5.4 - \frac{\pi_1^*}{2}; C_2: = \pi_1^* \leq 5.4 - \frac{\pi_2^*}{2} \tag{7.16}$$

where the second WRPE condition follows from symmetry.

Both conditions are drawn in Figure 7.2, from which it is evident that both conditions are non-binding. All payoff tuples in Π^{IR} are below the C_1 and C_2 line and hence $\overline{\Pi}^{SPE} = \overline{\Pi}^{WRPE}$. Thus, the PD game is an example where the WRPE concept does *not* reduce the set of *equilibria payoffs* compared to the SPE concept.

Theorem 7.2
In an infinitely repeated (ordinary) PD game with stage game payoffs as given in Matrix 3.2 and payoff structure $c_i > a_i > d_i > b_i$, $a_1 + a_2 > b_1 + c_2$, $a_1 + a_2 > b_2 + c_1$ and $a_1 + a_2 > d_1 + d_2$, $\overline{\Pi}^{SPE} = \overline{\Pi}^{WRPE}$ for $\delta_i \to 1\ \forall\ i \in I$.

Proof: See Appendix IV.2 for a general proof. QED

The reason for this result can easily be seen in Figure 7.2. For each payoff tuple in Π^{IR} there is a payoff tuple in Π which can be used to punish deviation and which gives the punisher more than in the cooperative phase. This is particularly important for tuples on the border of Π^{IR} which comprises lines L_1 and L_2 and the part of the Pareto frontier between π^{W1} and π^{W2}. Take for instance point $\pi^{W1} = (\pi_1^{W1}, \pi_2^{W1})$ which is defined as that payoff tuple which delivers the lowest weakly renegotiation-proof payoff on the Pareto frontier to player 1. Accordingly, the free-rider incentive for player 1 is high and a severe threat is needed to deter deviation. In the case of $\pi_i^* = \pi^{W1}$ one may choose $\pi = (1.4, 4.4)$ as a punishment tuple or any payoff tuple lying in the hatched area which is located at the upper left part of area A in Figure 7.2. More generally, in a PD game for any payoff tuple $\pi^* \in \Pi^{IR}$ there are strategy combinations available for which $\pi_i^C(s_i^i, s_j^i) \leq \pi_i^*(s_i, s_j)$ and $\pi_j^R(s_i^i, s_j^i) \geq \pi_i^*(s_i, s_j)$ hold – the conditions of Theorem 7.1.[9]

In contrast, in the game depicted in Figure 7.1, $\Pi = \Pi^{IR}$ and therefore it *must* be that payoff tuples along the border of Π^{IR} cannot be backed by WRPE strategies. For instance, consider point X. Punishing player 1 requires a punishment tuple in a north-west direction, which is not feasible because this would lie outside Π.

Chicken game
Drawing Π and Π^{IR} for a chicken game would also show that we can find for all $\pi^* \in \Pi^{IR}$ strategies such that $\pi_i^C(s^i) \leq \pi_i^*(s)$ and $\pi_j^R(s^i) \geq \pi_i^*(s)$ hold.

Theorem 7.3
In an infinitely repeated chicken game with stage game payoffs as given in Matrix 3.2 and payoff structure $c_i > a_i > b_i > d_i$, $a_1 + a_2 > b_1 + c_2$, $a_1 + a_2 > b_2 + c_1$ and $a_1 + a_2 > d_1 + d_2$, $\overline{\Pi}^{\text{SPE}} = \overline{\Pi}^{\text{WRPE}}$ for $\delta_i \to 1\ \forall\ i \in \text{I}$.

Proof: Follows along exactly the same lines as the proof of Theorem 7.2 which is provided in Appendix IV.2. Note that in the chicken game punishment is particularly easy because both stage game Nash equilibria are Pareto-efficient and satisfy the inequalities above. Therefore they can be used as punishments. QED

7.1.3 Discount Factors Smaller than 1

In this subsection we want to determine the minimum discount factor requirements for WRPE and compare them to SPE requirements. We consider the same games as in the previous sub-section.

Extended PD game V
Assume that the payoff tuple $\pi^* = (3.2, 3.2)$ shall be backed by WRPE strategies (see Matrix 7.1). For a start, consider only *pure strategies* and hence $\pi_i^{\text{D}} = 4.4$, $\pi^{\text{R}(i)} = (1.4, 4.4)$ (the superscript i refers to the player who gets punished and where $\pi^{\text{R}(i)} = (\pi_i^{\text{R}(i)}, \pi_j^{\text{R}(i)})$ and $\pi_i^{\text{C}} = 2$. To determine $\delta_i^{\min}$ in this symmetric game, we substitute (7.6) into (7.3) and assume for a start punishment time to be given by $t_i^{\text{P}} = 1$. Then, we have:

$$(1 - \delta_i)1.4 + \delta_i 3.2 \geq 2 \Leftrightarrow \delta_i^{(1)} \geq 1/3 \tag{7.17}$$

and (7.4) becomes:

$$3.2 \geq (1 - \delta_i)4.4 + \delta_i((1 - \delta_i)1.4 + \delta_i 3.2) \Leftrightarrow \delta_i^{(2)} \geq 2/3. \tag{7.18}$$

Since (7.18) is more restrictive than (7.17), a harsher punishment should be constructed so that π_i^{P} gets smaller and the $\delta_i^{(2)}$ requirement as well. For this we increase the punishment time to $t_i^{\text{P}} = 2$. Then (7.3) delivers $\delta_i^{(1)} \geq 0.5774$ and (7.4) $\delta_i^{(2)} \geq 0.4574$ and hence $\delta_i^{\min}(\text{WRPE}) = \max\ [\delta_i^{(1)}, \delta_i^{(2)}] = 0.5774$. Though any further increase of the punishment time would reduce $\delta_i^{(2)}$, it increases $\delta_i^{(1)}$ above 0.5774 and would therefore deliver a higher $\delta_i^{\min}$. Consequently, $\delta_i^{\min}(\text{WRPE}) = 0.5774$.

Next consider what will change if we require only SPE strategies to be played in this game. For this, we substitute (5.10) into (5.9), noting that $\pi_i^{\text{G}} = \pi_i^{\text{M}} = 0$, and find:

$$3.2\delta_i^{t_i^{\text{P}}} \geq 0 \tag{7.19}$$

which is trivially satisfied for any punishment time and discount factor. Hence, we may choose $t^P = \infty$ and $\pi_i^P = \pi_i^M = 0$. Inserting this into (5.8) reveals $\delta_i \geq 0.\overline{27} = \delta_i^{min}(SPE)$, which is obviously lower than the discount factor requirement for WRPE strategies. The reason is simple: for a WRPE less severe punishment strategies are available in this game because it must be ensured that the punishment payoff exceeds the cooperative payoff of the punisher.

Now allow for mixed strategies so that all payoff tuples in Π may be obtained. For $\delta_i^{min}(SPE)$ nothing changes because we have already used the harshest possible punishment. However, for $\delta_i^{min}(WRPE)$ there will be a modification.

From (7.4) it is evident that to determine the lowest $\delta_i^{min}(WRPE)$ we should look for a harsh punishment. However, the punishment is restricted by conditions (7.3) and (7.5). For the latter condition we know from the previous section that if, say, player 1 is punished, $p_1^1 = 1$ is the 'best' repentance strategy. Since, due to (7.5), $\pi_2^* = 3.2 \leq \pi_2^P$ must hold, we choose $q_1^1 = 0$, $q_2^1 = 0.\overline{72}$ and $1 - q_1^1 - q_2^1 = 0.\overline{28}$ so that (7.5) becomes binding and $\pi_2^P = 3.2$. From $q_2^1 = 0.\overline{72}$, $\pi_{1,}^C = 1.\overline{45}$ follows (if $q_1 = 1$) which, due to (7.3), constitutes a lower bound for the punishment. If player 1 complies with the punishment s/he receives a stage game payoff of $\pi_1^P = 1.0\overline{18}$ during the t_i^P periods.

Computing as above $\delta_i^{(1)}$ and $\delta_i^{(2)}$ reveals that $\delta_i^{min} = 0.447$ for $t_i^P = 2$. Thus, by enlarging the set of strategies to include mixed strategies reduces the discount factor requirement from $\delta_i^{min}(WRPE) = 0.577$ to $\delta_i^{min}(WRPE) = 0.477$. Nevertheless, also for mixed strategies the discount factor requirements for WRPE strategies are higher than for SPE strategies.

Ordinary PD game

Next, consider the minimum discount factor requirements in the ordinary PD game in Matrix 3.1 and assume $\pi^* = (3.2, 3.2)$. From Section 5.1, $\delta_i^{min}(SPE) = 0.5$ is known. To determine $\delta_i^{min}(WRPE)$ in the case of pure strategies is an easy task because the relevant payoffs are the same as in the extended PD game V, namely $\pi_i^D = 4.4$, $\pi_j^C = 2$ and $\pi^{R(i)} = (1.4, 4.4)$. Hence, $\delta_i^{min}(WRPE) = 0.577$, which is greater than $\delta_i^{min}(SPE) = 0.5$.

In the case of mixed strategies one can show the following:

Theorem 7.4

In an infinitely repeated two-player (ordinary) PD game with stage game payoffs as given in Matrix 3.2, $c_i > a_i > d_i > b_i \; \forall \; i \in I$, $a_1 + a_2 > b_1 + c_2$, $a_1 + a_2 > b_2 + c_1$ and $a_1 + a_2 > d_1 + d_2$, $\delta_i^{min}(WRPE) = \delta_i^{min}$ (SPE) $\forall$ $i \in I$ and $\pi^* \in \Pi^{IR}$ holds.

Proof: See van Damme (1989).[10]

The basic idea of the proof is to construct a harsh punishment so that $\pi_i^P = \pi_i^C$ at the limit. Van Damme shows that a *two-phase* punishment strategy profile will do the job. In the *first phase*, assuming player 1 to be the potential deviator, player 1 cooperates with probability $p^1 = p$ and player 2 with probability $q^1 = 0$ for t_1^P periods. This delivers a low pure strategy payoff $\pi_1^{R(1)}$ to player 1. In the *second phase*, which is played for one period only, player 1 again cooperates with probability $p^1 = p$ but player 2 with probability $q^1 = q$. This delivers an intermediate stage game payoff between $\pi_1^{R(1)}$ and the cooperative payoff π_1^*. The average continuation payoff of the two phases approaches π_1^C from above so that $\delta_1^{min}(SPE) = \delta_1^{min}(WRPE)$. For the example in Matrix 3.1, the details are provided in Appendix IV.3.

Chicken game

For the chicken game a similar result as for the PD game holds.

Theorem 7.5

In an infinitely repeated two-player chicken game with stage game payoffs as given in Matrix 3.2 and payoff structure $c_i > a_i > b_i > d_i \; \forall \; i \in I$, $a_1 + a_2 > b_1 + c_2$, $a_1 + a_2 > b_2 + c_1$ and $a_1 + a_2 > d_1 + d_2$, $\delta_i^{min}(WRPE) = \delta_i^{min}(SPE) \; \forall \; i \in I$ and $\pi^* \in \Pi^{IR}$ holds.

Proof: Theorem 7.5 is easily proved. First note that since (7.3) is never binding in a chicken game the harshest available punishment strategy (a_i, na_j) may be used to punish player i. This delivers the lowest individually rational payoff to him/her, π_i^C, and at the same time the highest obtainable payoff to punisher j, π_j^U, in this game. Since $s = (a_i, na_j)$ is a Pareto-undominated stage game Nash equilibrium, $t_i^P = \infty$ can be chosen. Then $\pi_i^P = \pi_i^C = \pi_i^N$ and $\delta_i^{min}(SPE) = \delta_i^{min}(WRPE) \; \forall \; i \in I$ holds. QED

7.2 STRONGLY RENEGOTIATION-PROOF AND STRONGLY PERFECT EQUILIBRIA

7.2.1 Strongly Renegotiation-proof Equilibria

In the previous section it became apparent that the WRPE are not necessarily Pareto-efficient. On the one hand, the average payoff tuples of a WRPE may not be Pareto-efficient. On the other hand, the continuation payoffs induced by punishment may not be Pareto-efficient. A WRPE only requires that no continuation payoff of an equilibrium strategy is Pareto-dominated by a payoff of another subgame of this strategy. Put differently,

a WRPE strategy is *internally stable* because no path specified by the strategy is Pareto-dominated.

However, there may exist other WRPE (besides that particularly chosen) which Pareto-dominate a particular WRPE and both players could agree to switch to such an equilibrium. In other words, a WRPE may not be *externally stable*. A concept which ensures such an external stability is that of a *strongly renegotiation-proof equilibrium* (SRPE). It requires that no subgame of a WRPE, including the whole game itself, shall be Pareto-dominated by another WRPE. Though this has to be qualified slightly below (see Theorem 7.8), we may write – following the convention of Chapter 6 – $\overline{\Pi}^{\mathrm{SRPE}} = \mathrm{Eff}(\overline{\Pi}^{\mathrm{WRPE}})$, where we recall that Eff denotes *weak* efficiency. Note, however, that because a WRPE does not require efficient punishments, an SRPE may also involve inefficient punishments.

In the following we discuss some theorems related to SRPE and their implications, though we do not prove these theorems because their proofs are quite involved.

Theorem 7.6 (Farrell and Maskin 1989a, Theorem 2)
Let $\mathrm{P}(\Pi^{\mathrm{IR}}) \in \Pi^{\mathrm{IR}}$ denote the Pareto frontier on which all payoff tuples lie for which Eff (π^*) is true. Then in any game of full dimensionality there exists a Pareto-efficient weakly renegotiation-proof equilibrium for discount factors close to 1, that is, $\overline{\Pi}^{\mathrm{WRPE}} \cap \mathrm{P}(\Pi^{\mathrm{IR}}) \neq \{\,\}$.

Proof: See Evans and Maskin (1989).

Figures 7.1 and 7.2 illustrate Theorem 7.6. All points between π^{W1} and π^{W2} are Pareto-efficient WRPE. Whereas in Figure 7.2 these efficient WRPE payoff tuples comprise all individually rational payoff tuples on the Pareto frontier, in Figure 7.1 the efficient WRPE payoff tuples are a (true) subset of $\mathrm{P}(\Pi^{\mathrm{IR}})$. Hence, we may write $\overline{\Pi}^{\mathrm{WRPE}} \cap \mathrm{P}(\Pi^{\mathrm{IR}}) = \mathrm{P}(\overline{\Pi}^{\mathrm{WRPE}})$ or $\mathrm{P}(\overline{\Pi}^{\mathrm{WRPE}}) = \overline{\Pi}^{\mathrm{SRPE}}$.

From Figures 7.1 and 7.2 the next theorem is intuitively appealing:

Theorem 7.7 (Farrell and Maskin 1989a, Theorem 3)
For discount factors close to 1 the set of Pareto-efficient weakly renegotiation-proof equilibria lying on the Pareto frontier, $\mathrm{P}(\overline{\Pi}^{\mathrm{WRPE}})$, is a closed set.

Proof: See Farrell and Maskin (1989a).

Theorem 7.7 states a result which is similar to the findings for finite games. There we found that for $\delta_i = 1\ \forall\ i \in \mathrm{I}$ the renegotiation-proof payoff space for large T is either a single point below the Pareto frontier or a closed

subset on the Pareto frontier. Now for infinite games only the second possibility is relevant.

A corollary of Theorem 7.7 is that for each payoff tuple $\pi^* \in P(\overline{\Pi}^{WRPE})$ there exists a strategy tuple such that $\pi_i^C(s^i) \leq \pi_i^{Wi}$ and $\pi_j^P(s^i) \geq \pi_i^{Wi} \geq \pi_j^*$ where we may recall that π_i^{Wi} denotes the lowest weakly renegotiation-proof payoff to player i located on the Pareto frontier and π_j^{Wi} is the payoff to player j associated with this point. For instance, in Figures 7.1 and 7.2, all points lying in the hatched areas satisfy these conditions and may be used as potential punishment tuples to back an SRPE. The following theorem summarizes the *sufficient conditions* for an SRPE:

Theorem 7.8 (Farrell and Maskin 1989a, Theorem 5)
If in a generic game $\pi_i^C < \pi_i^{Wi} < \pi_i^{Wj}\ \forall\ i = 1, 2$ and $\forall\ i \neq j$ is true, then any payoff vector $\pi \in P(\overline{\Pi}^{WRPE})$ is a strongly renegotiation-proof equilibrium for all discount factors sufficiently close to 1.

Proof: See Farrell and Maskin (1989a).

Condition $\pi_i^C < \pi_i^{Wi} < \pi_i^{Wj}$ is a rather mild requirement and is almost always satisfied for games which are of interest in economics. With respect to Figures 7.1 and 7.2, it requires that player 1 (2) receives more (less) at π^{W2} than at π^{W1}, which almost follows from the definition of this point. If this condition holds, then indeed $\overline{\Pi}^{SRPE} = Eff(\overline{\Pi}^{WRPE})$. Accordingly, all points lying between π^{W1} and π^{W2} on the Pareto frontier of the PD game (see Figure 7.2) and the extended PD game V (see Figure 7.1) are an SRPE.

7.2.2 Strongly Perfect Equilibria

As discussed in the context of finite games, there is an even stronger concept than renegotiation-proofness, namely that of strongly perfect equilibria. An SSPE requires that all continuation payoffs of any subgame are Pareto-efficient with respect to the *whole individually rational payoff space* and not only with respect to the *WRPE payoff space*.[11] In particular, this implies that punishment has to be conducted efficiently. Like SRPE, a necessary condition for SSPE is that they must lie on the Pareto frontier. Due to the higher requirement with respect to the efficiency of punishments, the set of SSPE is a subset of SRPE.[12]

However, in all three examples discussed above, SSPE are not a true subset of SRPE because the SSPE requirement is not binding. In the *PD game* the Pareto-efficient stage game strategy (a_i, na_j) can always be used to punish player i (provided discount factors are sufficiently high), so that even boundary payoff tuples like π^{W1} and π^{W2} can be backed as SSPE.

In the *extended PD game V* we have already ensured an efficient punishment when deriving the WRPE conditions C_1 and C_2. Recall that to go from inequalities (7.8) and (7.9) to (7.10) and (7.11), we assumed $p_1^1 = 1$. Subsequently, we set $q_1^1 = 0$, which implied playing the pure Pareto-efficient strategy tuple (a_i, na_j).

In the *chicken game* (a_i, na_j) is a Pareto-efficient stage game Nash equilibrium which can be used as punishment. Thus, in all three games $\overline{\Pi}^{SRPE} = \overline{\Pi}^{SSPE}$ if $\delta_i \to 1$. In other games, however, SSPE are a true subset of SRPE. We shall encounter such examples in Chapter 12.

NOTES

1. Sub-section 7.1.1 draws on Endres and Finus (1998a) and Finus and Rundshagen (1998b).
2. The origin of the idea can be traced back to the independent work of Farrell (1983) and Bernheim and Ray (1985). For similar concepts, see Abreu *et al.* (1993); Asheim (1991); Bergin and MacLeod (1993); Bernheim and Ray (1989); and Ray (1994). For more intuitive explanations of the gist of the concept, see Fudenberg and Tirole (1996, pp. 174ff.); and Mohr (1988, pp. 551ff.).
3. Farrell and Maskin (1989a) assume the deviation payoff π_i^D to be the maximax payoff of country i, $\pi_i^U = \max(s_j) \max(s_i) \pi_i(s_i, s_j)$ which constitutes an upper bound of the deviation payoff (see Section 4.3). Though this simplifies matters and is of course valid for considerations at the limit, it introduces a bias when computing the minimum discount factor requirements (δ_i^{min}) of a particular cooperative strategy combination. Since we shall be concerned below with computing δ_i^{min} for different 'cooperative' payoff tuples, we are already using π_i^D instead of π_i^U at this introductory stage.
4. Thus (7.3) follows from $\pi_i^P \geq (1-\delta_i)\pi_i^C + \pi_i^P$ after rearranging terms.
5. Condition (7.5) is equivalent to $\pi_j^* \leq \pi_j^P$ because of $\pi_j^P(1-\delta_j{}^{t_j^P})\pi_j^R(s_i^j, s_j^i) + \delta_j{}^{t_j^P}\pi_j^*(s_i, s_j)$.
6. For a derivation, see Appendix IV.1.
7. Note that Farrell and Maskin's proof also includes showing that the two strategy combinations s and s^i are sufficient to generate any payoff tuple in Π^{IR} (see Farrell and Maskin 1989a, pp. 332ff.). Since this part of the proof is not important for the central idea of the Theorem 7.1, we skip it.
8. Probabilities p_1^1 and p_2^1 in (7.9) imply that player 1 accepts the punishment, probabilities p_1 and p_2 in (7.8) refer to the defection strategy of player 1.
9. It should be kept in mind that $\overline{\Pi}^{SPE} = \overline{\Pi}^{WRPE}$ does *not* imply $\Sigma^{SPE} = \Sigma^{WRPE}$. For instance, the trigger strategy which calls for an infinite punishment by reverting to the stage game NE in case of defection is not a WRPE because (for any $\pi_j^* > \pi_j^N$) the punisher receives less than in the cooperative phase. This violates condition (7.5).
10. Van Damme (1989) proves Theorem 7.4 using a particular example. However, acknowledging that any more general proof would require a great amount of notation and would be rather messy in the context of mixed strategies, we refrain from giving one.
11. For more than two-player games, SSPE requires more than subgame-perfection and Pareto-efficiency of all continuation payoffs. See Chapter 15.
12. Farrell and Maskin (1989a), Theorem 4, specify sufficient conditions for the existence of an SSPE.

8. Issue linkage

8.1 INTRODUCTION

So far we have considered only single-issue games. However, relations between governments may also concern several issues. For instance, governments may negotiate on the reduction of sulfur emissions and on the reduction of greenhouse gases at the same time, or talk about the creation of a free-trade area and disarmament issues simultaneously. Issue linkage may be able to *stabilize an agreement* in three respects: first and most importantly, issue linkage may lead to a more *symmetric distribution of the gains from cooperation*; second, issue linkage may ease the *enforcement of an agreement* – a government can threaten to withdraw from *all* treaties if a country free-rides with respect to one issue; third, an agreement regulating a *public good*, like a global pollutant, may be linked to an agreement regulating the provision of a *club good*. Since the benefits of the 'club good agreement' can be made exclusive to its members, requiring participants to hold a simultaneous membership in both agreements may be able to stabilize the public good agreement.

The first aspect is the most frequently treated in the literature and is referred to in the following as the 'enlargement of payoff space' (Section 8.2). The typical framework is that of a supergame and the following exposition draws on work done by Cesar and de Zeeuw (1996); Folmer *et al.* (1993); Kroeze-Gil and Folmer (1998); and Ragland *et al.* (1996). The effect of issue linkage in the context of finite games will be treated in Section 8.3. Whereas the classical framework (Sections 8.2 and 8.3) assumes separable utilities in the linked games, the discussion of the second aspect of issue linkage is particular interesting in the context of *non-separable utility functions*, which is treated in Section 8.4 based on Spagnolo (1996). In contrast to the classical framework, in which it can be shown that issue linkage is always beneficial for an agreement, in the context of non-separable utility functions this must not always be the case. It will turn out that the form of governments' objective function is crucial for the success of issue linkage.

The third aspect of issue linkage will be discussed in Chapter 13 on coalition formation.

Asymmetric payoffs may occur, for instance, in the acid rain context.

Upwind countries such as Great Britain which produce a great bulk of sulfur emissions in Europe have no incentive to contribute much to the reduction of sulfur since the benefits mainly accrue to the northern countries like Sweden, Norway and Finland. In other instances payoffs are even more unevenly distributed, such as the pollution of rivers, since there is typically a *pure* upstream–downstream relationship between countries.

Generally, upwind or upstream countries can only be convinced to contribute to a cooperative solution if the beneficiaries compensate the polluter(s) for emission reductions in the sense of Coase (1960). However, looking at the record of IEAs, it is striking that almost exclusively all of the agreements signed until the 1980s have no provisions for transfers.[1] Only the more recently signed IEAs, for instance the Rio Declaration in 1992,[2] the Montreal Protocol in 1987 on the depletion of the ozone layer, or the 1992 Convention on Biological Diversity, propose the implementation of international funds to which industrialized countries are supposed to contribute. These funds are designed to cover the 'incremental costs' which accrue to developing countries from agreed abatement efforts. However, the proposed financial commitments are negligible when contrasted to expected abatement costs. Moreover, all these IEAs have no provisions which allow the enforcement of transfer obligations. Therefore, it seems fair to claim that in reality the willingness to pay transfers of any kind are more lip-service than actual policy.

From an economist's point of view, the lack of transfer payments is puzzling since possible advantages are all too obvious. In the literature four arguments have been presented to explain this phenomenon:

1. Each party tries to hide its 'true' preferences for strategic reasons. In particular, upstream/upwind governments will exaggerate their abatement costs in order to extract high compensation payments from downstream/downwind governments (Mäler 1990).
2. There seems to be a widespread consensus in international politics that the polluter-pays principle instead of the victim-pays principle should be applied to tackle international environmental problems (United Nations 1972). There is the fear that if the victim-pays principle were applied, the polluters would reduce their efforts at a preventative environmental policy.
3. There is the danger that governments which pay compensation payments may be judged weak bargaining partners, which diminishes their bargaining power with respect to other issues (Mäler 1990).
4. Following Heister (1997, pp. 247ff.) transfers themselves may be interpreted as prisoners' dilemma games or as games with a similar incentive structure. For both countries, the payer and the payee, there exists

> a free-rider incentive. On the one hand, the payer prefers the payee to increase its abatement efforts without paying the transfer. On the other hand, the payee would like to receive the transfer without increasing its abatement obligations in turn. Thus if transfers are not accompanied by appropriate punishments they constitute no equilibrium strategies and cannot be used to stabilize an abatement game.[3]

Though these arguments help partially to resolve the above-mentioned puzzle, these explanations basically rely on *ad hoc* arguments rather than on scientifical and empirical foundations. However, in the following we do not try to answer the question as to why governments do not use transfers, that is, operate within a restricted framework voluntarily, but recognize that this phenomenon influences the strategy set of players. In particular, we focus on the *connection of games* in order to compensate for the lack of monetary transfers. In the non-environmental context, issue linkage can be observed in the form of barter in the international exchange of goods. Countries which lack foreign exchange, such as many eastern European countries, pay for their imports in kind instead of in hard currency.

In the environmental context some regional agreements used issue linkage as a means of compensation. Examples include the 1944 International Boundary Waters Treaty between the United States and Mexico (see Kneese 1988; and Ragland *et al.* 1996). The treaty allocates the water rights of the Colorado River, where the USA is the upstream country, and the lower portion of the Rio Grande River, where Mexico is the upstream country, such that both countries benefit. Krutilla (1975) also suggests that the Columbia River Treaty of 1961 between the USA and Canada – which viewed as a single issue was to the disadvantage of the USA – was built on concessions by Canada involving North American defense.[4]

Reviewing the examples of issue linkage cited in the literature it is apparent that typically only two countries are involved. This suggests that although, on the one hand, issue linkage may improve the chances for cooperation, on the other hand, it makes negotiations more complex. In particular, one should expect that, with an increasing number of countries, transaction costs become a limiting factor, so that the advantages of issue linkage may be lost.

In the following we abstract from transaction costs but their impact should be kept in mind. However, their existence is the motivation to restrict the following analysis to two countries and two issues only.

8.2 THE ENLARGEMENT OF PAYOFF SPACE

Matrix 8.1 Asymmetric PD game I

	a_2	na_2
a_1	3 −1	−1 2
na_1	4 −3	**0** **0**

Consider the asymmetric PD game I in Matrix 8.1, the payoff space of which is indicated by the broken rectangle in Figure 8.1. The asymmetry refers to a low cooperative payoff to one player, which may be lower than in the status quo ($a_i < d_i$, see the General Payoff Matrix 3.2). As assumed for the ordinary PD game in Section 3.2, mutual cooperation is globally optimal ($a_1 + a_2 > b_1 + c_1,\ c_1 + b_2,\ d_1 + d_2$), each country has a free-rider incentive ($c_i > a_i\ \forall\ i \in \mathrm{I}$) and unilateral investment does not pay ($b_i < d_i\ \forall\ i \in \mathrm{I}$). As in an ordinary PD game, the payoffs in the stage game Nash equilibrium (NE) correspond to the minimax payoffs which are normalized to zero. Thus $= \Pi^{\mathrm{IR}} = \{\pi \mid \pi_i \geq 0\ \forall\ i \in \mathrm{I}\}$ and from Figure 8.1 it is evident that as long as mixed strategies cannot be played, cooperation fails even in a supergame framework and all discount factors close to 1. To make the following analysis interesting and to stress the importance of issue linkage in international policy coordination, we rule out mixed strategies in the remainder of this chapter. Moreover, we assume simultaneous moves within a stage game.

Matrix 8.2 Asymmetric PD game II

	a_2	na_2
a_1	−1 3	−3 4
na_1	2 −1	**0** **0**

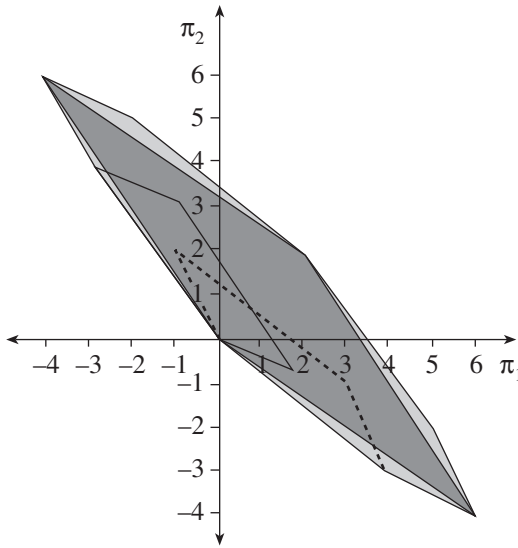


Figure 8.1 Issue linkage of asymmetric PD games I and II

Assume that there is a second issue with associated payoffs as given in Matrix 8.2. This game is the exact mirror image of asymmetric PD game I; however, as we shall show, this assumption is not crucial for the main result derived below. Again, cooperation in the isolated game is not possible since country 1 receives a payoff below its minimax payoff. However, if both issues are linked to each other, that is, the separate games are viewed as a single game, the payoff space is enlarged. The payoff space of the entire stage game is represented by Matrix 8.3 where the first subscript of a stage game strategy refers to the player and the second to the issue. In Figure 8.1 this payoff space comprises the lighter and darker shaded areas.

From Matrix 8.3 and Figure 8.1 it is evident that in the interconnected game the payoff tuple from mutual cooperation on both issues, that is, $\pi^* = (2, 2)$, is located in the positive quadrant and is therefore entailed in the individually rational payoff space. It is also apparent that if the countries agree to cooperate on both issues, a country has an incentive to deviate with respect to *both* issues simultaneously. Though a country improves upon its payoff by deviating with respect to one issue only, it gains even more by deviating on both issues. The harshest SPE punishment also involves stopping cooperation on both issues. The familiar trigger strategy delivers for the example $\delta_i^{\min} = 2/3 \; \forall \; i \in \mathrm{I}$, which stresses that for large enough discount factors mutual cooperation can be sustained as an SPE.

Matrix 8.3 Interconnected PD game

	a_{21}/a_{22}	a_{21}/na_{22}	na_{21}/a_{22}	na_{21}/na_{22}
a_{11}/a_{12}	2 2	0 3	−2 5	−4 6
a_{11}/na_{12}	5 −2	3 −1	1 1	−1 2
na_{11}/a_{12}	3 0	1 1	−1 3	−3 4
na_{11}/na_{12}	6 −4	4 −3	2 −1	**0** **0**

Of course, defining credible strategies with respect to renegotiation-proofness, an NE cannot be played as punishment. However, from Figure 8.1 it is evident that because in the interconnected game the payoff space Π is excessively larger than Π^{IR}, it should not prove difficult to sustain the cooperative solution as a weakly renegotiation-proof equilibrium (WRPE) for discount factors close to 1. This is particularly evident by noting the similarity of the payoff space of an ordinary PD game and that of a linked asymmetric PD game. As a thought experiment, connect payoff tuples (−4, 6), (0, 0), (6, −4) and (2, 2) in Figure 8.1 (darker shaded area). Then it is apparent that this 'imaginary payoff space' is that of a typical PD game and is *contained* in the entire payoff space of the linked game. Hence, Theorems 7.2 and 7.4 of Section 7.1 apply.[5] In fact, since the cooperative payoff tuple π^* lies on the Pareto frontier of the linked game (and of the imaginary payoff space), the result extends to *strongly renegotiation-proof equilibrium* (SRPE). This result should not come as a surprise since, as in the ordinary PD game, unilateral cooperation can be used in the linked game as a Pareto-efficient punishment.

From the example the natural question arises of how general the results obtained above are. To answer this question consider the General Payoff Matrix of the isolated game in Matrix 8.4 and of the linked game in Matrix 8.5 where the subscript $k \in \{1, 2\}$ denotes the game. The minimax payoffs

which are the sums of minimax payoffs to the players in the isolated games have been normalized to zero.

Matrix 8.4 Payoffs in the isolated PD game

	a_{2k}	na_{2k}
a_{1k}	a_{1k} a_{2k}	b_{1k} c_{2k}
na_{1k}	c_{1k} b_{2k}	**0** **0**

Matrix 8.5 Payoffs in the interconnected PD game

	a_{21}/a_{22}	a_{21}/na_{22}	na_{21}/a_{22}	na_{21}/na_{22}
a_{11}/a_{11}	$a_{11}+a_{12}$ $a_{21}+a_{22}$	$a_{11}+b_{12}$ $a_{21}+c_{22}$	$b_{11}+a_{12}$ $c_{21}+a_{22}$	$b_{11}+b_{12}$ $c_{21}+c_{22}$
a_{11}/na_{11}	$a_{11}+c_{12}$ $a_{21}+b_{22}$	a_{11} b_{21}	$b_{11}+c_{12}$ $c_{11}+b_{22}$	b_{11} c_{21}
na_{11}/a_{12}	$c_{11}+a_{12}$ $b_{21}+a_{22}$	$c_{11}+b_{12}$ $b_{21}+c_{22}$	a_{12} a_{22}	b_{12} c_{22}
na_1/na_1	$c_{11}+c_{12}$ $b_{21}+b_{22}$	c_{11} b_{21}	c_{12} b_{21}	**0** **0**

The Nash equilibrium payoffs are printed in bold. For an asymmetric PD game, which may be seen as the more general case of a PD game, the following relations hold:

1. $c_{ik} > a_{ik}$;
2. $0 > b_{ik}$;
3. $a_{1k}+a_{2k} > b_{1k}+c_{2k}$, $a_{1k}+a_{2k} > c_{1k}+b_{2k}$ and $a_{1k}+a_{2k} > 0 \; \forall \; i \in \mathrm{I}$ and k.

An implication of (3) is that $a_{ik} > 0$ must hold for at least one country and from (1) and (2) $\Sigma_{i=1}^{2} c_{ik} > \Sigma_{i=1}^{2} a_{ik} > \Sigma_{i=1}^{2} b_{ik}$ $\forall$ k follows. Now we claim the following:

Proposition 8.1
In an infinitely repeated PD game with payoff relations as defined above, the possibility of issue linkage either improves upon the chances that mutual cooperation can be sustained as a strongly renegotiation-proof equilibrium or leaves them unchanged.

Proof: In Theorem 8.1 cooperation implies the strategy combination (a_{1k}, a_{2k}) with associated payoff tuple (a_{1k}, a_{2k}). In order to sustain cooperation as an SPE in an isolated game k, a *necessary* and *sufficient* condition is:

$$a_{ik} - (1 - \delta_i)c_{ik} \geq 0 \ \forall \ i \in \mathrm{I} \text{ and } k \in \{1, 2\} \tag{8.1}$$

whereas in the linked game cooperation requires:

$$\sum_{k=1}^{2} a_{ik} - (1 - \delta_i) \sum_{k=1}^{2} c_{ik} \geq 0 \ \forall \ i \in \mathrm{I} \tag{8.2}$$

where, obviously, condition (8.2) is implied by (8.1) and hence condition (8.2) is more easily satisfied than (8.1) (for any $0 \leq \delta_i \leq 1$).[6] Since we demonstrated above that the linked game possesses a kind of PD payoff structure[7] $\delta_i^{\min}(\mathrm{SPE}) = \delta_i^{\min}(\mathrm{WRPE})$ holds by Theorem 7.4 (for $\delta_i \to 1$ Theorem 7.2 applies). Moreover, since we assumed mutual cooperation to be globally optimal $\delta_i^{\min}(\mathrm{SPE}) = \delta_i^{\min}(\mathrm{SRPE})$ is true. QED

It should be evident that Theorem 8.1 can easily be extended to cover the case where more than two PD games are linked. Linking issues either improves the chances for cooperation or leaves them unchanged. Note also that it is not necessary that in each game the strategy tuple (a_{1k}, a_{2k}) is played as a cooperative strategy. It is also conceivable that other strategy combinations are played as long as they deliver individually rational payoff tuples in the linked game. However, in this case Theorem 8.1 can only be stated with respect to SPE strategies because it is not possible to tell at a general level whether such a payoff tuple lies within the 'imaginary PD payoff space' of the linked game.

Proposition 8.2
In an infinitely repeated PD game the possibility of issue linkage improves the chances that cooperation can be sustained as a subgame-perfect equilibrium or leaves them unchanged.

Proof: Let $\pi_{ik}(s_k) > 0$ denote the payoff to player i from some cooperative strategy s_k in game k, and $\pi_{ik}^{D}(s_{ik}(s_{jk}), s_{jk})$ the deviation payoff in the stage game ($\pi_{ik}^{N}(s_k^N) = 0$). Then cooperation can be sustained in an isolated game provided:

$$\pi_{ik}(s_k) - (1 - \delta_i)\pi_{ik}^{D}(s_{ik}(s_{jk}), s_{jk}) \geq 0 \; \forall \; i \in \mathrm{I} \text{ and } k \in \mathrm{K} \tag{8.3}$$

holds. In contrast, in the linked game cooperation requires:

$$\sum_{k=1}^{\mathrm{K}} \pi_{ik}(s_k) - (1 - \delta_i) \sum_{k=1}^{\mathrm{K}} \pi_{ik}^{D}(s_{ik}(s_{jk}), s_{jk}) \geq 0 \; \forall \; i \in \mathrm{I} \tag{8.4}$$

where the number of games is K and (8.4) is again less restrictive than (8.3). QED

Conditions (8.5) and (8.6) stress the main reason why issue linkage has a positive effect on the possibilities for cooperation in infinite games: it is simply because the payoff space is enlarged. Of course, by varying some of the conditions above, results in the spirit of Theorems 8.1 and 8.2 may also be derived for other classes of games.

8.3 THE IMPACT ON STAGE GAME NASH EQUILIBRIA

Whereas in the context of infinite games the main focus of issue linkage was on the enlargement of the payoff space, in finite games the focus is on the number and the nature of stage game NE. Of course, also in finite games issue linkage may lead to more symmetric payoffs so that cooperation becomes attractive to all participants. However, since in finitely repeated games special requirements with respect to the stage game NE must be satisfied to establish cooperation, it seems more interesting to analyze how issue linkage affects these requirements. Recall that for SPE strategies two stage game NE – a good and bad NE – are sufficient to sustain a cooperative solution for sufficiently large discount factors whereas for renegotiation-proof strategies at least two Pareto-undominated stage game NE are required. Consequently, the main concern is to answer the question of how many stage game equilibria result from the linkage of two games.

First, note that the stage game NE of the linked game are the strategy combinations of the NE of the isolated games. This is so because it can never be a best reply in a linked game to play an NE strategy, say, in game I but not in game II. Hence, if games I and II each have a single stage game NE, the linked game will also have only a single NE. Thus, a necessary

condition to establish cooperation in a linked game is that one of the isolated games must have two stage game NE.[8] For instance, suppose we link the extended PD game I in Matrix 4.2, which has only one NE, with the extended PD game II, which has two NE. Then, there are two NE in the linked game, that is, $\{(na_{11}, na_{12}), (na_{21}, na_{22})\}$ and $\{(na_{11}, na_{12}), (p_{12}, p_{22})\}$, with associated payoff tuples (4, 4) and (3, 3). Thus, mutual cooperation can be sustained in the linked game as an SPE using the first as a good and the second as a bad NE provided discount factors are sufficiently high, though cooperation would not be possible in the isolated extended PD game I.

To establish cooperation in a linked game as an RPE we require two Pareto-undominated stage game NE and because of that at least one of the two single games must possess two of such equilibria on its own. For instance, if we link the extended PD game I and the extended PD game IV in Matrix 6.2, where the latter has two Pareto-undominated Nash equilibria, the linked game will also have two such equilibria. Thus cooperation may be sustained in the linked game as an RPE, though this is not possible in the extended PD game I.

Moreover, note that if we link two extended PD games IV we get four NE with payoff tuples (4, 8), (8, 4), (6, 6), (6, 6). Whereas in the isolated game it was necessary to play a three-stage sequence due to the lack of a symmetric NE (assuming $\delta_i = 1 \; \forall \; i \in \mathrm{I}$), now a two-stage sequence (as in the extended PD game III) is sufficient. This increases the average payoff to a player for large T from 4 to 4.5.

From the examples it is evident that in the context of finite games issue linkage improves upon the conditions for cooperation or leaves them unchanged since, according to the assumptions so far, linking occurs voluntarily. Since the results are intuitive and confirm our previous findings in Section 8.2, namely that issue linkage has a positive effect on the possibilities for cooperation, we refrain from stating theorems and giving proofs in this section. However, there is a slight difference between the effect that issue linkage has on the payoff space and on Nash equilibria. In Section 8.2 it became apparent that though *both* single games had *no* individually rational payoff vector, the linked game had at least one. Now in the context of finite games *at least one* of the games must possess the necessary properties with respect to the NE to establish cooperation in the linked game. In the case of SPE strategies at least one game must have a good and a bad NE; in the case of RPE at least one game must have two Pareto-undominated NE in order to establish cooperation in the linked game.

8.4 NON-SEPARABLE UTILITY FUNCTIONS

8.4.1 Introduction

In the previous sections we implicitly assumed that agents have separable utility functions. That is, though they may link and negotiate two issues together, they value the payoffs of each game independently. Of course, this might indeed be the case and this is also the standard assumption of the literature on issue linkage. However, a more natural assumption seems to be that governments' evaluation of an issue depends also on other issues. In this case, the effect of issue linkage is less straightforward and relies on the shape of governments' objective functions. To concentrate on the effect of issue linkage in the presence of non-separable utility functions we assume two *infinitely* repeated ordinary PD games. Hence, as laid out above, we can apply Theorem 7.4. That is, the minimum discount factor requirement derived from a subgame-perfect trigger strategy also delivers $\delta_i^{\min}$(WRPE).

The two PD games are called for short games I and II. We assume that the underlying 'physical payoffs' are those of Matrix 8.4 but that they are evaluated through the utility function u_i. If the utility function were separable in issues (or if the second issue did not exist), the 'transformed' payoffs would be given for game k by Matrix 8.6. If, however, utility is non-separable in the two issues, payoffs are given for game I by Matrix 8.7. $x_{12} \in \Pi_{12} = \{a_{12}, b_{12}, c_{12}, 0\}$ and $x_{22} \in \Pi_{22} = \{a_{22}, b_{22}, c_{22}, 0\}$ refer to possible payoffs in game II to players 1 and 2 respectively. Thus the notation implies that the first subscript refers to the player, the second to the game. By symmetry we have $x_{11} \in \Pi_{11} = \{a_{11}, b_{11}, c_{11}, 0\}$, $x_{21} \in \Pi_{21} = \{a_{21}, b_{21}, c_{21}, 0\}$ in game II which is represented by Matrix 8.8.

Matrix 8.6 Game k*: single issues*

	a_{2k}	na_{2k}
a_{1k}	$u_1(a_{1k})$ $u_2(a_{2k})$	$u_1(b_{1k})$ $u_2(c_{2k})$
na_{1k}	$u_1(c_{1k})$ $u_2(b_{2k})$	$u_1(0)$ $u_2(0)$

Matrix 8.7 Game I: non-separable utility

	a_{21}	na_{21}
a_{11}	$u_1(a_{11}, x_{12})$ $u_2(a_{21}, x_{22})$	$u_1(b_{11}, x_{12})$ $u_2(c_{21}, x_{22})$
na_{11}	$u_1(c_{11}, x_{12})$ $u_2(b_{21}, x_{22})$	$u_1(0, x_{12})$ $u_2(0, x_{22})$

Matrix 8.8 Game II: non-separable utility

	a_{22}	na_{22}
a_{12}	$u_1(a_{12}, x_{11})$ $u_2(a_{22}, x_{21})$	$u_1(b_{12}, x_{11})$ $u_2(c_{22}, x_{21})$
na_{12}	$u_1(c_{12}, x_{11})$ $u_2(b_{22}, x_{21})$	$u_1(0, x_{11})$ $u_2(0, x_{21})$

For simplicity we shall focus on symmetric strategy combinations in the following. That is, when analyzing the possibilities for cooperation we assume (a_i, a_j). This symmetry also implies that x_{1k} will be either a_{1k} or 0 and x_{2k} will be either a_{2k} or 0. Moreover, it will turn out to be convenient to normalize utility such that $u_i(0) = 0$ and $u_i(x_{ik}, 0) = u_i(x_{ik})$. The utility of the linked game can be constructed as a 4×4 matrix based on Matrices 8.7 and 8.8 similar to Matrix 8.5. Since the basic procedure is the same as described above we skip the derivation of such a matrix here.

For the properties of the utility function we make the standard assumptions, namely $\partial u_i/\partial x_{ik} > 0$, $\partial^2 u_i/\partial x_{ik}^2 < 0$ $\forall$ i and k. Moreover, we define:

Definition 8.1
Issues are substitutes in agents' objective functions if $\partial^2 u_i/\partial x_{i1}\partial x_{i2} < 0$ and they are complements if $\partial^2 u_i/\partial x_{i1}\partial x_{i2} > 0$ holds.

That is, if issues are substitutes, additional utility derived from increased physical payoffs in game I will be lower if the physical payoffs of game II are already high than if they are low. By symmetry, the same holds for game II. In contrast, if issues are complements, marginal utility will be higher in game I if the payoffs in game II are also high. If $\partial^2 u_i/\partial x_{i1}\partial x_{i2} = 0$ we are led back to the case we discussed in previous sections.

8.4.2 Possibilities of Cooperation and Delegation of Policy Coordination

Basically, there are two main cases to be considered. First, linking a new issue to an existing issue: that is, issue 1 already exists and has already been negotiated. After the negotiation the second issue emerges. The new issue

may or may not be linked to the first issue. Second, 'linking two existing issues': that is, governments first seek cooperation on each issue separately and then consider linking both issues in negotiations.

In both cases four subcases have to be distinguished.

1. Cooperation can be sustained on each single issue.
2. Cooperation cannot be sustained on issue 1, but on issue 2.
3. Cooperation can be sustained on issue 1 but not on issue 2.
4. Cooperation on both issues cannot be sustained in the isolated games.

Now one can show the following:

Proposition 8.3
If two infinitely repeated (ordinary) prisoners' dilemma games are linked to each other, mutual cooperation becomes easier to sustain as a strongly renegotiation-proof equilibrium than in the isolated games, provided issues are substitutes in governments' objective functions. The opposite holds if issues are complements in governments' objective functions.

Proof: See Appendix V. QED

Roughly speaking, the reason for this result is the following. Once a country deviates in the interconnected game it is punished with respect to both issues. If issues are substitutes, the loss through the punishment is particularly severe as the marginal utility from cooperation is high at low payoff levels. Moreover, if countries are cooperating and receiving a relatively high level of utility, free-riding pays less than at lower levels of utility. Thus, though in the linked game a country deviates with respect to both issues and is punished by terminating cooperation on both issues, the latter effect is stronger than the former, implying additional enforcement power in the linked game. By symmetry, exactly the opposite holds if issues are complements.

Consequently, if issues are complements an obvious countervailing measure would be to separate issues. This could be done if a government delegates decision-making power on one or both issues to separate independent agencies. For instance, in Germany monetary policy is independently conducted by the central bank, economic competition is enforced by an antitrust agency, and economic and fiscal policy is conducted by the government itself. Of course, a basic prerequisite for such an 'isolation strategy' to be successful is that the delegation contract must be based on a long-term relationship between the government and the agencies; otherwise, delegation is not credible and can always be reversed. In the

above-mentioned example this long-term relationship is ensured by the constitution.

Of course, by the same token if issues are substitutes delegation of decision-making power would have a negative effect on policy cooperation. This may be summarized as follows:

Proposition 8.4
In an infinitely repeated PD game, if issues are substitutes in governments' objective function long-term delegation of decision-making power on one or two policy issues to independent agencies with the same objective function as the government will make it more difficult to sustain cooperation. If issues are complements, delegation will make cooperation easier.

Proof: Is obvious and therefore omitted.

The result of Proposition 8.4 is based on the assumption of a long-term relationship between principal and agent. Note that this is not an *ad hoc* assumption but, in fact, an essential implication of our previous assumption of an infinite game. Hence, stating Proposition 8.4 without proof requires the assumption that the agency also views policy coordination as an infinite game and that the principal–agent relationship can also be seen as an infinite game.

It is interesting to note that Proposition 8.4 holds even though the objectives of the government and the agency are the same with respect to the delegated issue. This contrasts with the typical assumption in the principal–agent literature where such a result is generated due to the fact that delegation distorts the objective function or the incentives of the agent. However, if utility functions between government and agency differ, then the following rule applies:

Proposition 8.5
Suppose that the relation between two governments is described by an infinitely repeated PD game. Then if a government decides to delegate decision-making power to an agency and has the option to choose between several agencies, then it should select that agency with the most concave utility function to sustain international cooperation.

Proof: Assume material payoffs of the General Payoff Matrix 3.2. Then cooperation can be sustained in the infinitely repeated game k provided:

$$\delta_i \geq \delta_i^{\min} = \frac{c_{ik} - a_{ik}}{c_{ik} - d_{ik}} \tag{8.5}$$

holds. For any monotone transformation this condition reads:

$$\delta_i \geq \delta_i^{\min\#} = \frac{\Omega_i(c_{ik}) - \Omega_i(a_{ik})}{\Omega_i(c_{ik}) - \Omega_i(d_{ik})} \tag{8.6}$$

where Ω_i is agency i's utility function. This transformation will ease cooperation if $\delta_i^{\min} \geq \delta_i^{\min\#}$. Using (8.5) and (8.6) this implies:

$$\frac{\Omega_i(a_{ik}) - \Omega_i(d_{ik})}{a_{ik} - d_{ik}} \geq \frac{\Omega_i(c_{ik}) - \Omega_i(d_{ik})}{c_{ik} - d_{ik}}$$

after rearranging terms, which is true since concavity implies $\partial^2\Omega_i/\partial x_{ik}^2 < 0$. QED

Taken together, the examples show that, in the context of non-separable utility functions of governments, the implications of issue linkage are less straightforward than in the 'classical' approach. Taking these complications into consideration, then, issue linkage is only conducive to cooperation if the issues are viewed as substitutes by governments.

NOTES

1. Exceptions include the North Pacific Seal Treaty signed in 1957 which requires the USA and the former USSR to pay Canada and Japan 15 percent of their annual harvest of pelts. Another instance is the 1972 agreement on the reduction of the salinity of the Rhine, where the Netherlands agreed to compensate France for 35 percent of its costs.
2. The Rio Declaration comprises mainly the Climate Framework Convention in which states declare their concern about the global warming problem. Therefore, it may be viewed mainly as a declaration of good will rather than an actual IEA.
3. Basically, this implies that the abatement game is extended to include transfer strategies, where each strategy set may be interpreted as belonging to a separate game. Heister (1997) calls the combinations of several games, like the combination of abatement and transfer games, hypergames. According to this definition the issue linkage games considered below, where two no-transfer games are linked, could also be termed hypergames.
4. Other interesting examples where issue linkage has played some role may be found in Ragland *et al.* (1996).
5. Note that it is important that the cooperative payoff tuple lies on the boundary of this imaginary payoff space. Other individually rational payoff tuples of the linked game may lie outside this payoff space and hence Theorems 7.2 and 7.4 *cannot* be applied.
6. There are three cases to consider: (1) Inequality (8.1) is satisfied for both isolated games, consequently, inequality (8.2) will hold as well; issue linkage does not improve upon the chances for cooperation but also does no harm. (2) Inequality (8.1) does not hold for both isolated games, consequently, inequality (8.2) is not satisfied either; again, issue linkage has no effect. (3) Inequality (8.1) is satisfied for one issue, but not for the other; this leads to two subcases: (3a) inequality (8.2) is also not satisfied. In this case cooperation is not possible in the linked game but countries can cooperate on that issue for which (8.1) holds; (3b)

inequality (8.2) is satisfied and cooperation can be sustained with respect to both issues. Due to this last subcase Theorem 8.1 holds.

7. In the general case, this is an implication of $\Sigma_{i=1}^{2} c_{ik} > \Sigma_{i=1}^{2} a_{ik} > \Sigma_{i=1}^{2} b_{ik}$ $\forall$ k as derived above.
8. We assume that each game possesses at least one NE.

9. Static games with continuous strategy space: global emission game

9.1 INTRODUCTION

Up to now we have assumed that players have discrete action sets so that the normal form of the game could conveniently be displayed in a matrix. Though for many policy problems modeling decisions as a discrete choice seems adequate, other situations may be better modeled as a continuous choice problem such as the amount of emission reduction in a global policy game, for example, greenhouse gases (see also the discussion in Section 2.3).

A continuous strategy set allows finer tuning of actions and reactions and therefore leads to some interesting results which are absent in discrete policy games. This is true at least as long as mixed strategies are ruled out for discrete policy games. Though we dealt with mixed strategies in Chapter 3 and also mentioned some instances in which one can expect players to use mixed strategies, they were introduced mainly for technical reasons; that is, mixed strategies were required in the discrete strategy context to capture the entire feasible payoff space when deriving folk theorem type results. In a continuous strategy setting it suffices to consider only pure strategies. This is true at least as long as games with a convex payoff space are considered. Since all games in the remainder of this book satisfy this condition, we no longer have to bother about mixed strategies.

The following analysis is based on a rather simple emission model. In particular, the payoff or net benefit functions contain emissions as the only argument. Thus, the aspect of issue linkage as well as agents' choice between environmental quality and other goods is not considered. The reasons are the following:

1. Issue linkage has been already extensively covered in Chapter 8 and a continuous strategy set would not add substantially new aspects to the problem.
2. Payoff functions containing more than one argument do not affect the main conclusions derived in the subsequent analysis.

That is, the strategic interplay of players with respect to international environmental issues would remain unchanged. Although, of course, such an extended framework might be more convincing from an economic point of view, it would only complicate the analysis.[1] Only if 'joint production' were considered, *could* the results of the analysis change. Joint production implies that the consumption or production of private goods is directly related to environmental quality. That is, a (representative) consumer's choice concerns not only how much money to allocate to the purchase of private goods versus the public good 'environment', but the consumption of private goods affects environmental quality and vice versa. If joint products are substitutes (the consumption of the private good causes environmental damage which, *ceteris paribus*, decreases the demand for the private good) the main conclusions of the simple model continue to hold. Only if joint products are complements (for example, an increase in carbon emissions, which increases the demand for refrigerators, which in turn creates more global warming and so on), this may partially lead to qualitatively different results (see Sandler 1996). Though such a complication could be handled in a static framework, treatment in a dynamic context would certainly be beyond the scope of this book.[2]

Since 'several-goods models' are typically set up as maximization problems subject to budget constraints, the influence of a *change* in income and a *redistribution* of income on equilibrium emissions can be analyzed. Whereas the first aspect can easily be integrated in the simple model by interpreting an increase in income as equivalent to a reduction in abatement costs, transfers can only be analyzed by assuming that the tastes of players (or the willingness to pay for environmental quality) are unaffected by these transfers. Though this is certainly a disadvantage of the model used here, working with a simple model at the present stage allows us to relate the setting of this chapter to the more advanced game theoretical later chapters which require, by their nature, such a simplification anyway.

Another simplification worth mentioning is that the subsequent analysis is restricted to a *global emission game*. An extension to cover *transboundary pollutants* – though it is conceptually straightforward – would require us to deal with additional variables, namely transportation coefficients reflecting different spillover patterns between countries. Since most qualitative results remain valid for transboundary pollutants also, this extension is not considered in what follows. Extension may be found, for example, in Kuhl (1987) and Nentjes (1994).

As in the previous chapters, we take a stepwise approach. We start the analysis in this chapter in a static framework, which we shall continue to assume in parts of Chapters 10 and 11. We also integrate two-stage games into the analysis in Chapters 10 and 11. Finally, we extend the analysis to

an infinite dynamic framework in Chapters 12 and 14. Due to the complexity, we restrict the discussion of infinite games in a first step to two countries (Chapter 12) and extend the analysis in a second step to $N>2$ countries (Chapter 14). Though the analytical parts of Chapters 9 to 10 will mainly cover N countries, the graphical illustrations are mainly confined to two countries for expositional simplicity. Chapters 11 and 12 are exclusively restricted to two countries but the bargaining solutions analyzed in these chapters will form the basis for the coalition model laid out in Chapter 14. Until Chapter 13, the strategic aspect of coalition formation will be discarded. Chapter 13 presents simple static or two-stage coalition models, whereas Chapter 14 discusses coalition formation in a supergame framework. Chapter 15 presents some new theoretical concepts to analyze the process of coalition formation.

9.2 FUNDAMENTAL FUNCTIONS AND ASSUMPTIONS

9.2.1 General Case

Let the payoff function (net benefit function) of country i, π_i, comprise the benefits from emissions, $\beta_i(e_i)$, and the damages caused by global emissions, $\phi_i(\Sigma e_i)$. In particular assume:

$$\pi_i = \beta_i(e_i) - \phi_i\left(\sum_{j=1}^{N} e_j\right) \;\forall\; i \text{ and } j \in N. \tag{9.1}$$

Since π_i describes the welfare implication from emissions at an aggregate level, emissions may be viewed as an input to the production and consumption of goods from which benefits $\beta_i(e_i)$ are derived (Welsch 1993). Emissions in country i, e_i, may be a production or/and a consumption externality causing environmental damage in country i and also in the other countries. What only matters for damage are aggregate emissions, Σe_i, due to the assumption of a global externality (pure public bad; emissions disperse uniformly in the atmosphere).

The following properties of the benefit and damage cost functions are henceforth assumed, if not stated otherwise, though some implications if they are violated will be discussed:

$$\begin{aligned} &A_1\text{: } \phi_i' \geq 0 \text{ and } \phi_i'' \geq 0 \;\forall\; \Sigma e_i \geq 0,\ \phi_i(0) = 0;\ \pi_i'' < 0 \Rightarrow \beta_i'' < \phi_i'' \;\forall\; e_i \geq 0 \\ &\beta_i' \geq 0 \;\forall\; 0 \leq e_i \leq e_i^0,\ \beta_i' <> = 0 \;\forall\; e_i > e_i^0,\ \beta_i'' \leq 0 \;\forall\; e_i \geq 0,\ \beta_i(0) = 0. \end{aligned} \tag{9.2}$$

The assumptions regarding the *first derivative of the benefit functions* imply that benefits increase in emissions up to a level e_i^0. If emissions are further

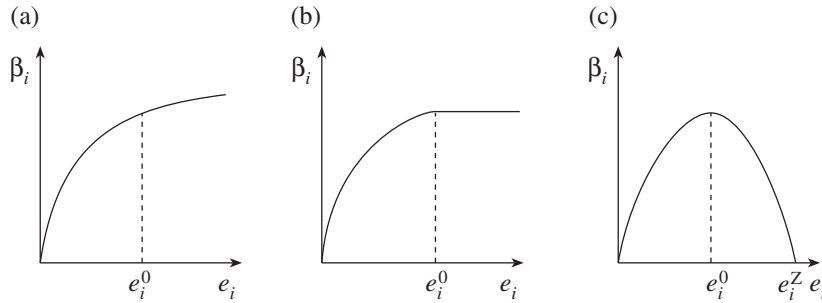


Figure 9.1 Possible curvatures of benefit functions

increased, benefits may either continue to increase (Figure 9.1a), remain constant (Figure 9.1b) or decrease (Figure 9.1c). The example which is used below exhibits the curvature of Figure 9.1c, though it turns out that the properties of β_i' above e_i^0 are not important.

The benefit function in Figure 9.1c has the advantage that it suggests two natural candidates for the upper bound of a player's strategy space, $e_i^{\max}$, namely e_i^0 or e_i^Z. Apart from the need to specify an upper bound of the strategy space in order to define the normal form of the emission game, the upper bound determines the threat and punishment options (see Section 9.5). An argument in favor of $e_i^{\max} = e_i^0$ is that even if damages are ignored at all, governments would not choose emissions above e_i^0. Provided one accepts harsher punishments, an upper bound for a credible threat could be e_i^Z since emissions above this level would not only imply high damages but also negative benefits.

In contrast, the assumption regarding the *second derivative of the benefit function* will turn out to be crucial for the relations investigated below. $\beta_i'' < 0$ reflects the standard assumption of decreasing marginal economies of scale in the production and/or decreasing marginal utility in the consumption of goods. However, even if $\beta_i'' > 0$ were allowed (for example, due to economies of scale), (strict) concavity of the *net benefit function* with respect to own emissions would still be ensured as long as $\beta_i'' < \phi_i''$.[3] Apart from being a sufficient condition for a unique global maximum with respect to the own decision variable, it will become apparent that this condition also determines the slope of countries' reaction functions.

Note that the *benefit functions* may alternatively be interpreted as the *opportunity costs of abatement* (where higher abatement levels imply lower emissions). Accordingly, the opportunity cost of abatement increases at an increasing rate at higher reduction levels. This assumption is also standard in the environmental economics literature (for example, Baumol and Oates

1990; Endres 1994; Turner *et al.* 1994) and acknowledges the fact that raising emission removal levels requires successively sophisticated and costly abatement technologies.

With respect to the *damage function* assumptions, A_1 describe the fact that damage increases in emissions at an increasing rate. The last property reflects the diminishing self-purification of environmental systems at higher rates of contamination. Of course, there might be an upper bound of aggregate emissions above which environmental systems collapse. In the following, however, it is assumed that such an upper bound lies well above $\Sigma e_i^{\max}$ and therefore its implications for equilibrium considerations can be discarded.

Another interpretation of $\phi_i'' > 0$ follows if the evaluation of environmental damage is interpreted as society's willingness to pay for emission reductions. Then at higher levels of aggregate emissions this willingness to pay is higher than at low levels and decreases more than proportionally if environmental quality improves due to lower emissions.[4, 5]

9.2.2 Example

In order to demonstrate the (explicit) derivation of the Nash equilibrium (NE) and other benchmarks in a global emission game, we consider an example with the following specification of the net benefit functions:

$$\pi_i = b_i\left(de_i - \frac{1}{2}e_i^2\right) - \frac{c_i}{2}\left(\sum_{j=1}^{N} e_i\right)^2;\ b_i > 0,\ c_i > 0,\ d > 0\ \forall\ i \text{ and } j \in \mathrm{I} \quad (9.3)$$

where we assume for simplicity symmetric countries in the following, that is, $b_i = b_j = b$ and $c_i = c_j = c\ \forall\ i$ and j, $i \neq j$, and restrict the number of players to two, that is, $N = 2$. It is easily checked that the net benefit and the damage cost functions possess the properties of assumption A_1. In particular, we have $\beta_i' \geq 0\ \forall\ 0 \leq e_i \leq e_i^0 = d$, $\beta_i' < 0\ \forall\ e_i > e_i^0 = d$. We restrict the strategy space of player i to $S_i = E_i = [0, d]$, that is, $e_i^{\max} = e_i^0 = d$.

9.3 BEST REPLY FUNCTIONS, NASH EQUILIBRIUM AND PARAMETER VARIATIONS

The derivation of the *Nash equilibrium* proceeds in two steps. In the first step the best reply of each country i for *given* strategies of the other players is established. In the second step that strategy tuple for which *no* country likes to modify its choice given the *optimal strategies* of the other countries

is determined. This is an immediate application of the NE concept as defined in Chapter 3. For illustrative purpose, we first derive the NE for the example ((9.3)) and subsequently discuss the more general case ((9.1)).

9.3.1 Example

The *best response* or *reaction functions* for the example are derived by differentiating (9.3) with respect to e_i, setting the first derivative equal to zero and solving for e_i (assuming $N=2$, $b_1=b_2=b$, $c_1=c_2=c$):

$$r_i: = e_i(e_j) = \frac{bd - ce_j}{b+c} \Leftrightarrow \mathrm{E}_i = \frac{\gamma d - e_j}{\gamma + 1};\ \gamma = b/c \text{ and } b > c. \tag{9.4}$$

It will turn out that most of the later results can be expressed in terms of γ, which may be interpreted as a parameter describing the *benefit–cost ratio from emissions*. From (9.4) it is evident that the best reply of country *i* depends on the strategy of country *j* and hence *no* equilibrium in dominant strategies exists. Country *i*'s optimal emission level is negatively related to emissions in country *j* due to the global character of the negative externality. The best reply functions of countries 1 and 2, r_1 and r_2, are drawn in Figure 9.2.

Both best reply functions start from $e_i=(bd)/(b+c)$ where $e_j=0$. The

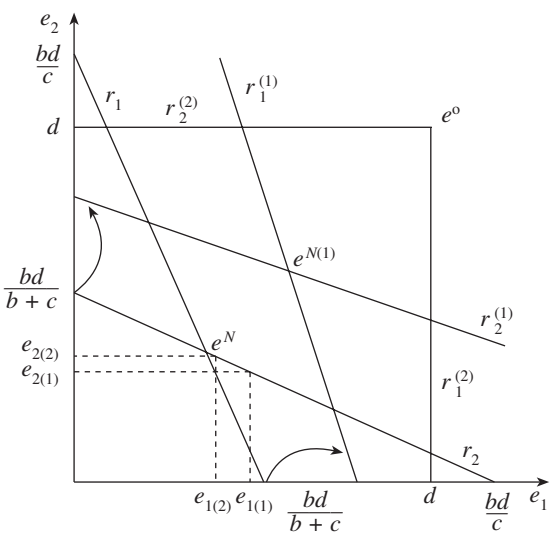


Figure 9.2 Nash equilibrium and parameter variations in the global emission game

optimal choice of e_i becomes smaller at higher levels of e_j. For $e_j=(bd)/c$, $e_i=0$. One possibility, to rule out negative emissions as a best reply, that is, $r_i(e_j)\geq 0$ (since without further assumptions it is not clear whether $e_i^{\max}=e_i^0=d$ is greater or smaller than $(bd)/c$), is to assume $b\geq c$ or $\gamma\geq 1$, as is done in (9.4). Alternatively, one could define piecewise reaction functions:

$$r_i(e_j):=\begin{cases} e_i=\dfrac{\gamma d-e_j}{\gamma+1} & \forall e_j\leq\gamma d \\ e_i=0 & \forall e_j>\gamma d \end{cases}. \tag{9.5}$$

This would imply that reaction functions do not stop at bd/c but continue along the e_1 and e_2 axes in Figure 9.2 (not drawn).

From Figure 9.2 it appears that the NE emission tuple, e^N, is determined as the intersection of the best reply functions r_1 and r_2. Due to the assumption of complete information, each player can form expectations about the best reply of the other player and s/he knows that the only stable strategy tuple is e^N. To see this, consider that country 1 chooses $e_{1(1)}$ instead of e_1^N. The best reply would have country 2 choosing emission level $e_{2(1)}$. This in turn would motivate country 1 to correct its previous choice, $e_{1(1)}$, to $e_{1(2)}$. Continuing with this kind of reasoning one derives at $e_{2(2)}$, and, finally, as with a cobweb, the NE is reached. This method neatly stresses the logic behind the NE as the result of the *convergent expectations of rational players* (see Section 3.2).[6]

Mathematically, the NE is derived by substituting the reaction functions mutually in each other and solving for the remaining variable. The following is obtained:

$$e_i^N=\frac{bd}{b+2c}\Leftrightarrow e_i^N=\frac{\gamma d}{\gamma+2} \tag{9.6}$$

which upon substitution in the net benefit functions (9.3) delivers:

$$\pi_i^N=\frac{b^3d^2}{2(b+2c)^2}. \tag{9.7}$$

It can easily be seen from (9.6) that as long as environmental damage is not neglected, that is, $c>0$ in the example, $e_i^N<e_i^0=d\ \forall\ i\in \mathrm{I}$. In particular, from (9.4) (or (9.5)) it is evident that even if country j chose $e_j=0$, the maximum emission level of country i would only be $e_i=bd/(b+c)<d$.

With the help of (9.4) and (9.6) the implications of a variation of the benefit–cost ratio γ on equilibrium emissions can be analyzed. Differentiating NE emissions with respect to γ, we find $\partial e_i^N/\partial\gamma>0$. That is, the higher the opportunity cost of abatement compared to environmental

damage, the higher will be equilibrium emissions. In this case reaction functions move outward, as indicated by the arrows in Figure 9.2, with the 'new' reaction functions $r_i^{(1)}$ and the 'new' Nash equilibrium $e^{N(1)}$. Thereby, the movement of the reaction functions may be broken down into two parts. First, the starting point of the reaction function ($e_i = bd/(b+c)$, $e_j = 0$) of country 1 (country 2) moves to the right (up) on the abscissa (ordinate).[7] Second, the slope of the reaction function becomes less steep since environmental damages are now valued at less than the opportunity costs of abatement. This follows from:

$$r_i' = \partial e_i / \partial e_j = -1/(\gamma + 1) \Rightarrow -1 < r_i' < 0. \tag{9.8}$$

from which it is evident that if either abatement is extremely costly (b is large) and/or environmental damages are neglectable (c is small) the optimal strategy of a country is not sensitive to emissions in the neighboring country. Hence, at the limit, as γ goes to infinity, both reaction functions are *orthogonal* to each other (see functions $r_i^{(2)}$ in Figure 9.2), intersecting at $e_i^0 = d$. $\gamma \to \infty$ implies $e_i^N = e_i^0$ and that a country has a dominant strategy. However, below, we shall encounter other reasons why reaction functions might be orthogonal to each other.

9.3.2 General Case

What has been derived for the example can be shown for the general case too. The first-order conditions (FOC) in the NE (assuming an interior solution[8]) are given by:

$$\pi_i' = \beta_i'(e_i) - \phi_i'\left(\sum_{j=1}^{N} e_i\right) = 0 \ \forall \ i \text{ and } j \in \mathrm{I}\,. \tag{9.9}$$

Condition (9.9) *implicitly* defines the N reaction functions. From (9.9) (and recalling $\beta_i'' \leq 0$ and $\phi_i'' \geq 0$) it immediately follows that for a given level of emissions in the other countries, e_{-i}, country i emits more, the higher the marginal opportunity cost of abatement, β_i', and the lower the marginal environmental damage, ϕ_i', are. Recall, in the example, that a high value of β_i' and/or a low value of ϕ_i' was associated with a high value of $\gamma = b/c$.

Differentiating the expression in (9.9) and rearranging terms, the *slopes* of these reaction functions are derived:[9]

$$r_i' = \frac{\partial e_i}{\partial e_{-i}} = \frac{\phi_i''}{\beta_i'' - \phi_i''}; \ -1 < r_i' < 0; \ e_{-i} = \sum_{j \neq i}^{N} e_j. \tag{9.10}$$

Since it is irrelevant to country i from which country emissions stem, all emission sources originating from outside i are summarized to e_{-i}.[10] The

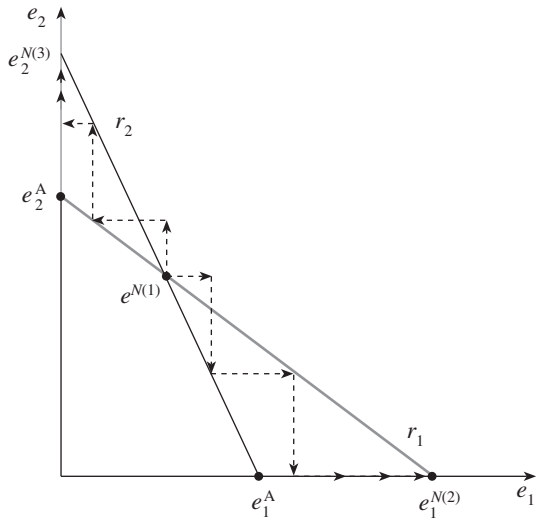


Figure 9.3 Reaction functions of slope less than −1

slopes of the reaction functions are less than 1 in *absolute terms* since $\beta_i'' < 0$ and are negative due to the assumption of the concavity of the net benefit functions, that is, $\beta_i'' - \phi_i'' < 0$ and $\phi_i'' > 0$. The greater β_i'' (the less negative or smaller in absolute value), the more negatively sloped are the reaction functions. At the limit, for β_i'' close to zero, reaction functions have a slope of -1. The same is true for large values of ϕ_i'', that is, $\phi_i'' \to \infty$.

Recall in the example that $\beta_i'' = b$ and $\phi_i'' = c$ so that an increase of $\gamma = b/c$ implies now in the general case that $|\beta_i''/\phi_i''|$ is increasing.

There are six interesting cases to note with respect to the properties of the reaction functions, four of which depart from assumptions A_1.

First, allowing for $\beta_i'' > 0$ (contrary to A_1) but still assuming $\beta_i'' - \phi_i'' < 0$ (according to A_1), the slope of the reaction functions would still be negative, though smaller than -1 (greater than 1 in absolute terms). In Figure 9.3 such an example is shown where *both countries' reaction functions have a slope less than −1*. In this example there are three NE, two at the boundaries, that is, $e^{N(2)}$ and $e^{N(3)}$, and one in the interior of the strategy space, that is, $e^{N(1)}$. As above, the equilibrium $e^{N(1)}$ is determined as the intersection of both best reply functions. To see that there are two more equilibria, consider that country 1 chooses emission level e_1^{A} to which country 2's best reply is $e_2 = 0$. Given $e_2 = 0$, country 1 will respond by choosing $e_1^{N(2)}$ which results in the equilibrium $e^{N(2)}$. By symmetry, equilibrium $e^{N(3)}$ is derived where country 1 emits nothing. Of course, this requires

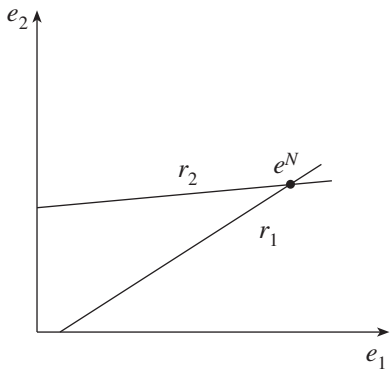


Figure 9.4 Positively sloped reaction functions

reaction functions to be piecewisely defined (see (9.5)) and that $e_i^{\max} > e_i^{A}$ holds, otherwise $e^{N(1)}$ would be the only NE.

From the example one would expect that the question of whether a game has a unique equilibrium or multiple equilibria would be closely related to the slope of the reaction functions. This is indeed the case and this issue will be taken up in more detail in Section 9.3.

Note that an immediate implication of the arguments above is that the equilibrium conditions (9.9) hold *only* for an *interior* NE. In Figure 9.3, though (9.9) holds for country 1 at $e^{N(2)}$, $\beta_2' - \phi_2' < 0$ must be true for country 2. An 'adjustment' to $\beta_2' - \phi_2' = 0$ would imply negative emissions, which are ruled out by the definition of the strategy space, that is, $e_i \geq 0\ \forall\ i \in I$. By the same token, at $e^{N(3)}$ $\beta_2' - \phi_2' = 0$ but $\beta_1' - \phi_1' < 0$.[11]

Second, from (9.10) it appears that one possibility to generate *positively sloped reaction functions* would be to assume $\beta_i'' > 0$ and $\beta_i'' - \phi_i'' > 0$ (contrary to A_1). However, $\beta_i'' - \phi_i'' > 0$ implies a *convex* payoff function for which no interior maximum exists. In particular, as will be set out in Section 9.3, the sufficient conditions for the existence of a Nash equilibrium would not be met. Therefore, this possibility is discarded.

Another *theoretical* possibility to obtain positively sloped reaction functions would be to assume $\beta_i'' < 0$ (according to A_1) but $\phi_i'' < 0$ (contrary to A_1). In particular, one would need $|\beta_i''| > |\phi_i''|$. This assumption would be in line with a concave payoff function, though, of course, in the present setting one has to think hard to find a case where $\phi_i'' < 0$ would hold true.[12] Nevertheless, this case is illustrated in Figure 9.4, since we refer to it later, in Section 10.3. As it is drawn, country 1's reaction function has a slope of greater than 1 and that of country 2 has a slope of less than 1.

Third, *best reply functions may not intersect* for any emission tuple in the

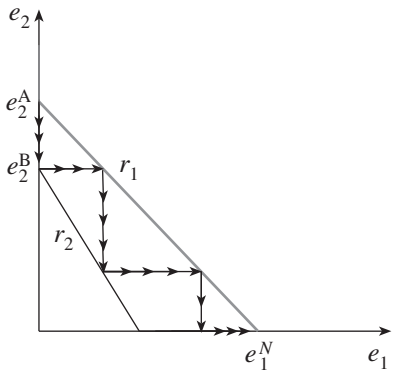


Figure 9.5 Non-intersecting reaction functions

positive quadrant. Such an example is shown in Figure 9.5 (see Hoel 1991) where reaction functions are drawn as if they were defined piecewise and where e^N is the NE tuple. To see this, pick e^A ($e_1 = 0$, $e_2 = e_2^A$) as the 'initial situation' which is also the 'origin' of country 1's reaction function. A best reply to $e_1 = 0$ has country 2 choosing e_2^B. This move and the subsequent adjustment reactions are indicated by arrows. It is apparent that finally $e^N = (e_1^N, e_2^N = 0)$ is reached which, in contrast to Figure 9.3, is a unique equilibrium.

Fourth, consider the possibility of *linear damage cost functions*, which implies $\phi_i'' = 0$ (contrary to A_1) and hence $r_i' = 0$. Consequently, *reaction functions are orthogonal to each other* as, for instance, $r_{1(2)}$ and $r_{2(2)}$ in Figure 9.2. That is, countries have *dominant strategies*. Though linear damage cost functions are sometimes chosen for mathematical simplicity (see Bauer 1992; Hoel 1992a; Mäler 1994), they may be criticized on two grounds: (a) they ignore the ecological relations mentioned at the beginning of this section and may only be justified in the range of low emission levels; (b) a linear specification does not depict a typical feature of global pollution control, namely that of the interaction and dependency of countries.

Fifth, another interesting case occurs if $\beta_i'' = 0$ (contrary to A_1) because then $r_i' = -1$. That is, *reaction functions run parallel* to each other, as shown in Figure 9.6. In Figure 9.6(a), country 1's reaction function lies to the left of country 2's reaction function[13] and therefore in equilibrium country 1 emits nothing. (The arguments to derive e^N are basically the same as those used to derive the equilibrium in Figure 9.5.)

In Figure 9.6(b), net benefit functions of all countries are assumed identical (and $\beta_i'' = 0$). Then $r_i = r_j$ and there are multiple NE which comprise all emission tuples on these reaction functions.

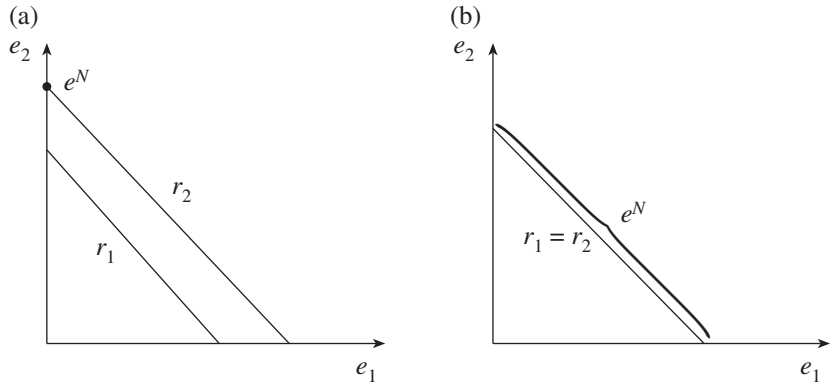


Figure 9.6 Parallel reaction functions

Sixth, though all reaction functions have been drawn as straight lines for simplicity in the previous figures, this is not true in general. Whenever *second-order derivatives are not constant*, *best reply functions are not linear*. Then the curvature of the reaction functions depends on third-order derivatives. Basically, reaction functions may be concave (Figure 9.7a), convex (Figure 9.7b) or may contain some segments which are convex and some which are concave (Figure 9.7c).[14]

In order to analyze the curvature of these reaction functions, the second-order derivatives of these functions have to be determined. This is demonstrated with respect to country 1's reaction function assuming a two-dimensional emission space. Defining convexity or concavity with respect to the origin of the $e_1 - e_2$ space, country 1's slope of the reaction function may be written (based on (9.10)) as:

$$\frac{1}{r_1'} = \frac{\partial e_2}{\partial e_1} = \frac{\beta_1'' - \phi_1''}{\phi_1''} \tag{9.11}$$

from which follows the second-order derivative (see Appendix VI.1 for details):

$$\frac{1}{r_1''} = \frac{\partial^2 e_2}{\partial e_1^2} = -\frac{\beta_1''' \phi_1''^2 - \beta_1''^2 \phi_1'''}{(-\phi_1'')^3} \Rightarrow \frac{\partial^2 e_2}{\partial e_1^2} > (<)\ 0 \text{ if } \beta_1''' \phi_1''^2 - \beta_1''^2 \phi_1''' > (<)\ 0. \tag{9.12}$$

So far, we have not specified our expectations regarding the third-order derivatives and hence (9.12) cannot be signed. However, it seems plausible to assume $\beta_1''' \geq 0$ and $\phi_1''' \geq 0$. $\beta_1''' \geq 0$ implies that the marginal opportunity

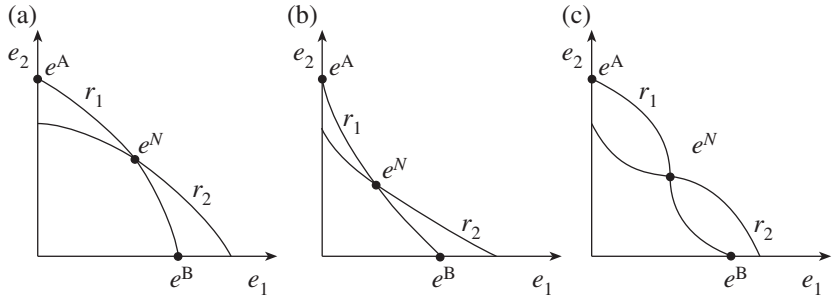


Figure 9.7 Non-linear reaction functions

cost of abatement increases more than proportionally at lower emission levels.[15] $\phi_1''' \geq 0$ implies that at high levels of emissions marginal damage increases more than proportionally.[16] This assumption is in line with the fact that most environmental systems (almost) collapse at very high emission levels (see Section 9.2). Applying these assumptions to (9.12) we have:

$$\frac{\partial^2 e_2}{\partial e_1^2} < (>)\ 0 \text{ if } \frac{\beta_1''^2}{\beta_1'''} > (<) \frac{\phi_1''^2}{\phi_1'''}. \tag{9.13}$$

Consequently Figure 9.7a implies the first constellation in (9.13) where reaction functions are *concave* with respect to the origin. Accordingly, Figure 9.7b reflects the second constellation in (9.13), implying *convex* reaction functions. In Figure 9.7c, the first constellation holds within the segment between e^A and e^N and the second constellation within the segment between e^N and e^B of country 1's reaction function.[17] In the example based on payoff function (9.3), $\beta_1''' = 0$ and $\phi_1''' = 0$ and hence reaction curves may either be called convex or concave.

9.4 EXISTENCE AND UNIQUENESS OF NASH EQUILIBRIUM

From the discussion of the slope of the reaction functions and of the point(s) where they intersect the question arises when we can expect a unique Nash equilibrium. As preliminary information one may find out first whether an equilibrium exists at all. The sufficient conditions for the existence of a Nash equilibrium are readily checked in the present context (Friedman 1986, p. 23):

Definition 9.1: Sufficient conditions for a Nash equilibrium
(1) The strategy space for each player, S_i, is compact and convex; (2) $\pi_i(s)$ is defined, continuous and bounded for all $s \in S$ and $i \in I$; and (3) $\pi_i(s)$ is concave with respect to s_i for all $s \in S$.

The first property is satisfied. The strategy space is compact, that is, $E_i = [0, e_i^{\max}]$ and the strategy space of player i is convex because any strategy lies within this interval. For instance, any strategy tuple for which $E_i = [0, d]\ \forall\ i \in I$ holds in Figure 9.2 can be connected by a straight line which is entirely contained in the rectangle defined by the emission tuples $(0, d)$, $e^0 = (d, d)$ and $(d, 0)$ and the origin $(0, 0)$.

The second property also holds since in (9.1) a continuous payoff function for each player is defined. Payoffs are bounded since an upper and lower bound for the strategy space has been defined (see also Section 9.4).

The third property is implied by A_1 but also, as we know from the previous section, by $\beta_i'' > 0$ as long as $\beta_i'' - \phi_i'' < 0$ holds and by upward-sloping reaction functions requiring $\phi_i'' < 0$ and $|\beta_i''| > |\phi_i''|$.

Though the sufficient conditions for the existence of a Nash equilibrium are rather general, this is not true for the uniqueness of a Nash equilibrium. In fact, one may need more than one theorem to show uniqueness, depending on the particular circumstances. We shall introduce two of them.

If best response functions are twice differentiable and continuous, then uniqueness can be checked by the following theorem (see also Friedman 1986, pp. 44ff. for a proof):

Theorem 9.1 (Rosen 1965)
Let $\Gamma(N, S, \Pi)$ be a strictly smooth game. That is, $\partial\pi_i/\partial s_{jk}$ and $\partial^2\pi_i/\partial s_{jk}\partial s_{je}$ exist and are continuous for all $s_i \in S$ (including strategies at the boundary of S) where $i, j, k, l \in I$, and $i, j, k, l = 1, ..., N$. Then if the Jacobian matrix of the first-order derivatives of the implicit reaction functions is negatively quasi-definite for all $s_i \in S$, and the sufficient conditions for the existence of a Nash equilibrium are satisfied, then there is unique Nash equilibrium.

Since $\partial\pi_i/\partial s_j = 0$ are the *implicit reaction functions*, the Jacobian matrix takes the following form:

$$J = \begin{bmatrix} \frac{\partial^2\pi_1}{\partial s_1^2} \cdots \frac{\partial^2\pi_1}{\partial s_1 \partial s_N} \\ \cdots\cdots\cdots\cdots\cdots \\ \frac{\partial^2\pi_N}{\partial s_1^2} \cdots \frac{\partial^2\pi_N}{\partial^2 s_N} \end{bmatrix}. \tag{9.14}$$

Negative quasi-definiteness requires that $J^A = (J + J^T) > 0$ where the superscript T denotes the transpose. In the two-country case J^A is given by:

$$J = \begin{bmatrix} \beta_1'' - \phi_1'' & -\phi_1'' \\ -\phi_2'' & \beta_2'' - \phi_2'' \end{bmatrix}, \; J^T = \begin{bmatrix} \beta_1'' - \phi_1'' & -\phi_2'' \\ -\phi_1'' & \beta_2'' - \phi_2'' \end{bmatrix} \Rightarrow$$

$$J^A = J + J^T = \begin{bmatrix} 2(\beta_1'' - \phi_1'') & -(\phi_1'' + \phi_2'') \\ -(\phi_1'' + \phi_2'') & 2(\beta_2'' - \phi_2'') \end{bmatrix} \tag{9.15}$$

in the global emission game and therefore negative quasi-definiteness requires:

$$4(\beta_1'' - \phi_1'')(\beta_2'' - \phi_2'') - (\phi_1'' + \phi_2'')^2 > 0. \tag{9.16}$$

If we try to establish that the inequality sign in (9.16) holds generally in the emission game by applying assumptions A_1 we shall be disappointed. At such a general level the opposite sign in (9.16) cannot be ruled out *a priori*. Only by introducing restrictive assumptions, for example, $\phi_1'' = \phi_2''$, will the above inequality sign generally hold. Thus, for instance, in the case of orthogonal reaction functions uniqueness of the Nash equilibrium would follow due to $\phi_i'' = 0 \; \forall \; i \in I$.

This lack of predictability suggests the need to search for other theorems to establish a unique Nash equilibrium for the general payoff functions (9.1) and assumptions A_1. Before doing so we apply Theorem 9.1 to example (9.3). In order to discuss some general features, it will prove helpful to depart from the previous assumption of symmetric countries. Moreover, it will turn out to be convenient to rearrange the implicit reaction functions slightly to have:

$$\partial \pi_i / \partial e_i = b_i d - b_i e_i - c_i(e_i + e_j) \leq 0 \Rightarrow \frac{\gamma_i d}{1 + \gamma_i} - e_i - \frac{1}{1 + \gamma_i} e_j \leq 0; \; \gamma_i = \frac{b_i}{c_i} \tag{9.17}$$

from which matrix J^A can be computed:

$$J^A = J + J^T = \begin{bmatrix} -2 & -\frac{1}{1+\gamma_1} - \frac{1}{1+\gamma_2} \\ -\frac{1}{1+\gamma_1} - \frac{1}{1+\gamma_2} & -2 \end{bmatrix} \tag{9.18}$$

and hence negative quasi-definiteness requires:

$$-2 < 0 \text{ and } 4 - \left(-\frac{1}{(1+\gamma_1)} - \frac{1}{(1+\gamma_2)} \right)^2 > 0. \tag{9.19}$$

Since $|-1/(1+\gamma_i)| < 1$, where the LHS term is the slope of a country's reaction function (see (9.8)), (9.19) is always satisfied. From (9.19) it is also

evident that it is *not* necessary that the slopes of both reaction functions are smaller than 1 but it suffices if their *sum* is less than 2. Figure 9.4 represents such a case (though reaction functions are positively sloped!) where the slope of country 1's reaction function is greater than 1 and that of country 2's reaction function less than 1; however, the sum of both slopes is less than 2 and therefore there is a unique Nash equilibrium.

Thus in the case of linear reaction functions, Theorem 9.1 may be specified as follows:

Corollary 9.1
In a strictly smooth two-player game $\Gamma(N, S, \Pi)$ with linear reaction functions and where the sufficient conditions for the existence of a Nash equilibrium hold, there is a unique Nash equilibrium provided $|\partial r_1/\partial s_2 + \partial r_2/\partial s_1| < 2$.

Proof: Is obvious and therefore omitted. QED

Note that Corollary 9.1 also covers the case where the reaction functions have opposite signs. In order to make progress in establishing uniqueness of the equilibrium for the general payoff functions in (9.1), some definitions on the way to another uniqueness theorem are needed (see, for example, Friedman 1986, pp. 42ff.):

Definition 9.2: Distance
Let x and y be two vectors of real numbers and denote the distance from x to y by $d(x, y) = \max |x_i - y_i|$.

Definition 9.3: Contraction
Let $f(x)$ be a function and λ be a positive scalar. If for any x and x' in the domain of the function and for any $0 \leq \lambda < 1$ $d(f(x), f(x')) \leq \lambda d(x, x')$ holds, then $f(x)$ is a contraction.

Thus, roughly, a contraction leaves the images of two points closer than the original points. If $f(x)$ is a differentiable function (which is not necessary for a contraction and Theorem 9.2 below), then a contraction implies $\Sigma|\partial f_i(x)/\partial x_j| \leq \lambda$ for each component $f_i(x)$ of $f(x) = (f_1(x), ..., f_N(x))$.

Theorem 9.2 (Friedman 1986)
Let a game be described by $\Gamma(N, S, \Pi)$, then if best reply functions are contractions the game has a unique equilibrium.

Proof: The proof follows by contradiction (see Friedman 1986, pp. 44ff.). Suppose that there are two equilibria instead of one equilibrium, that is,

$s_1^{*(1)}$ and $s_1^{*(2)}$ and thus $s_1^{*(1)} = r(s_1^{*(1)})$ and $s_2^{*(2)} = r(s_2^{*(2)})$. Then $d(r(s_1^{*(1)}), r(s_1^{*(2)})) = d(s_1^{*(1)}, s_1^{*(2)})$, but r being a contraction means that $d(r(s_1^{*(1)}), r(s_1^{*(2)})) \leq \lambda d(s_1^{*(1)}, s_1^{*(2)})$ for $0 \leq \lambda < 1$. This requires, however, $s_1^{*(1)} = s_1^{*(2)}$ which implies that the equilibrium is unique. QED

Generally, the advantage of Theorem 9.2 is twofold. First, it does not require twice differentiable payoff functions. Second, if best reply functions are contractions, then the existence of a Nash equilibrium is automatically ensured and does not have to be checked separately. Of course, in our context we have already established the existence of an NE and payoff functions are twice differentiable over the whole domain of the strategy space. Hence, the upshot of Theorem 9.2 in the emission game is that reaction functions must approach each other in the N dimensional strategy space by a slope *consistently* less than 1. This is known to be true for the general payoff functions (9.1) and assumptions A_1 (see in particular (9.10)) as well as for the example in (9.3) (see in particular (9.8)).

The logic behind Theorem 9.2 may best be seen in Figure 9.3, where there are two reaction functions with slope less than -1. Hence, these reaction functions are *no* contraction. In equilibrium $e^{N(1)}$ any small perturbation of the strategies leads the players away from this equilibrium, either to $e^{N(2)}$ or $e^{N(3)}$ which is indicated by the arrows.[18]

Figure 9.3 may be contrasted with Figure 9.5 where there are non-intersecting reaction functions, both with a negative slope less than 1 in absolute terms. Hence, Figure 9.5 is covered by Theorem 9.2 and confirms our finding of a unique Nash equilibrium. Also in Figures 9.7a, b and c, though reaction functions are non-linear, they all have a slope less than 1 in absolute terms over the *whole* domain of the strategy space and therefore there is a unique equilibrium according to Theorem 9.1 in all three cases.

However, the case of parallel reaction functions with slope -1 due to $\beta_i'' = 0$ (recall A_1 is violated) which has been illustrated in Figure 9.6a is covered neither by Theorem 9.1 nor by Theorem 9.2, though from the previous analysis it is known that there is a unique equilibrium. The example in Figure 9.4 is also not covered by Theorem 9.2 since the reaction function of country 1 has a slope greater than 1 and this country's reaction function is therefore not a contraction. However, as argued above, the sum of the slopes of the two linear reaction functions is less than 2 and hence Corollary 9.1 applies.

The findings stress the character of the above theorems: they provide 'only' *sufficient* but *not necessary* conditions for the existence of a unique equilibrium. Thus, if those conditions do not hold in a particular game, one *cannot* conclude that a unique equilibrium does not exist. The examples also illustrate that there is no universal theorem to prove uniqueness

of an equilibrium and that it is convenient to have a bunch of theorems available.

9.5 CHARACTERIZATION OF PAYOFF SPACE AND NORMAL FORM REPRESENTATION

All relevant information is now available to determine the boundaries of the payoff space, some important benchmarks and the normal form of the global emission game. Under each heading the general case is discussed first and then the application to the example is illustrated. We start by determining the upper and lower bounds of the payoff space.

9.5.1 Upper Bound of Payoff Space

The upper bound of the payoff space in this game is the maximax payoff, π_i^{U}. Since $\partial\pi_i/\partial e_j<0$ is true, the maximax payoff is determined by assuming that all countries emit nothing and country i maximizes its payoff, that is, $\max\pi_i(e_i, 0)$, which delivers $\pi_i^{\mathrm{U}}=\pi_i(e_i(0), 0)$.

For the example it is known that $e_i=bd/(b+c)$ is country i's best reply to $e_j=0$. Substituting these emissions in (9.3) gives:

$$\pi_i^{\mathrm{U}}=\frac{b^2d^2}{2(b+c)}. \tag{9.20}$$

9.5.2 Lower Bound of Payoff Space

From $\partial\pi_i/\partial e_j<0$ it follows that the lowest payoff to a country is obtained if all other countries choose their highest emission level $e_i^{\max}$. If $\partial\pi_i(e_i, e_i^{\max})/\partial e_i>0$ for small values of e_i, country i's lowest payoff is given by $\pi_i^{\mathrm{L}}=(0, e_{-i}^{\max})$. If the first derivative is negative for all $e_i>0$, the lowest payoff is $\pi_i^{\mathrm{L}}=(e_i^{\max}, e_{-i}^{\max})$.

In the example $e_j^{\max}=e_j^0=d$ and since $\partial\pi_i(e_i, d)/\partial e_i>0$ if $b>c$, which we assume to hold, $\pi_i^{\mathrm{L}}=(0, d)$. For the function in (9.3) we get:

$$\pi_i^{\mathrm{L}}=-\frac{cd^2}{2}. \tag{9.21}$$

9.5.3 Minimax Payoff

To punish country i, all countries will choose their maximum emission level $e_j^{\max}$ to which country i chooses its best reply, that is, $\pi_i^{\mathrm{M}}=(e_i(e_{-i}^{\max}, e_{-i}^{\max})$.[19] Obviously, the assumption regarding the upper bound of the strategy space

is crucial for the level of π_i^{M}. For any $e_{-i}^{\max} > e_{-i}^{N}$, $\pi_i^{\mathrm{M}} < \pi_i^{N}$ holds and trivially $\pi_i^{\mathrm{M}} = (e_i(e_{-i}^{N})\ e_{-i}^{N}) = \pi_i^{N}$ is true if $e_{-i}^{\max} = e_{-i}^{N}$. As discussed in Section 9.2, if one is more specific regarding the curvature of the benefit curve, then some candidates for $e_i^{\max}$ may suggest themselves.

In the example $e_j^{\max} = e_j^0$ holds by assumption and country *j* minimaxes country *i* by choosing $e_j^{\mathrm{M}(i)} = e_j^0 = d$ to which country *i* reacts by choosing its best reply $e_i(d) = e_i^{\mathrm{M}(i)} = d(b-c)/(b+c)$ according to (9.4). Consequently, we have:

$$\pi_i^{\mathrm{M}(i)} = \frac{bd^2(b-3c)}{2(b+c)};\ \pi_j^{\mathrm{M}(i)} = \frac{bd^2(b-c)^2}{2(b+c)^2}. \tag{9.22}$$

With this information we can now write $\Pi = \{\pi | \pi_i \geq \pi_i^{\mathrm{L}}\ \forall\ i \in \mathrm{I}\}$ and $\Pi^{\mathrm{IR}} = \{\pi | \pi_i \geq \pi_i^{\mathrm{M}}\ \forall\ i \in \mathrm{I}\}$. The *normal form* of this game is given by $\Gamma = (N, \mathrm{S}, \Pi)$ where $N = \{1, ..., N, \{, \mathrm{S} = \{\mathrm{S}_1, ..., \mathrm{S}_N\}, \mathrm{S}_i = [0, e_i^{\max}], \Pi = \{\Pi_1, ..., \Pi_N\}$ and $\Pi_i = [\pi_i^{\mathrm{L}}, \pi_i^{\mathrm{U}}]$. In the following it is assumed that all relevant emissions are not restricted by the definition of the strategy space, that is, $e_i^{\max} > e_i^{N}$, $e_i^{\mathrm{M}(i)}$, $e_i^{\mathrm{S}} \geq 0$ where socially optimal emission levels, e_i^{S}, are derived in the next section.

For the example $\mathrm{S}_i = \mathrm{E}_i \in [0, d]$, $N = 2$ and the lower and upper bounds of the payoff set of player *i* are given by (9.20) and (9.21) respectively.

9.6 SOCIAL OPTIMUM

Before illustrating the benchmarks of Section 9.5, socially optimal emission levels have to be derived. We start out by considering the general case first and then determine the social optimum in the example.

9.6.1 General Case

The *social optimum* in the global emission game follows from maximizing global net benefits according to:

$$\max_{e_1, ..., e_N} \sum_{i=1}^{N} \pi_i = \sum_{i=1}^{N} \beta_i(e_i) - \sum_{j=1}^{N} \phi_j\left(\sum_{k=1}^{N} e_k\right);\ i, j, k \in \mathrm{I}, \text{ s.t. } e_i^{\mathrm{S}} \geq 0\ \forall\ i \in \mathrm{I} \tag{9.23}$$

which leads to:[20]

$$\mathrm{CBA}: = \begin{cases} (1)\ \beta_i'(e_i^{\mathrm{S}}) = \Sigma\phi_k'(\Sigma e_k^{\mathrm{S}}) \text{ if } e_i^{\mathrm{S}} > 0 \\ (2)\ \beta_i'(e_i^{\mathrm{S}}) = \Sigma\phi_k'(\Sigma e_i^{\mathrm{S}} + 0) \wedge \beta_j'(0) < \Sigma\phi_k'(\Sigma e_i^{\mathrm{S}} + 0) \text{ if } e_i^{\mathrm{S}} > 0 \wedge e_j^{\mathrm{S}} = 0 \\ (3)\ \beta_i'(0) < \Sigma\phi_k'(0) \text{ if } e_i^{\mathrm{S}} = 0 \end{cases} \quad \forall\ i, j, k \tag{9.24}$$

where the superscript S stands for social optimum.[21] That is, a country considers not only the damage caused by its emission in its own country, as in the non-cooperative Nash equilibrium, but also those in the neighboring country. Since (9.24) states the optimality conditions of a *cost–benefit analysis*, these conditions are abbreviated CBA. The first condition in (9.24) assumes an interior solution. That is, emissions in all countries are positive in the social optimum. In this case social optimality implies that *marginal opportunity costs of abatement are equal across all countries*. The second condition takes care of the possibility that for some countries (the *j* countries) emissions might be zero in the social optimum. This could be the case if marginal opportunity costs of abatement are very asymmetric in various countries. The *third condition* reflects the possibility that damages are so high that no emissions are advisable in the social optimum.

Comparing aggregate emissions in the social optimum with those in the Nash equilibrium we find:

Proposition 9.1
Let the global emission game be described by the payoff functions in (9.1) and assumption A_1 in (9.2), then $\Sigma e_i^N > \Sigma e_i^S$.

Proof: This assertion is readily proved by contradiction. Compare the FOC (9.24) with (9.9), where the latter assumes an *interior* Nash equilibrium, and assume that $\Sigma e_k^N < \Sigma e_k^S$ would be true. Then, (1) $\phi_i'(\Sigma e_k^N) < \Sigma \phi_k'(\Sigma e_k^S)\ \forall\ i, j$ and k. Consequently, from (9.9) and the first constellation in (9.24), we must have (2) $\beta_i'(e_i^N) < \beta_i'(e_i^S)\ \forall\ i \in I$, which is only possible provided $e_i^N > e_i^S\ \forall\ i \in I$, and hence $\Sigma e_k^N > \Sigma e_k^S$, which contradicts the initial assumption. For the second constellation in (9.24), again, assume $\Sigma e_k^N < \Sigma e_k^S$ and (1) above follows. Then for a country i for which $e_i^S > 0$ holds, this implies $\beta_i'(e_i^N) = \phi_i'(\Sigma e_k^N) < \Sigma \phi_k'(\Sigma e_k^S) = \beta_i'(e_i^S)$ which is only possible if $e_i^N > e_i^S$ for all i countries which, again, contradicts the initial assumption. For the third constellation $\Sigma e_k^N > \Sigma e_k^S$ is obvious because $e_i^N \geq 0\ \forall\ i \in I$ by assumption. The proof in the case of a corner solution in the Nash equilibrium proceeds exactly along the same lines and is therefore omitted here. QED

Note that though $\Sigma e_k^N > \Sigma e_k^S$ always holds, this should *not* be mistaken to imply $e_i^N > e_i^S\ \forall\ i \in I$. For instance, assume that abatement cost functions are equal across all countries, that is, $\beta_i' = \beta_j'\ \forall\ i$ and $j \in I$, but that environmental damages are perceived differently. Consequently, countries with high damages will have low emissions in the Nash equilibrium and vice versa. In the social optimum, where, due to $\beta_i' = \beta_j'$, $e_i^S = e_j^S$ holds and where the countries with low damage also have to consider high damage in the

neighboring countries, the high damage countries may very well have to increase emissions compared to the Nash equilibrium.[22]

Moreover, note that $\pi_i^S < \pi_i^N$ and even $\pi_i^S < \pi_i^M$ may be possible for some countries, though $\Sigma\pi_i^S > \Sigma\pi_i^N$ and $\Sigma\pi_i^S > \Sigma\pi_i^M$ always hold. For instance, if a country has low opportunity costs of abatement compared to other countries and perceives damages as less severe than its neighbors, a *globally rational solution* may imply that the *individual rationality constraint* of a country is violated.[23] Since this result is important for subsequent chapters it is summarized in the following proposition:

Proposition 9.2
Let the global emission game be described by the payoff functions in (9.1) and assumption A_1 in (9.2), then $\pi_i^N > \pi_i^S$ and $\pi_i^M > \pi_i^S$ is possible if $e_i^{\max} \geq e_i^N$, though $\Sigma\pi_i^{M(i)} < \Sigma\pi_i^N < \Sigma\pi_i^S$.

Proof: The first part of the assertion has already been proved by example. The second part follows trivially from:

$$\max_{e_1, \ldots, e_N} \Sigma\pi_i \geq \max_{e_1} \pi_1 + \ldots + \max_{e_N} \pi_N$$

and by the definition of a minimax payoff. QED

9.6.2 Example

For the example socially optimal emission levels are derived by differentiating $\Sigma\pi_i$ with respect to e_1 and e_2, setting both derivatives equal to zero and solving for the respective variable. Thus, one derives a sort of 'cooperative reaction functions' (that is, $e_i^S(e_j^S)$ or in implicit form $\beta_i' - \Sigma\phi_j' = 0$) which have to be mutually inserted in each other in order to solve for socially optimal emission levels:

$$e_i^S = \frac{bd}{b + 4c} \Leftrightarrow e_i^S = \frac{\gamma d}{\gamma + 4}. \tag{9.25}$$

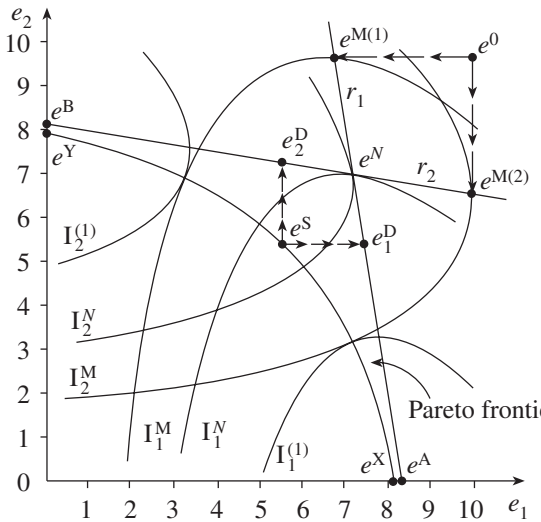
Inserting these emission levels into the net benefit functions in (9.3) gives:

$$\pi_i^S = \frac{b^2 d^2}{2(b + 4c)}. \tag{9.26}$$

A routine check confirms $e_i^S < e_i^N$ and $\pi_i^S > \pi_i^N \; \forall \; i \in I$ due to the assumption of symmetric countries.

9.7 INDIFFERENCE CURVES, PAYOFF STRUCTURE AND PARETO FRONTIER

All information derived in the previous sections is visualized for the example in Figure 9.8, assuming $b=5$, $c=1$ and $d=10$. The particular results of the example will be put into perspective to the general case in the course of the discussion.



Note: Payoff functions (9.3) apply, assuming $b=5$, $c=1$ and $d=10$.

Figure 9.8 Nash equilibrium and social optimum in the global emission game

9.7.1 Indifference Curves

In Figure 9.8 the *Nash equilibrium emission tuple* is denoted by e^N. Moreover, the *socially optimal emission tuple* is indicated by e^S. $e^{M(1)}$ ($e^{M(2)}$) denotes the *minimax emission tuple* if country 1 (country 2) is minimaxed by country 2 (country 1) and e^0 is the emission tuple if no emissions are reduced at all.

Additionally, *indifference curves* for different payoff levels have been drawn. An indifference curve depicts all emission combinations for a given constant payoff level of a country. For instance, I_1^N reflects all emission tuples which give country 1 the same payoff as in the Nash equilibrium, π_1^N. I_i^M denotes the indifference curve of country i which ensures this country its minimax payoff, π_i^M.

From country 1's perspective indifference curves which lie in a more south-easterly direction represent higher welfare levels. Thus indifference curve $I_1^{(1)}$ implies a higher welfare to country 1 than in the Nash equilibrium. This is so since for indifference curves which are located in a more south-easterly direction emissions in country 2 are lower for a given emission level in country 1 and $\partial\pi_i/\partial e_j < 0 \;\forall\; i \in I$ holds. By the same token, indifference of country 2 which lies in a more north-westerly direction implies higher welfare levels to country 2. The *concavity* (*convexity*) of country 1's (country 2's) indifference curves with respect to the origin follows from the assumptions of the second-order derivatives of the benefit and damage cost functions as specified in assumption A_1. But even if B_i'' was assumed to be positive, the indifference curves would have the same curvature as long as B_i'' is sufficiently small. Appendix VI.2 gives a formal proof of this assertion.

Moreover, from Figure 9.8 it is evident that the *reaction function* of country 1 connects all the peaks of its indifference curves. By symmetry, the same holds for country 2, but with the axes reversed. That is, reaction functions run through all points where the slope of the indifference curve is zero. To understand this consider for instance country 1's indifference curve for which $I_1 := \pi_1(e_1, e_2) = \bar{\pi}_1$ or, alternatively, $\pi_1(e_1, e_2) - \bar{\pi}_1 = 0$ is true where $\bar{\pi}_1$ indicates some constant utility level. Differentiating this implicit function gives the *slope of country 1's indifference curve*:

$$I_1' = \frac{\partial e_2}{\partial e_1} = -\frac{\partial \pi_1/\partial e_1}{\partial \pi_1/\partial e_2} = \frac{\beta_1' - \phi_1'}{\phi_1'}. \tag{9.27}$$

This slope is zero for $\beta_1' - \phi_1' = 0$ which is exactly the condition along the reaction function of country 1 in the interior of the emission space (see (9.9).

Accordingly, *country 2's slope of the indifference curve* is given by:

$$I_2' = \frac{\partial e_2}{\partial e_1} = -\frac{\partial \pi_2/\partial e_1}{\partial \pi_2/\partial e_2} = \frac{\phi_2'}{\beta_2' - \phi_2'}. \tag{9.28}$$

9.7.2 Payoff and Incentive Structure

From the previous chapters we know that in a static context no other emission tuple than the (unique) Nash equilibrium is stable. If, for instance, countries agreed to reduce emissions to e^S, then each country would have an incentive to deviate to $e_i^D = e_i(e_j^S)$ which is indicated by the arrows in Figure 9.8. More generally, for any emission tuple – which is not necessarily socially optimal – for which $\beta_i'(e_i) > \phi_i'(\Sigma e_j)$ holds, a country has an incentive to increase emissions. Thus, even if a country has to increase

emissions in the social optimum compared to the Nash equilibrium, an incentive to expand emissions exists since $\phi_i'(\Sigma e_j^S) < \beta_i'(e_i^S) \leq \Sigma \phi_k'(\Sigma e_j^S)$ holds.

By the same token, if country j (which may comprise more than one country in the general case) does not reduce emissions at all and hence emits e_j^{max} (in the example $e_j^{max} = e_j^0 = d$), country i has an incentive to reduce emissions (see the arrows in Figure 9.8). In fact, in the example a best reply has country i choosing its minimax emission level $e_i^{M(i)} = e_i(e_j^0 = e_j^{M(i)})$. More generally, for any emission tuple for which $\beta_i'(e_i) < \phi_i'(\Sigma e_j)$ is true a country has an incentive to reduce emissions. A simple way of summarizing this incentive structure and to relate it to the discrete strategy games discussed in Chapters 3–8 is suggested in Matrix 9.1, where the parameters of Figure 9.8 have been assumed.

Matrix 9.1 Payoff structure in the global emission game[a]

	e_2^S	$e_2(e_1)$	e_2^0
e_1^S	138.9 138.9	116.6 149.2	79.6 129.0
$e_1(e_2)$	149.2 116.6	**127.6** **127.6**	*83.3* *111.1*
e_1^0	129.0 79.6	*111.1* *83.3*	50 50

Note: [a] Based on payoff functions (9.3), $N = 2$ and symmetric countries, that is, $b = b_1 = b_2$, $c = c_1 = c_2$. In particular $b = 5$, $c = 1$, $d = 10$ are assumed.

As in the PD and the chicken game there is an incentive to free-ride if more ambitious abatement targets than in the Nash equilibrium are realized. As in the chicken game a country has an incentive to reduce emissions unilaterally as long as $\beta_i'(e_i) < \phi_i'(\Sigma e_j)$ holds. However, in contrast to the pure strategy chicken game, a player does *not* face the decision whether to contribute to abatement or not but, due to the continuous strategy space, is confronted instead with the question of *how much* to contribute. Starting

from strategy tuple e^0 countries will adjust their emissions as long as e^N ($e^N = (e_1(e_2), e_2(e_1))$) has not yet been reached. As in the mixed strategy NE of the chicken game, in the NE of the emission game each country contributes something to abatement, that is, $e_i^N < e_i^0$. Payoffs in the NE are printed in bold in Matrix 9.1.

A typical feature apparent from Matrix 9.1 is that $\pi_i^{M(i)} < \pi_i^N$ ($\pi_i^{M(i)}(e_i(e_j^0), e_j^0)$) holds. That is, the minimax payoff (the minimax payoff tuples are printed in italic in Matrix 9.1) to a country is lower than its payoff in the Nash equilibrium. This property is a typical feature of the extended PD game I in Matrix 4.2, where the punishment option p_i in the emission game is represented now by the emission level e_i^0.[24] Moreover, extended PD game I also depicts the feature that a best reply to the punishment e_i^0 (strategy p_i in Matrix 4.2) is not e_j^0 (strategy p_j in Matrix 4.2) but an increased abatement effort, that is $e_j(e_i^0) < e_j^0$ (strategy a_j in Matrix 4.2). What is not apparent from Matrix 9.1 but is known from Proposition 9.2 is that $\pi_i^S < \pi_i^{M(i)}$ may be true if countries have asymmetric payoff functions. Thus, in accordance with the asymmetric PD game in Chapter 8, the emission game can approximately be described as either a symmetric or an asymmetric extended PD game I. However, $\pi_i^{M(i)} < \pi_i^N$ is also true in the chicken game if one allows for mixed strategies. In any of the pure NE, which constitutes a bad NE and at the same time the minimax payoff to one country, payoffs to the punished player are lower than in the mixed strategy NE. Moreover, as is known from Chapter 3, asymmetries can also be accommodated in a chicken game. Hence, taken together, it seems reasonable to view the incentive structure in an emission game as resembling that in an extended PD game and/or a chicken game. The chicken game interpretation, however, requires the play of mixed strategies. We shall return to this issue in more detail in Chapter 13, where we discuss coalition models.

9.7.3 Pareto Frontier

In Figure 9.8 the *Pareto frontier* has also been drawn, on which by definition the socially optimal emission tuple must lie. The Pareto frontier represents all those emission tuples for which it is not possible to increase the payoff to one country without reducing the payoff of any other country. That is, for a given level of welfare of country j (or countries $-i$ in the $N > 2$ case), one determines that emission tuple which maximizes country i's welfare. Thus, in contrast to the social optimum which implies an *unconstrained optimization* (with respect to welfare, see (9.23)), finding the Pareto optima in this game requires a *constraint optimization procedure*.

Constructing the Lagrangian L for this problem, restricting the number

of countries to two for simplicity,[25] and taking the relevant derivatives, the Kuhn–Tucker conditions can be derived:

$$L = \beta_1 - \phi_1 + \lambda(\beta_2 - \phi_2 - \bar{\pi}_2), \tag{9.29}$$

$$\frac{\partial L}{\partial e_1} = \beta_1' - \phi_1' - \lambda\phi_2' \leq 0,\ e_1 \geq 0,\ e_1 \frac{\partial L}{\partial e_1} = 0,$$

$$\frac{\partial L}{\partial e_2} = -\phi_1' + \lambda\,(\beta_2' - \phi_2') \leq 0,\ e_2 \geq 0,\ e_2 \frac{\partial L}{\partial e_2} = 0,$$

$$\frac{\partial L}{\partial \lambda} = \beta_2 - \phi_2 - \bar{\pi}_2 \geq 0,\ \lambda \geq 0,\ \lambda \frac{\partial L}{\partial \lambda} = 0$$

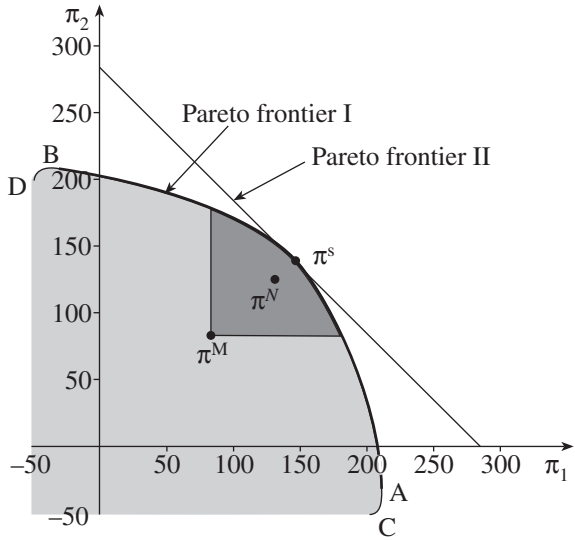
from which the following FOC along the Pareto frontier can be derived:[26,27]

$$\begin{aligned}
&\text{(a)}\ e_1 > 0,\ e_2 > 0,\ \lambda > 0\text{:}\ \frac{\beta_1' - \phi_1'}{\phi_2'} = \frac{\phi_1'}{\beta_2' - \phi_2'} \Leftrightarrow \frac{\beta_1' - \phi_1'}{\phi_1'} = \frac{\phi_2'}{\beta_2' - \phi_2'},\\
&\text{(b)}\ e_1 > 0,\ e_2 = 0,\ \lambda > 0\text{:}\ \frac{\beta_1' - \phi_1'}{\phi_2'} = \lambda,\\
&\text{(c)}\ e_1 = 0,\ e_2 > 0,\ \lambda > 0\text{:}\ \frac{\phi_1'}{\beta_2' - \phi_2'} = \lambda,\\
&\text{(d)}\ \lambda = 0\text{:}\quad \text{(i)}\ \beta_1' - \phi_1' = 0,\ \beta_2' - \phi_2' \leq 0 \Rightarrow e_1 > 0,\ e_2 = 0,\\
&\qquad\qquad \text{(ii)}\ \beta_1' - \phi_1' \leq 0,\ \beta_2' - \phi_2' \leq 0 \Rightarrow e_1 = 0,\ e_2 = 0,\\
&\text{(e)}\ \lambda \to \infty\text{:}\quad \text{(i)}\ \beta_1' - \phi_1' \leq 0,\ \beta_2' - \phi_2' = 0 \Rightarrow e_1 = 0,\ e_2 > 0,\\
&\qquad\qquad \text{(ii)}\ \beta_1' - \phi_1' \leq 0,\ \beta_2' - \phi_2' \leq 0 \Rightarrow e_1 = 0,\ e_2 = 0.
\end{aligned} \tag{9.30}$$

Condition (a) holds in Figure 9.8 along the entire segment of the Pareto frontier between e^X and e^Y (where the emission tuples themselves are excluded). Note that the LHS term of the equality to the right of the arrow in condition (a) corresponds to the slope of the indifference curve of country 1 and that the RHS term is identical to the slope of country 2's indifference curve (see (9.27) and (9.28)). Hence that part of the Pareto frontier connects all those emission tuples where the slopes of countries' indifference curves are identical.[28] Since iso-net benefit curves are strictly concave (convex), this implies tangency of these curves along the Pareto frontier. This is also evident from Figure 9.8, where it is shown for indifference curves I_1^M, $I_1^{(1)}$ and I_2^M, $I_2^{(1)}$ respectively.

Rearranging condition (a) in (9.30) we get:

$$\beta_1'\beta_2' = \beta_2'\phi_1' + \beta_1'\phi_2'. \tag{9.31}$$



Note: Payoff functions (9.3) apply, assuming $b = 5$, $c = 1$ and $d = 10$.

Figure 9.9 Pareto frontier in the global emission game

Comparing (9.31) with (9.24) with respect to an interior solution in the social optimum it is evident that only if $\beta'_1 = \beta'_2$ is global efficiency ensured in a Pareto optimum (because then $\beta'_i = \Sigma\phi'_j$). Generally, for the various emission combinations along the Pareto frontier $\beta'_1 = \beta'_2$ does *not* hold and aggregate welfare will be lower at any point along the Pareto frontier (except at e^S) compared to the social optimum. This is particularly obvious if we transfer the global emission game in the payoff space. This is done in Figure 9.9 for the example and for the parameter values assumed in Figure 9.8. Pareto frontier I in Figure 9.9 corresponds to the Pareto frontier as drawn in the emission space. It thus assumes non-transferable utility. In contrast, Pareto frontier II assumes transferable utility (and constant marginal utility equal to 1). Thus, the difference between both Pareto frontiers measures the loss from a deviation from the socially optimal emission allocation.

Condition (b) corresponds to the segment between e^X and e^A in Figure 9.8 where e^A is excluded. Similarly, *condition (c)* is the segment between e^Y and e^B where e^B is excluded. At point e^X, $\lambda = \phi'_1/(\beta'_2 - \phi'_2)$. holds. Moving from e^X towards e^A, λ decreases gradually, implying that damage in country 2 is given less weight (see (9.29)). Similarly, at e^Y $\lambda = (\beta'_1 - \phi'_1)/\phi'_2$ holds where the value of λ gradually increases when moving from e^Y towards e^B.

Comparing conditions (b) and (c) in (9.29) with those in the social optimum as given in (9.24) for a boundary solution involving $e_i^S > 0$ for some i and $e_j^S = 0$ for some j, it is, again, evident that aggregate welfare will generally be lower than in the social optimum as long as $\lambda \neq 1$.

Conditions (d)(i) and (e)(i) correspond to emission tuples e^A and e^B respectively in Figure 9.8. At e^A country 1's payoff is maximized, which requires $e_2 = 0$ due to $\partial\pi_1/\partial e_2 < 0$. Thus, e^A gives country 1's maximax payoff, $\pi_1^U(e^A)$. By the same token, e^B represents the emission tuple generating country 2's maximax payoff. Hence, e^A and e^B imply that one country emits nothing and the other chooses its best reply which correspond to the origins of the reaction functions. e^A and e^B constitute the boundaries of the Pareto frontier in Figure 9.8 and correspond to π^A and π^B respectively in Figure 9.9.

Conditions (d)(ii) and (e)(ii) are not reflected in Figure 9.8. These conditions would imply that marginal damages are already so high at low emissions compared to marginal benefits as to call for no emissions at all. For the example in (9.3) we would need $bd \leq c$ to generate this result.[29]

Turning now again to the relations in the payoff space (see Figure 9.9) it is evident that payoff tuples π^C and π^D are at the edges of the boundary of the payoff space where one country chooses $e_i = 0$ and the other $e_j = d$. This implies that at these points one country derives its lowest payoff π_i^L. Of course, these points are *not* Pareto-efficient and therefore do *not* belong to the Pareto frontier. The whole payoff space of the emission game, Π, is more lightly shaded than the individually rational payoff space, Π^{IR}.

The concavity of the Pareto frontier (in the payoff space!) follows from the *envelope theorem*. To see this, note that for the Lagrangian $L = \pi_1(e_1, e_2) + \lambda(\pi_2(e_1, e_2) - \bar{\pi}_2)$ in (9.29) the envelope function may be written as $ev = \pi_1(\Theta)$ where Θ indicates the amount by which the constraint $\pi_2(E_1, E_2) \geq \bar{\pi}_2)$ must be relaxed to obtain one unit more of π_1. Thus the Pareto frontier can be interpreted as the function ev. Since it is known that if $\pi_1(e_1, e_2)$ and $\pi_2(e_1, e_2)$ are strictly concave, the envelope function is also strictly concave (see Birchenhall and Grout 1984, pp. 176ff.; Varian 1984, pp. 322ff.), we only have to establish these properties for these functions. This, however, has already been done in note 3 based on A_1.[30,31]

NOTES

1. More sophisticated models may be found, for instance, in Arnold (1984); Kuhl (1987); Pethig (1982); and Sandler (1996). However, the main results of these works can also be accommodated in the subsequent simple model. This will also be apparent in the core-coalition models in Section 13.3, where the original setting is more sophisticated but where *all* conclusions remain valid in the 'simple' framework as well.

2. To the best of our knowledge, a several-goods model, let alone a joint-production model, still awaits game theoretical treatment in a dynamic context.
3. This condition follows from differentiating the net benefit function twice with respect to the *own* strategy e_i, that is, $\partial^2\pi_i/\partial e_i^2 < 0$. In case one requires *strict* concavity of the net benefit function in the two-dimensional emission space, $\beta_i'' < 0$ must be assumed since the Hessian determinant is only negatively definite if $\beta_i'' - \phi_i'' < 0$ and $(\beta_i'' - \phi_i'')(-\phi_i'') - (-\phi_i'')^2 > 0 \Leftrightarrow \beta_i''\phi_i'' > 0 \; \forall \; i \in \mathrm{I}$ hold (see, for example, Chiang 1984, pp. 307ff.). In the N-dimensional emission space ($N > 2$) only concavity but no strict concavity holds since there are many e_j–e_{-j} combinations for a given level of e_i which deliver the same net benefit to country i. Recall, in a global emission game, only the sum of emissions in other countries is relevant to country i, not its composition.
4. For the evaluation and measurement of environmental damage, see Endres and Holm-Müller (1998); Finus (1992a, b).
5. An alternative way of modeling a global emission game would assume the payoff function to be given by $\pi_i = \varphi_i(r_i) + \phi_i(\Sigma r_j)$ where φ_i is an abatement cost function, ϕ_i a damage cost function and r_i emission reduction in country i. Thus the optimization problem involves minimizing costs. Alternatively, one could assume $\pi_i = \beta_i(\Sigma r_j) - \varphi_i(r_i)$ where β_i are benefits from emission reduction (for example, Stähler 1998a). Assuming the appropriate first- and second-order conditions would deliver the same quantitative results obtained below.
6. Of course, this illustration is only an additional device to make the NE plausible and is, strictly speaking, only valid for dynamic games because an adjustment process requires at least two periods by definition.
7. This follows from $e_i = bd/(b+c) = \gamma d/(\gamma+1)$ and $\partial e_i/\partial\gamma = d/(\gamma+1)^2 > 0$.
8. The case of corner equilibria will be taken up below.
9. This is an application of the implicit function rule. See, for example, Chiang (1984, pp. 204ff.).
10. Put differently, since $\partial e_{-i}/\partial e_k = 1$, $\partial e_i/\partial e_k = \partial e_i/\partial e_{-i}$.
11. Setting up the maximization task subject to $e_i \geq 0$ and deriving the Kuhn–Tucker conditions would deliver the above FOC.
12. A simple way of modeling positively sloped reaction functions would be to assume $\pi_i = b(de_i - \frac{1}{2}\, e_i^2) - c(a\Sigma e_j - \frac{1}{2}\, \Sigma e_j^2)$ instead of (9.3). If $b > c$, then $\beta_i'' - \phi_i'' < 0$ and $1 > r_i' > 0$, which delivers an interior unique Nash equilibrium.
13. This may be due to the higher marginal opportunity costs of abatement and/or the lower marginal damage costs in country 1 compared to country 2.
14. As it is drawn, all examples in Figure 9.7 assume an interior Nash equilibrium.
15. For example, the benefit function could be given by the logarithmic function $\beta_i = b\ln(1 - (e_i^{\mathrm{I}} - e_i)/e_i^{\mathrm{I}}) + be_i$ where b is some positive scaling parameter, e_i^{I} denotes some initial emission level, and hence the term in brackets denotes the fraction of emission reduction. It is easily checked that $\beta_i' > 0$, $\beta_i'' < 0$ and $\beta_i''' > 0 \; \forall \; e_i > 0$.
16. An example of a damage cost function featuring these properties is $\phi_i = c(\Sigma e_j)^3$.
17. Unfortunately, there is not much more one can say about when to expect the first, second or a mix of both constellations. For the functions as specified in notes 15 and 16, $\beta_i''^2/\beta_i''' = b/(2e_i^{\mathrm{I}})$, and hence this term is constant, whereas $\phi_i''^2/\phi_i''' = 6c(e_1 + e_2)^2$. Thus, as long as b is not too big, $\beta_i''^2/\beta_i''' < \phi_i''^2/\phi_i'''$ implies a convex reaction function.
18. Hence, the predictability of the equilibrium $e^{N(1)}$ may be regarded as rather low considering the possibility that players may make small errors ('they tremble') when choosing their best reply. Any small error would upset this equilibrium. An equilibrium which is immune against such errors is called a trembling-hand perfect equilibrium (Selten 1975), which is a refinement of the NE concept.
19. It is easily checked that minimax and maximin payoff coincide in this game.
20. Thus, (9.24) reflects the relevant Kuhn–Tucker conditions of maximization problem (9.23) where the Lagrangian multipliers have been omitted for convenience.
21. Constructing the Hessian determinant for this maximization problem and using assumptions A_1 confirms that the aggregate net benefit function is strictly concave and hence $e^{\mathrm{S}} = (e_1^{\mathrm{S}}, \ldots, e_N^{\mathrm{S}})$ constitutes a unique global maximum.

22. A simple way of seeing this would be to assume cost coefficients $c_1 \neq c_2$ for country 1 and 2 instead of $c_1 = c_2 = c$ in (9.3). Assume, for instance, $b = 10$, $c_1 = 9$, $c_2 = 1$ and $d = 10$. Then, $e_1^N = 1$, $e_2^N = 9$, $e_1^S = e_2^S = 3.3$, which demonstrates the assertion that $e_i^N < e_i^S$ is possible, though $\Sigma e_i^N > \Sigma e_i^S$ holds.
23. For instance, consider the example in the previous note for which we find $\pi_1^N = -355$, $\pi_2^N = 445$, $\pi_1^S = 77.7$, and $\pi_2^S = 255.5$, which stresses that $\pi_i^N > \pi_i^S$ is possible, though $\Sigma\pi_i^N > \Sigma\pi_i^S$ is true. Computing the minimax payoffs delivers $\pi_1^M = -447.4$ and $\pi_2^M = 318.2$, which demonstrates that even $\pi_i^M > \pi_i^S$ may be true, though $\Sigma\pi_i^M > \Sigma\pi_i^S$ holds.
24. In the emission game minimaxing another country leads to a higher payoff to a country than if it is minimaxed itself, that is, $\pi_i^{M(i)} < \pi_i^{M(j)}$ whereas in Matrix 4.2 $\pi_i^{M(i)} > \pi_i^{M(j)}$ has been assumed. However, the relation of $\pi_i^{M(i)}$ and $\pi_i^{M(j)}$ is not important for the classification of an extended PD game.
25. The extension to $N > 2$ countries should be apparent but is omitted here because the subsequent derivations are mainly conducted to support the discussion of Figure 9.8.
26. An alternative way to derive the Pareto optimality conditions is to maximize the weighted welfare function $\lambda\pi_1 + (1-\lambda)\pi_2$ for different values of λ, $0 \leq \lambda \leq 1$, subject to the constraints of the strategy space, that is, $e_i \geq 0 \; \forall \; i \in I$. See Friedman (1986, p. 14).
27. The second-order sufficient conditions for a maximum hold since (1) the objective function $\beta_1 - \phi_1$ is twice differentiable and concave in the non-negative orthonant; (2) the 'constrain function' $\beta_2 - \phi_2 - \bar{\pi}_2$ is twice differentiable and concave; and (3) the FOC for each emission tuple in the emission space can be satisfied (see, for example, Chiang 1984, pp. 738ff.). The concavity of these functions has been established in note 3.
28. The similarity to the Edgworth box should be obvious.
29. From the discussion it should be evident that, due to the restriction of the lower bound of the strategy space to $e_i \geq 0$, the Pareto frontier *cannot* be simply derived from the Lagrangian conditions, which would only deliver the FOC in the segment e^X to e^Y in Figure 9.8, but would miss those along the segments e^A to e^X and e^B to e^Y.
30. By analogy with the arguments presented in note 3, strict concavity of the Pareto frontier holds only in the $\pi_1 - \pi_2$ space. With respect to the N-dimensional payoff space ($N > 2$) only concavity of the Pareto frontier holds.
31. If instead of the Lagrangian the weighted welfare function $\lambda\pi_1 + (1-\lambda)\pi_2$ is used to derive the Pareto frontier (see note 26), strict concavity follows from the simple fact that the sum of two strictly concave functions is a strictly concave function (see Chiang 1984, pp. 342ff.).

10. Finite dynamic games with continuous strategy space and static representations of dynamic games

10.1 INTRODUCTION

From the previous chapter it became evident that in a static setting there is *underprovision* of the public good 'environmental quality'. In the Nash equilibrium global emissions are too high from a global point of view. Thus one may wonder whether more positive results could be obtained in a dynamic, though *finite* time horizon. This question will be analyzed within two approaches.

The first approach remains in the tradition of repeated games, as encountered in previous chapters. More precisely, we proceed as in Chapter 4: first the equilibrium (or the equilibria) of the constituent game is determined; and second, one investigates whether the finitely repeated play of the stage game leads to more optimistic results. Again, a simultaneous and a sequential move version of the constituent game can be distinguished. In the former case the analysis is simple. If the constituent game is the global emission game described in Chapter 9, where there is a unique Nash equilibrium (NE) due to assumption A_1, the repeated play of this stage game NE is the only equilibrium in a finite time horizon. This is an immediate implication of Theorem 4.2. In the case of sequential moves, it is shown in Section 10.2 that there is a unique subgame-perfect equilibrium (SPE) of the emission stage game and hence, again, by Theorem 4.2, the finitely repeated play of this stage game equilibrium is the only equilibrium of the overall game. Thus, the outcome in the finitely repeated emission game does not differ from that if the game is played only once. Hence, the main point to be clarified in the subsequent analysis is whether the outcome in the simultaneous or in the sequential move game is preferable from a global point of view. In Section 10.2 this is done for a *filterable externality* (as assumed in Chapter 9) and for the case of a *transferable externality* in Section 10.3. Crudely speaking, the term transferable externality refers to a 'policy of high stacks' where emissions are not truly reduced but merely transferred to third parties.

The second approach departs from the assumptions of previous chapters and may be called the *static representations of dynamic games* or *conjectural variation models*. Here we discuss work done by scholars in the field of *public goods economics*. Among the many contributions in the literature we focus on the theory of *non-Nash behavior*, also called *hybrid behavior* (Cornes and Sandler 1983, 1984a, b, 1985a, b) in Section 10.4, *strategic matching* (Guttman 1987, 1991; Guttman and Schnytzer 1992) in Section 10.6, and the *principle of reciprocity* (Sudgen 1984) in Section 10.7.

Section 10.5 describes an *auctioning game* on emission reduction. On the one hand, this section serves as a preparation in that it provides preliminary information used in Section 10.6 and also in that it builds a 'methodological bridge' connecting Sections 10.4 and 10.6. On the other hand, fundamental results obtained for the auction equilibrium will be reconsidered in the context of infinite dynamic games in Chapter 12. That is, the auctioning equilibrium will be interpreted as the outcome of negotiations leading to an IEA where the agreed terms have to be stabilized by appropriate threats of punishments within a supergame framework.

The *motivation* of the papers pursuing the second approach is twofold: first, from a theoretical point of view, the authors apparently feel unhappy that a model can only explain the underprovision of public goods but is not capable of depicting the possibility of a sufficient or overprovision of public goods; second, from an empirical point of view, they argue that in some cases the negative predictions of a static public goods model are not confirmed. Instances cited in this literature are fund-raising for charities where agents contribute more than projected by 'Nash behavior'. Following this line of argument, in the present context this implies that some more or less successful IEAs could not be explained within a static framework.[1]

In particular, these scholars argue that contributions to a public good by some agents should be expected to influence the behavior of others. Provided own contributions are based on a positive *conjecture* with respect to the response by others, agents would increase their efforts to provide a public good. If contributions can be mutually observed, then positive conjectures may mitigate the free-rider problem.

Though it has to be stressed that not all these papers pay particular attention to the game theoretical interpretation of their models, we shall scrutinize them carefully with respect to their consistency employing game theoretical methodology. We shall argue that these models possess three main shortcomings: first, regardless of whether they are interpreted in a static or in a dynamic setting, they are not consistent; second, if they are interpreted in a dynamic context (which seems to be the more 'natural' interpretation), their static representation is deficient in capturing the main

forces of the strategic interaction between agents; third, the assumptions of some of these models are not in accordance with the notion of non-cooperative game theory. Taken together, we pose the question *whether a dynamic process would not be better modeled as a dynamic game, rather than confine its analysis to a static framework* (see, for example, Friedman 1986; Makowski 1983).

Despite these reservations, the conjectural variation models will be discussed in this chapter for two reasons: (a) most of these approaches capture an interesting phenomenon in a simple manner; and (b) these models have been influential in the literature on the provision of public goods. In a modified version they appear in many coalition models which will be discussed in Chapter 13.

The first point criticized above can be discussed at this preliminary stage (and the reader is invited to confirm this point in the later presentation), whereas the other points will be treated in the course of the presentation of these models. We first show the inconsistency of the assumptions for the static interpretation and then for the dynamic interpretation.

In the terminology of the conjectural variation literature, 'Nash behavior' is a special type of conjecture, namely a *zero conjecture*. In the NE each agent maximizes his/her payoff, given the strategies of the other players. Hence, this *could* be interpreted as if each player assumes that his/her strategy choice has *no* influence on the choice of the others. Consequently, as long as reaction functions are positively or negatively sloped, this conjecture is *not* confirmed, which *may* be taken as inconsistent behavior. Put differently, why should players be ignorant of the reaction of other players when choosing their best strategy? Would it not be more 'natural' to expect that they already incorporate this information in their strategy choice?

Though at first thought this may be a compelling argument, it is false in a strictly game theoretical context. Due to the assumption of complete information each player knows the best reply of each player and, what is important, s/he also knows that the strategy combination in the NE is a *stable equilibrium*. Though we used a cobweb type of argument as an auxiliary device to illustrate an NE, after all, in the very meaning of a *static* framework this is a one-shot game. Consequently, the interpretation of reaction functions in a static game *must* therefore be very narrow: a reaction function contains information about the best strategy of player i, for all *possible* strategy combinations of other players. However, possessing complete information, a player can identify the best replies of other players with respect to his/her own strategy and, as a rational player, will match his/her strategy accordingly. The previous comparative static analysis of reaction functions is therefore *only* to be interpreted as the influence of exogenous parameters on the location of an equilibrium.

From this it may appear that non-zero conjectures are more convincing in the context of infinite dynamic games and *incomplete information* where 'players learn gradually about their economic environment' (Cornes and Sandler 1985a, p. 125) and the conjectural equilibrium may be interpreted as a long-term steady state (Sabourian 1989, pp. 86ff.). However, this interpretation is also not very convincing for two reasons: (a) an explicit learning story is missing in these models; (b) all models assume complete information. If they assumed incomplete information, then a natural way of modeling this would be to assume that players assign a probability distribution over different possible strategies of their fellows when choosing their best reply. The probabilities would be revised in the course of the game if new information became available to players. However, as will become evident, all models are deterministic in that players expect a very specific reaction from their fellow players throughout the game, which is not revised (though new information becomes available).

10.2 SEQUENTIAL MOVE EMISSION GAME: FILTERABLE EXTERNALITIES

Consider a stage game where the payoff functions of countries are given by (9.1). Furthermore, suppose that assumption A_1 holds and assume without loss of generality only two countries for simplicity. Let country i move in the first stage and country j in the second stage. Since country i knows country j's best reply of the second stage, that is, $e_j(e_i)$, it can incorporate this knowledge in its objective function when choosing its equilibrium emissions, e_i^{ST}, in the first stage (see (10.1)). Given these emissions, country j chooses $e_j = e_j(e_i^{ST})$ in the second stage. Borrowing a term of oligopoly theory, player i may be called the *Stackelberg leader*, player j the *Stackelberg follower* and the SPE of this game a *Stackelberg equilibrium*, which also explains the superscript ST.[2] Thus, the objective functions of the leader and follower read:

$$\pi_i = \beta_i(e_i) - \phi_i(e_i + e_j(e_i)) \text{ and } \pi_j = \beta_j(e_j) - \phi_j(e_i + e_j). \tag{10.1}$$

from which the FOC in the Stackelberg equilibrium (assuming an interior solution):[3]

$$\begin{aligned} \pi_i' &= \frac{\partial \beta_i}{\partial e_i} - \frac{\partial \phi_i}{\partial e_i} - \frac{\partial \phi_i}{\partial e_j}\frac{\partial e_j}{\partial e_i} = 0 \Leftrightarrow \frac{\partial \beta_i}{\partial e_i} - \frac{\partial \phi_i}{\partial e_i}(1 + k_i) = 0,\ k_i = \frac{\partial e_j}{\partial e_i}; \\ \pi_j' &= \frac{\partial \beta_j}{\partial e_j} - \frac{\partial \phi_j}{\partial e_j} = 0 \end{aligned} \tag{10.2}$$

follow.[4] From (10.2) it is apparent that for the follower the maximization problem does not change compared to the assumption of simultaneous moves. The FOC of the leader, however, now contains an additional term (see (9.9)). In the conjectural variation context, k_i could be interpreted as the conjecture of the leader about the reaction of the follower.[5] In the present context, this parameter is nothing else than the slope of country *j*'s reaction function. Due to A_1, $-1<k_i<0$ holds. The implications in the Stackelberg equilibrium may be summarized as follows:[6]

Proposition 10.1
In the Stackelberg (stage game) equilibrium of the global emission game with payoff functions (9.1) the Stackelberg leader *i* increases emissions and the follower *j* reduces emissions compared to the Nash equilibrium, assuming conditions A_1 to hold. Global emissions will exceed those in the Nash equilibrium of the simultaneous move game. Global welfare will be lower than in the Nash equilibrium if the follower has the same or higher marginal abatement costs than the leader in the Nash equilibrium. The leader gains and the follower loses from the strategic move.

Proof: See Appendix VII.2. QED

Proposition 10.1 is intuitively appealing and is illustrated with the help of Figure 10.1.

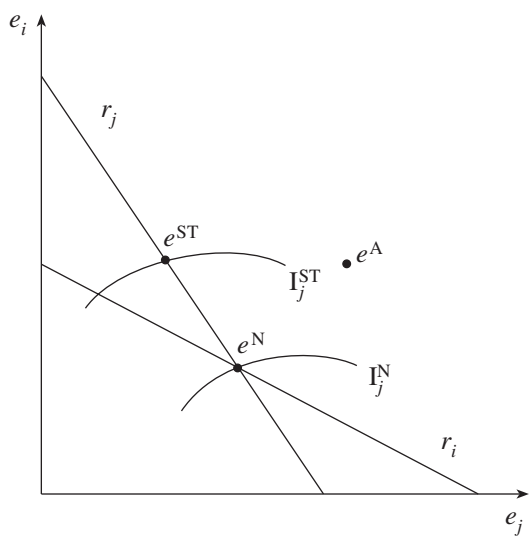


Figure 10.1 Stackelberg equilibrium

First, note that the Stackelberg equilibrium is located on the follower's reaction function r_j since his/her FOC have not changed compared to the simultaneous move game.

Second, due to $k_i < 0$, the leader expands emissions compared to the NE since s/he values marginal damage less (see (10.2)). Accordingly, the follower's best reply implies reducing emissions compared to the NE. Since $-1 < k_i$, global emissions are higher than in the NE.

Third, the emission tuple e^{ST} implies a lower welfare to the follower compared to the NE – his/her indifference curve in the Stackelberg equilibrium, I_j^{ST}, lies north-west of his/her indifference curve in the NE, I_j^{N}. By the same token, the leader's welfare is higher in the Stackelberg equilibrium than in the NE (country i's indifference curves are not drawn in Figure 10.1).

Fourth, the direction of global welfare is generally undetermined. However, if the leader has lower (or equal) marginal abatement costs than the follower in the NE global welfare will be lower in the Stackelberg equilibrium. The strategy tuple e^{ST} implies that the gap between marginal abatement costs widens compared to e^{N}, implying a higher inefficiency. Moreover, since global damages are higher in the Stackelberg compared to the Nash equilibrium, welfare must be lower. Consequently, one can deduce immediately that in the case of symmetric countries global welfare will be lower in the Stackelberg equilibrium than in the NE. In contrast, if the leader is that country with the higher marginal abatement costs in the NE, the sequential move *may* increase efficiency, which could compensate for higher damages, so that overall welfare could be higher in the Stackelberg equilibrium than in the NE.[7]

As pointed out in the Introduction, since the Stackelberg stage game equilibrium is unique (see also note 3), only this equilibrium will be played throughout the finitely repeated play of the emission game. Thus, also for the finitely repeated version of the emission game, one should hope from an ecological perspective that countries would move simultaneously. However, as demonstrated above, from a welfare perspective such general superiority cannot be established. From a single country's perspective sequential moves are clearly preferable, provided it can move first.

A general problem of Stackelberg equilibria is the motivation of the assumption that one player (group of players) moves first, that is, has superior information compared to the follower(s). Since each player has an incentive to be the leader, one would expect players to compete for this position. Therefore, as long as this leader–follower relationship is not obvious from the investigated problem itself, some doubts about the reason for such an asymmetry remain.[8] If each country assumes the supposed leadership and chooses a high emission level, both countries may end up neither on country i's nor on j's reaction function (see, for example, point e^{A} in

Figure 10.1), which cannot be an equilibrium. Hence, it seems plausible to expect that in this 'competition case' countries are led back to the NE. In particular in the context of repeated games one would expect 'on average' both countries to hold the leader position with an equal chance, and hence simultaneous moves seem a 'natural' assumption.

10.3 SEQUENTIAL MOVE EMISSION GAME: TRANSFERABLE EXTERNALITIES

Shogren and Crocker (1991) and Shogren *et al.* (1992) point out that sometimes countries or regions reduce their emissions at the cost of *higher* emissions in other countries and regions. For instance, this could happen if a government does *not* pursue an environmental policy which either fosters preventative measures or 'truly' reduces emissions based, for instance, on an end-of-pipe technology, but induces firms to build high stacks. Such 'abatement measures' do not reduce externalities but merely *transfer* them to third parties. Following the terminology of Shogren and Crocker this type of externalities are called *transferable externalities* in contrast to *filterable externalities* which we have considered so far. Of course, a transfer strategy can only work in the case of impure public pollution, such as acid rain, since for pure global pollutants only total emissions are relevant to damage in a country and not their regional distribution. Hence, in this section (and only in this section) we depart from our previous assumption of a global pollutant.

A simple explanation for a 'transfer policy' could be that the costs of transferring externalities is cheaper than truly curbing emissions. This case causes some fundamental changes compared to the previous analysis. Without loss of generality, we again restrict the analysis to two countries.

With respect to A_1 the assumptions regarding the benefit functions do not have to be changed. The same holds true for the effect of own emissions on damage costs. That is, $\partial\phi_i/\partial e_i > 0$ and $\partial^2\phi_i/\partial e_i^2 > 0$ are assumed. Now, however, we have $\partial\phi_j/\partial e_i < 0$, $\partial^2\phi_j/\partial e_i^2 > 0$ and $\partial^2\phi_i/\partial e_i\partial e_j < 0$. That is, *reducing* emissions in country i increases damage in country j at an increasing rate. At low levels of foreign emission reductions, marginal damage caused by own emissions is lower than at higher levels. These modifications lead to the following results which are clearly distinct from those obtained in Chapter 9 and Section 10.2.[9]

Proposition 10.2

In a transferable externality game with payoff function $\pi_i = \beta_i(e_i) - \phi_i(e_i, e_j)$, where the benefit function has the properties enumerated in

assumptions A_1 and for the damage cost function $\partial\phi_i/\partial e_i>0$, $\partial\phi_j/\partial e_i<0$, $\partial^2\phi_i/\partial e_i^2>0$, $\partial^2\phi_j/\partial e_i^2>0$ and $\partial^2\phi_i/\partial e_i\partial e_j<0$ hold and own effects are stronger than foreign effects, that is, $|\partial\phi_i/\partial e_i|>|\partial\phi_i/\partial e_j|$ and $|\partial\phi_i^2/\partial e_i^2|>|\partial\phi_i/\partial e_i\partial e_j|$, global emissions in the (stage game) Nash equilibrium are lower than in the Stackelberg equilibrium and the social optimum.

Proof: See Appendix VII.3. QED

There are three remarks to be made with respect to this result:

1. The above result relies on a *given technology* and does not take into consideration the possibility of a switch to another abatement policy which might be globally more efficient. In other words, the transferable externality technology is viewed as the only option to obtain a socially optimal solution.
2. Whether aggregate emissions in the Stackelberg equilibrium exceed or fall short of socially optimal emissions depends on the specific functions. As in the case of filterable externalities, emissions in the Stackelberg equilibrium are higher than those in the Nash equilibrium. A general comparison of global welfare in the NE with that in the Stackelberg equilibrium is not possible in the case of asymmetric countries for the same reason as laid out in Section 10.2. Thus, again, from an ecological perspective whether the game is played once or repeatedly, simultaneous moves are superior to sequential moves. With respect to global welfare such a clear-cut conclusion is not possible.
3. It is interesting to note that in the case of transferable externalities reaction functions are upward-sloping in emission space, as illustrated in Figure 9.4 in the previous chapter. To ensure a unique Nash equilibrium, $|\partial\phi_i^2/\partial e_i^2|>|\partial\phi_i/\partial e_i\partial e_j|$ has been assumed in Proposition 10.2, which is a sufficient condition so that the slopes of the reaction functions have a slope less than 1 and Theorem 9.2 can be applied. The details are laid out in Appendix VII.3.

Taken together, in a transferable externality game of the Shogren and Crocker type the main conclusions of an ordinary international pollution game (filterable externality game) with respect to emissions in the social optimum and the Nash equilibrium are reversed. Whereas in filterable externality games aggregate emissions are too high in the Nash equilibrium, in a transferable externality game they are too low from a global point of view. Though, of course, transferable externality games may be regarded as an exception rather than the rule in international politics, this

case neatly motivates positively sloped reaction functions, as already discussed in Chapter 9.

10.4 NON-NASH OR HYBRID BEHAVIOR

10.4.1 Technical Preliminaries

The assumption that players hold conjectures about how other players react to their strategies may be written in its most general form as (see Cornes and Sandler 1983, 1984b):

$$\frac{\partial e_{-i}}{\partial e_i} = f^{C}_{-i}(\Lambda, e_i, e_{-i}) \tag{10.3}$$

where the responsiveness by others is a function of a bunch of factors which are summarized in the parameter (or vector of parameters) Λ and the level of the strategy of players i and $-i$. To demonstrate how the non-zero conjectural approach develops its full potential in capturing the interaction of agents, it is helpful to adopt an example by Cornes and Sandler (1983) to the present context. The example assumes:

$$\frac{\partial e_{-i}}{\partial e_i} = \left(\frac{e_i}{e_{-i}}\right)^{\Theta} \tag{10.4}$$

where the relation in brackets expresses the conjecture that low emissions in country i will induce other countries to reduce their emissions as well. In particular, large emitters have a greater influence on others to follow suit than if a country's share of total emissions is small.[10] The parameter Θ may be interpreted as the *elasticity of the conjectured response with respect to the relative importance of a country's emission level*.[11]

Assuming a symmetric game allows us to study the effect of the number of countries on the outcome.[12] Since for symmetric countries $e_{-i} = (N-1)\cdot e_i$, (10.4) becomes:

$$\frac{\partial e_{-i}}{\partial e_i} = \left(\frac{1}{(N-1)}\right)^{\Theta}. \tag{10.5}$$

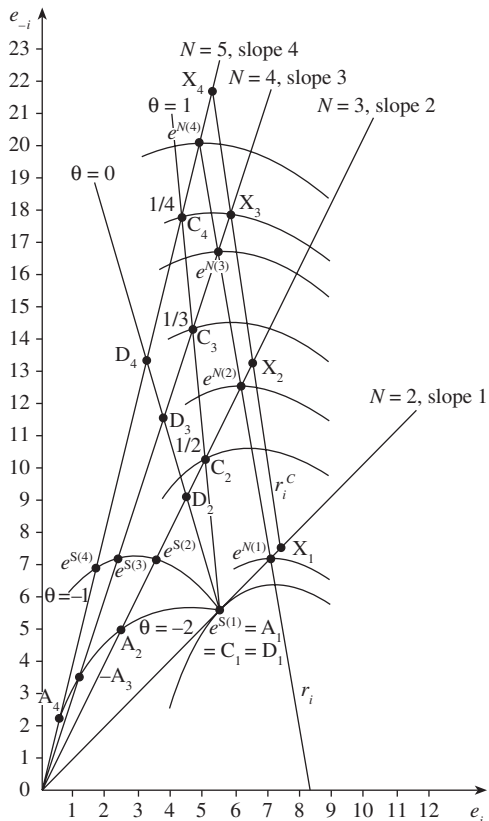
That is, the expected conjectured reaction diminishes with the number of countries. In other words, the greater the number of countries involved in the pollution problem, the less is the emission reduction of a single country noticed and the less is the free-riding incentive mitigated. The first-order conditions implied by (10.5) are (assuming an interior solution):[13]

$$\frac{\beta_i' - \phi_i'}{\phi_i'} = \left(\frac{1}{(N-1)}\right)^{\Theta}. \tag{10.6}$$

That is, in the optimum the slope of a country's indifference curve (LHS term in (10.6); see also (9.27)) is equal to the conjecture (RHS term in (10.6)). Since in (10.6) a positive conjecture is assumed, the optimum is located on the ascending part of a country's indifference curve.

10.4.2 Illustration

Figure 10.2 illustrates the effect of the elasticity parameter Θ on the conjecture of countries and global emissions. On the abscissa, emissions of a representative country i, e_i, are measured, along the ordinate those of the other $N-1$ countries, e_{-i}. The figure is based on the payoff functions in (9.3) and assumes the parameter values of Chapter 9 (without restricting N to two countries).



Note: Payoff functions (9.3), $b=5$, $c=1$ and $d=10$ are assumed.

Figure 10.2 Conjectural variation outcomes

For different values of Θ and N the conjectural variation outcomes are drawn. For instance, for $\Theta = 1$, C_1, C_2, C_3 and C_4 are the outcomes if $N = 2$, $N = 3$, $N = 4$ and $N = 5$ respectively. The slopes of the indifference curve at these points are 1, 1/2, 1/3 and 1/4 respectively, which reflects condition (10.6). For $\Theta = 0$ outcomes D_1, ..., D_4 are derived and $\Theta = -1$ implies socially optimal emission levels which are denoted by $e^{S(1)}$, ..., $e^{S(4)}$.[14] For values below $\Theta = -1$, for example, $\Theta = -2$, emissions in the conjectural outcome (A_1, ..., A_4) are even below socially optimal levels. The Nash equilibrium for different values of N is indicated by $e^{N(1)}$, ..., $e^{N(4)}$ where r_i denotes country i's Nash reaction function, as has already been encountered in Chapter 9. Due to the symmetry assumption, the Nash equilibrium and any other outcome lie on the same ray for a given N. The slope of such a ray is $N - 1$.

From Figure 10.2 it is evident that the higher the elasticity of the response (the smaller the value of Θ), the lower will be global emissions in the conjectured outcome, Σe_i^C, where $\Sigma e_i^C \to 0$ if $\Theta \to -\infty$. By the same token, the lower the response elasticity (the greater the value of Θ), the closer are global emissions to those in the Nash equilibrium, that is, $\Sigma e_i^C \to \Sigma e_i^N$ if $\Theta \to \infty$ implying $\partial e_{-i}/\partial e_i$ approaches zero).

Moreover, it is apparent that for any $\Theta \geq 1$, $\Sigma e_i^C \to \Sigma e_i^N$ if $N \to \infty$. Thus, this specification can capture Olson's assertion that free-riding increases in group size (Olson 1965; see also Sandler 1992, pp. 23ff.). In fact, this relation holds not only when comparing a conjectured outcome and the Nash equilibrium for different values of N but also if one compares social optimal emissions and those in the Nash equilibrium. Defining an index of free-riding as the ratio e^N/e^S, then it is apparent from Figure 10.2 that this ratio (measured as $\overline{0e^N}/\overline{0e^S}$ in Figure 10.2) increases in N. That is, the underprovision of the public good 'clean environment' is particularly devastating if many countries are involved in the problem. This central result of public goods economics will be given particular focus in Chapters 13 and 14 on coalition formation.

10.4.3 Welfare and Equilibrium Analysis

Generally, any *positive* conjecture will lead to a welfare improvement as long as emissions do not exceed those in the social optimum. For symmetric countries this is true as long as $0 < \partial e_{-i}/\partial e_i \leq N - 1$ holds or with reference to specification (10.5) $-1 \leq \Theta < \infty$ is true.[15] Thus moderate positive conjectures lead to welfare improving outcomes. However, the question arises whether such outcomes are stable. To answer this question, we first have to define a conjectural variation equilibrium (see Sabourian 1989, pp. 86ff.).

Definition 10.1: Rational conjectural variation equilibrium
Let $c_j(a_i', a_{-i})$ denote what i conjectures j will do in response to a deviation by i from action combination $a = (a_i, a_{-i})$ to a_i' and let the best response of a player be given by $b_i(a, c_j) = \arg\max_{a_i' \in A_i} \pi_i(a_i', c_j(a_i', a_{-i}))$, then a rational conjectural variation equilibrium (RCVE) is a pair of actions and conjectures for which (1) $a_i^* \in b_i(a^*, c)$ and (2) $c_j(a_i', a_{-i}^*) \in b_j(a_i', c_j(a_i', a_{-i}^*))$ $\forall$ i, j; $i \neq j$ and $\forall$ $a_i' \in A_i$ hold.

1. The action combination a may be interpreted as the status quo and hence a_i' describes a deviation from the status quo. The status quo plays the role of a *state variable* and players believe that the response of the other depends on it. The best replies are thus functions of the state variable and the conjectured responses.
2. Players are concerned only with the immediate payoffs after other players have responded to a deviation. Thus players are either myopic or the new status quo has to be interpreted as a new long-run steady state. The latter interpretation seems more appealing and allows us to view the whole conjectural variation story as resembling an infinitely repeated game.
3. Players' strategies depend only on the actions chosen in the previous round. Thus some kind of bounded recall is implicitly assumed in the conjectural variation context, though an explicit justification is missing.
4. There are no transitory gains to be obtained since by assumption players can instantaneously respond to a deviation. This assumption is difficult to justify since it basically renders the free-rider incentive in international pollution control a problem of minor importance. Sabourian (1989, pp. 87ff.), noticing this deficiency, suggests a way out of this motivational dilemma. He tells a time-explicit story in which adjustment costs for the deviator are prohibitively high so that deviation does not pay. Then, however, the question arises as to why the other players who are assumed to react to a deviation do not also face these costs. More generally, it seems to us, by assuming adjustment costs to be sufficiently great any deviation from a status quo can be ruled out and therefore any equilibrium can trivially be supported. Thus, taken together, we believe that the conjectural variation story captures the main forces of the strategic interaction between agents in international pollution control only deficiently.[16]
5. The requirement that conjectures must be confirmed in equilibrium (second requirement in the definition above) is a very similar condition to a Nash equilibrium in 'conventional' dynamic games. If this requirement were changed in the definition above to $c_j(a_i', a_{-i}) \in b_j(a_i', c_j(a_i',$

a_{-i})) $\forall\, i, j;\ i \neq j$ and $\forall\, a_i' \in A_i$, that is, the rationality requirement holds at every status quo and not only in equilibrium, then the RCVE would even resemble a subgame-perfect equilibrium in a 'conventional' dynamic game. Hence, by imposing the rationality requirement on a conjectural equilibrium the time-implicit story of the conjectural variation setting is placed in a consistent framework.

With the help of Definition 10.1, we can now give an answer to the question whether a conjectural variation outcome is stable as stated above:

Proposition 10.3
In a global emission game there is only a non-zero conjectural variation equilibrium with negative conjectures. Global emissions will be higher than in the Nash equilibrium. In a symmetric game, country-specific and global welfare in the conjectural variation equilibrium are lower than in the Nash equilibrium.

Proof: The *first assertion* is proved by applying the definition of an RCVE. In the present context, country i's best reply in the conjectural variation outcome is defined by its reaction function, r_i^C, and hence part (2) of Definition 10.1 requires (see also Cornes and Sandler 1983, 1984b):[17]

$$r_i^{C\prime} = \frac{\phi_i''(1 + k_i)}{\beta_i'' - \phi_i'' k_i (1 + k_i)} = k_i \tag{10.7}$$

to hold where for simplicity we assume that $\partial e_{-i}/\partial e_i = k_i$ does not depend on the level of e_i and e_{-i} (as in the example (10.5)) and is therefore constant, that is, $f_i^C(\Lambda, e_i, e_{-i}) = k_i$. In other words, we linearize the slope of the reaction function around the conjectured equilibrium. Since $r_i^{C\prime} = 0$ for $k_i > 0$, equality (10.7) cannot hold for positive conjectures. Only for negative conjectures can the equality be satisfied.

The *second assertion* of $\Sigma e_i^C > \Sigma e_i^N$ can be shown by comparing the FOC in the conjectured equilibrium for $k_i < 0$ (see note 17) and those in the NE (see (9.9)) using a similar argument in the spirit of the proof of Proposition 10.1. The *third assertion* follows from strictly concave payoff functions, $\partial \pi_i/\partial e_i < 0$ and $\partial \pi_i/\partial e_{-i} < 0$ at emissions above e^N and the fact that for symmetric games $\Sigma e_i^C > \Sigma e_i^N > \Sigma e_i^S$ and $e_i^C > e_i^N > e_i^S\ \forall\, i \in I$. QED

Thus, in the context of global pollutants the expectation that positive conjectures lead to more optimistic results is disappointed as long as *consistency of conjectures* is required in a 'non-Nash-behavior equilibrium' (see Cornes and Sandler 1985a; Sudgen 1984, p. 773). In fact, non-Nash behavior makes it more difficult to explain voluntary contributions to the

provision of the international public good clean environment.[18] This can also be seen in Figure 10.2 where the stable conjectured equilibrium is indicated by X_1, ..., X_4 for various assumptions of N.[19,20]

From Proposition 10.2 it follows that the Nash equilibrium is generally not stable in the conjectural variation context. That is, if reaction functions are interpreted as a dynamic adjustment process, the NE must (correctly) be rejected as a stable outcome since $k_i=0$ and $-1<r_i'<0$ by assumption A_1. Only for the case of ϕ_i'' close to zero and/or $\beta_i'' \to\infty$, which implies orthogonal reaction functions ($r_i' \to 0$), would conjecture $k_i=0$ be confirmed at the limit. Also the Stackelberg equilibrium is not stable in the conjectural variation context. Though the leader's negative conjecture is confirmed, this is not true for the follower's zero conjecture.[21]

10.5 AUCTIONING EMISSION REDUCTIONS[22]

10.5.1 General Considerations and Technical Preliminaries

In the previous section we showed that governments choose their emissions according to the conjectures they hold about how other countries respond to their abatement policy. For constant conjectures the responsiveness was measured by the parameter $k_i=\partial e_j/\partial e_i$. In terms of emission reductions, r_i, this could alternatively be written as $k_i=\partial r_j/\partial r_i$. Thus if country i reduces emission by Δr_i country j is supposed to adjust its emission by $k_i\cdot\Delta r_j$. Hence, k_i may be viewed as an *exchange rate* by which r_i and r_j (or e_i and e_j) are linked. From country i's perspective $1/k_i$ is the price to obtain 1 unit of reduction by country j. Accordingly, one could imagine an *auction* where an auctioneer calls up different exchange rates k_i (prices $1/k_i$) and where countries respond by offering reduction levels. This process continues until the offers of countries match which may be called the *auction equilibrium*.

In the following it will prove convenient to restrict the number of countries to two and to denote the exchange rate by μ. Further we define $r_2=\mu\cdot r_1$ where r_i denotes the fraction of emission reduction from some initial emission level e_i^{I}. Hence, emissions are given by $e_i^{\mathrm{A}}=e_i^{\mathrm{I}}(1-r_i)$ where $0\le r_i\le 1$. $100\cdot r_i$ measures emission reduction as a percentage from an initial emission level. The superscript A stands for auction. Consequently, countries 1 and 2's net benefit functions may be written as:

$$\begin{aligned}\pi_1&=\beta_1(e_1^{\mathrm{I}}(1-r_1))-\phi_1(e_1^{\mathrm{I}}(1-r_1)+e_2^{\mathrm{I}}(1-\mu\cdot r_1))\\ \pi_2&=\beta_2(e_2^{\mathrm{I}}(1-r_2))-\phi_2(e_1^{\mathrm{I}}(1-r_2/\mu)+e_2^{\mathrm{I}}(1-r_2))\end{aligned}\qquad(10.8)$$

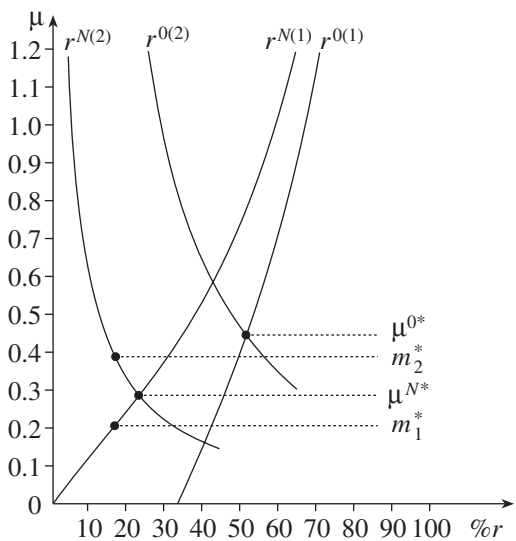
and the *implicit offer curves* of countries follow from differentiating (10.8) with respect to r_i:[23]

$$\frac{(\beta_1' - \phi_1')e_1^{\mathrm{I}}}{\phi_1' e_2^{\mathrm{I}}} = \mu, \quad \frac{\phi_2' e_1^{\mathrm{I}}}{(\beta_2' - \phi_2')e_2^{\mathrm{I}}} = \mu. \tag{10.9}$$

For an example (10.9) can be solved explicitly.

10.5.2 Illustration

The example illustrated in Figure 10.3 assumes the net benefit functions of (9.3), $N=2$, $b_1=b_2=10$, $c_1=5$, $c_2=1$ and $d=10$. In the figure use has been made of the fact that global reduction r, that is, $r=r_1+r_2$, implies $r=(\Sigma e_i^{\mathrm{I}} - \Sigma e_i^{\mathrm{A}})/\Sigma e_i^{\mathrm{I}}$ or $r=r_1(e_1^{\mathrm{I}}+\mu\cdot e_2^{\mathrm{I}})/(e_1^{\mathrm{I}}+e_2^{\mathrm{I}})$. Country i's offer may thus be



Note: Offer curves are based on payoff functions (9.3), assuming $b_1=b_2=10$, $c_1=5$, $c_2=1$, $d=10$.

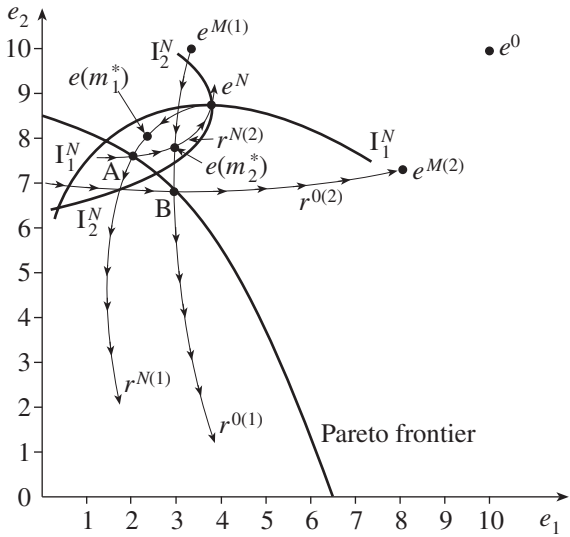
Figure 10.3 Emission reduction offer curves in abatement space

written as $r^{(i)}=r(\mu)$ or, as in Figure 10.3, expressed as a percentage reduction, that is, $r^{(i)}=\%r(\mu)$. $r^{0(i)}$ implies initial emissions $e_i^{\mathrm{I}}=e_i^0$, $r^{\mathrm{N}(i)}$ denotes the offer if initial emissions are given by $e_i^{\mathrm{I}}=e_i^N$.

From Figure 10.3 it is evident that country 1's offer increases in the exchange rate, whereas country 2's offer decreases. At the *equilibrium exchange rates* μ^{N*} and μ^{0*} offers of both countries match. In the example $\mu^{N*}=0.285$ which implies an emission reduction of 23.2 percent from Nash

equilibrium emissions. For $\mu^{0*} = 0.4475$ equilibrium emission reduction is 51.2 percent based on initial level $e_i^{\mathrm{I}} = e_i^0 = d$.

It is also possible to depict the offer curves in the e_1–e_2 space. This is done in Figure 10.4. The arrows indicate an increase of the exchange rate μ. For example, if $e_i^{\mathrm{I}} = e_i^0 = d \; \forall \; i \in \mathrm{I}$ and $\mu = 0$, country 1 assumes that country 2 does not contribute anything to emission reduction. Consequently,



Note: Offer curves are based on payoff functions (9.3), assuming $b_1 = b_2 = 10$, $c_1 = 5$, $c_2 = 1$, $d = 10$.

Figure 10.4 Emission reduction offer curves in emission space

country 1's best reply delivers the minimax emission tuple $e^{\mathrm{M}(1)}$. Raising μ leads country 1 to reduce emissions because the effective price for the public good clean environment falls. By symmetry, a small μ, that is, $\mu < 1$, constitutes a favorable situation for country 2 since each unit of emission reduction is accompanied by a large reduction in country 1 (left part of the $r^{0(2)}$ curve). Therefore, country 2's offer implies low aggregate emissions. If μ is raised gradually, country 2 has to contribute a higher portion to joint emission reduction and is therefore only prepared to accept a lower reduction level (right part of the $r^{0(2)}$ curve).[24]

From Figure 10.4 it is evident that at the points where the offer curves intersect (see points A and B), that is, in equilibrium (and only in equilibrium) the conditions along the Pareto frontier hold. That is, the *auction leads to a Pareto-efficient emission allocation*. Emissions in each country will be lower in the auction equilibrium compared to the initial situation. That is, at point A (B) emissions are lower than at $e^N(e^0)$. Moreover, payoffs

in the auction equilibrium are higher than in the initial situation. This becomes apparent from Figure 10.3, assuming $e_i^{\mathrm{I}} = e_i^N$, where the auction equilibrium emission tuple lies within the lens formed by indifference curves I_1^N and I_2^N.

10.5.3 Results and Discussion

The main result illustrated in Figure 10.3 may be summarized in the following proposition:

Proposition 10.4
Let e^{A} be the equilibrium emission tuple of an auction of emission reductions as described above and let e_i^{I} denote initial emission levels, then e^{A} constitutes a Pareto optimum. In equilibrium, emissions in each country and global emission are lower than in the initial situation, that is, $e_i^{\mathrm{A}} < e_i^{\mathrm{I}}$ and $\Sigma e_i^{\mathrm{A}} < \Sigma e_i^{\mathrm{I}}$ and individual and aggregate welfare are higher than in the status quo, that is, $\pi_i(e^{\mathrm{A}}) > \pi_i(e^{\mathrm{I}})\ \forall\ i \in \mathrm{I}$, and $\Sigma\pi_i(e^{\mathrm{A}}) > \Sigma\pi_i(e^{I})$ provided $e_i^{\mathrm{I}} \geq e_i^N > 0\ \forall\ i \in \mathrm{I}$. Each country receives a payoff in excess of its minimax payoff, $\pi_i(e^{\mathrm{A}}) > \pi_i^{\mathrm{M}}(e_i^{\mathrm{M}(i)}, e_j^{\mathrm{M}(i)})\ \forall\ i \in \mathrm{I}$.

Proof: The first part of Proposition 10.3 follows simply by comparing (10.9) with (9.31). The second part is proved in Appendix VII.5. QED

There are four remarks with respect to the *auction equilibrium*:

1. Each country has a strong incentive to misrepresent its preferences. Knowing that the final agreement is defined by $r^{N(1)}(\mu) = r^{N(2)}(\mu)$ or $r^{0(1)}(\mu) = r^{0(2)}(\mu)$ respectively, country 1 and country 2 will bias their offers downward for a given μ. That is, country 1 likes to shift its offer curve in a north-westerly direction and country 2 in a south-westerly direction in Figure 10.3 so that an equilibrium exchange rate is reached which is in a country's favor (high μ^* from country 1's perspective, low μ^* from country 2's perspective). If both countries behave strategically, the final outcome may differ from the 'true' equilibrium. The direction of the bias of the equilibrium exchange rate will depend on the extent of the misrepresentation.

   However, misrepresentation may pay only to some extent. For instance, if country 2 shifts its offer curve south-west for strategic reasons, the equilibrium exchange rate will, *ceteris paribus*, decrease and the implied global emission reduction will fall. If the latter effect becomes too strong, environmental damage may become so severe as to overcompensate the favorable terms of the exchange rate to

country 2. By symmetry, a similar argument holds for country 1's strategic considerations. Hence, one has to reckon with strategic offers within the vicinity of the equilibrium exchange rate with a 'natural' limit of biased offers.[25]

Thus taken together, strategic offers may cause the exchange to be biased upward or downward, though equilibrium emission reduction will undoubtedly be biased downward (the equilibrium reduction level in Figure 10.3 moves to the left). Then equilibrium emissions will be located above the Pareto frontier. In Chapter 11 we shall take up again the issue of strategic offers and discuss a decision mechanism which gets around this problem.

2. The auction equilibrium is *not* stable in a static context by the definition of a Nash equilibrium. It is also not stable in the conjectural variation context since it has been shown in Section 10.4 that positive conjectures (=exchange rates) are not compatible with negatively sloped reaction functions. Thus an auction equilibrium is only an equilibrium in that offers match with μ^* the clearing price in the emission market; however, it is not an equilibrium in the sense of a stable IEA.
3. Despite this qualification one could imagine that the auction is a first step where parties agree on an emission reduction which will then be backed within the wider context of a dynamic game. In other words, the auction could depict negotiations leading to an IEA where parties agree on an emission tuple within the large set of stationary supergame equilibria which must be enforced in subsequent periods by appropriate threat strategies. For the possibilities of enforcement it is important that payoffs of an agreement are not too asymmetrically distributed, otherwise threats may be insufficient to deter free-riding. Though the details will be discussed in Chapter 12 on dynamic games, Proposition 10.4 provides useful background information in stating that the auction equilibrium is a Pareto optimum and implies a Pareto improvement compared to the status quo.
4. As is apparent from Figure 10.4, the basis on which exchange rates are defined is crucial in determining which emission tuple is reached in equilibrium. Clearly, in the example, country 1 prefers the auction to be based on reductions from emission $e_i^{\mathrm{I}} = e_i^0$, whereas country 2 would like to see reductions to be related to $e_i^{\mathrm{I}} = e_i^N$. Unfortunately, conclusions about the preferences of countries at a more general level do not seem possible. The same is true with respect to global emissions in the auction equilibrium. In the example $\Sigma e_i(\mu^{N*}) < \Sigma e_i(\mu^{0*})$ but this does not have to be generally true.[26]

10.6 STRATEGIC MATCHING

10.6.1 Introduction

The *strategic matching* model is either set up as a two-stage game in its simultaneous move version or may have more than two stages when viewed sequentially. However, all stages are subsumed into one stage game and hence instantaneous reactions by all participants are assumed as in the conjectural variation context.

The model of Guttman (1987, 1991) and Guttman and Schnytzer (1992) is *supposedly* capable of explaining *positive contributions* to a public good *above* Nash levels within a *non-cooperative* framework without resorting to any altruistic or intrinsic motivation. Their model builds on the observation that sometimes individuals contribute to the provision of a public good if they observe that others have already contributed some amount or promise to do so. For instance, regional governments may invest in infrastructure if there is a promise by the central government of financial support. The central government in turn may make its promise conditional on the commitment of the regional government to contribute its share to the project.

10.6.2 Technical Preliminaries

Technically, each party i determines *matching rates* m_{ij} towards player j in a first step, and then, based on these matching rates, determines its *flat contribution* in a second step, that is, its actual investment. Thus the matching rate is some kind of a commitment in order to induce the other players to make a positive contribution. Denoting total emission reductions of all countries by r, total emission reduction of country i by r_i, the flat contribution of country i by $\bar{r}_i$ and the matching rate m_{ij} where country j commits itself to match country i's contribution by m_{ij} units, then the following simple relations hold:

$$r_i = \bar{r}_i + \sum_{i \neq j}^{N} m_{ij}\bar{r}_j;\; r = \sum_{i}^{N} r_i = \sum_{i}^{N} \bar{r}_i + \sum_{i}^{N}\sum_{i \neq j}^{N} m_{ij}\bar{r}_j. \tag{10.10}$$

Thus total emission reductions in country i depend on its flat contribution and the matching rates of all other countries. Payoffs are assumed to be a function of reduction levels, that is, $\pi_i(r_1, \ldots, r_N)$, and the standard FOC and SOC are presumed to apply. The derivation of the equilibrium will be explained for the *two-stage simultaneous move* version of the game – the extension to *sequential moves* should be obvious.

Due to the finite time horizon, the game can be solved by *backwards*

induction. That is, first, equilibrium flat actions are determined, assuming the matching rates to be *given*. The solution may be denoted by $\bar{r}^* = (\bar{r}_1^*(\mathrm{M}), \ldots, \bar{r}_N^*(\mathrm{M}))$ where M is an $(N) \times (N-1)$ matrix of matching rates where in each row the matching rates of a country *vis-à-vis* the other $N-1$ countries is listed and $\bar{r}^*$ is a vector of flat contributions of dimension N.[27] Hence, $\bar{r}_i^*(\mathrm{M})$ are the best replies of the second stage.

Second, the solution of the second stage is inserted into the payoff functions and one solves for the optimal set of matching rates, M^*. Substituting M^* into the best replies determined above, $\bar{r}_i^*(\mathrm{M})$, delivers the 'subgame-perfect equilibrium' of the overall game which is given by the two strategy sets $\bar{r}^*$ and M^*. Together with (10.10) overall equilibrium reduction levels r_i^* and r^* can be computed.

Once the best replies have been determined, it suffices to check stability of a strategy combination by focusing on M^* solely. Stability is then defined in the standard fashion of a Nash equilibrium in which no player has an incentive to change its matching rates. That is, $\pi_i(r(\mathrm{M}^*)) \geq \pi_i(r(\mathrm{M}^*_{-i}, m_{ij}))$ $\forall\ i$ and $j \in \mathrm{I}$, $i \neq j$ must hold where M^*_{-i} denotes the matrix of equilibrium matching rates except that one element is different, that is, $m_{ij} \neq m_{ij}^*$.

Of course, other specifications than the linear relation as given in (10.10) may also seem plausible but are far more difficult to work with. For most games of economic interest a linear relation between activity levels of different agents can hardly be solved by *simulations*, let alone *analytically*. Therefore, Guttman and Schnytzer devote most of their attention to show that, for a broad class of positive and negative externality games, socially optimal matching rates constitute a subgame-perfect equilibrium if moves occur simultaneously. However, it appears to us that this is trivially satisfied in almost all games by the construction of the equilibrium concept. For sequential move games, the equilibrium concept exhibits even more severe shortcomings. In particular, it seems difficult to establish the existence of an equilibrium without imposing some restriction on the strategy space.

10.6.3 Equilibrium Determination and Discussion

Taking a *sequential view* for simplicity first, a government, say, 2, could suggest to government 1 a matching rate m_{12}, promising that it would match government 1's reduction effort by a reduction of m_{21} times $\bar{r}_1$. Since in the initial stage $\bar{r}_1 = 0$ and $\bar{r}_2 = 0$, this would imply $r_1 = \bar{r}_1$ and $r_2 = \bar{r}_2 = m_{12} \cdot \bar{r}_1$. Hence, m_{12} is *de facto* the exchange rate μ encountered in the auction described in the previous section. Consequently, due to the assumption of complete information, government 2 knows country 1's *offer curve* and can suggest an *optimal matching rate* m_{21}^* which, due to the restriction to a two-country world, may be denoted m_1^* for short. If initial emissions are given

by $e^{\mathrm{I}} = e^N$, $r^{N(1)}$ is country 1's offer curve and the resulting emission tuple is $e(m_1^*)$ as shown in Figure 10.4. By the same token, one can determine government 1's matching rate m_2^* which it offers to country 2 provided it has the first move. This would deliver point $e(m_2^*)$ in Figure 10.4. The associated matching rates m_1^* and m_2^* are also depicted in Figure 10.3.

To highlight the main forces at work, parts of Figure 10.4 have been reproduced in Figure 10.5 on a greater scale. In order to keep the relations

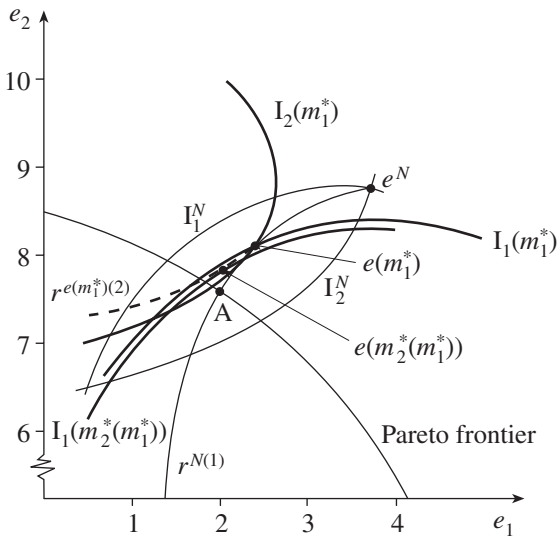


Figure 10.5 Sequential move matching game

in Figure 10.5 tractable, only country 1's offer curve, $r^{N(1)}$, is shown. Now it is evident how the offer m_1^* and therefore $e(m_1^*)$ is determined. Country 2 chooses m_1^* such that $e(m_1^*)$ is located on the highest indifference curve of country 2, $\mathrm{I}_2(m_1^*)$, which is tangential to country 1's offer curve $r^{N(1)}$. Country 1's welfare level at $e(m_1^*)$ is represented by indifference curve $\mathrm{I}_1(m_1^*)$ and that of country 2 by $\mathrm{I}_2(m_1^*)$. Since both indifference curves lie within the lens formed by the Nash equilibrium indifference curves, I_1^N and I_2^N, both countries have obtained a higher welfare level than in the initial situation. It is also apparent that the indifference curves $\mathrm{I}_1(m_1^*)$ and $\mathrm{I}_2(m_1^*)$ themselves form a lens which indicates that there is room for further mutually beneficial offers.

In the next step, *based on* $e(m_1^*)$, country 1 can offer country 2 that it will match an additional reduction by m_2^*. To determine m_2^*, country 1 chooses

that point on country 2's offer curve which delivers country 1 the highest obtainable welfare. This is $e(m_2^*(m_1^*))$ where indifference curve $I_1(m_2^*(m_1^*))$ is tangential to the offer curve $r^{e(m_1^*)(2)}$. Again, if one were to draw country 2's indifference curve (not shown in Figure 10.5), it would turn out that a further Pareto improvement is possible. Hence, country 2 has an incentive to commit itself to a matching rate based on $e(m_2^*(m_1^*))$.

It seems *intuitively* plausible that by extrapolating this process both countries will finally reach the Pareto frontier. An exact statement is difficult since, when determining matching rate offers at some stage, the first-order conditions deliver more than one solution (in the example, four). Though some of these solutions can be discarded since they define a minimum for the proposing country, negative matching rates may well be in the interest of a country.[28] Hence, if negative matching rates are not ruled out *a priori*, then there might be cyclical offers, and it is not clear whether a final equilibrium will be reached.

However, if negative matching rates are deleted from the solution set, the Pareto frontier will be reached. By the definition of emission tuples along the Pareto frontier, there is no positive matching rate which could improve welfare of *both* countries and no further offer will be accepted. Thus, in this *narrow sense* (ignoring negative matching rates) the matching equilibrium would be stable though this procedure stretches the equilibrium concept to the limit (and hence renders it almost meaningless). From the discussion it is also evident that for a complete determination of the equilibrium the *sequence of moves* (matching path) and the *initial stage* must be specified *ex ante*. Both aspects of the practical determination of an equilibrium (restriction of strategy space and specification of the details of the game) seem to be unnoticed by Guttman and Schnytzer.

In the case of *simultaneous moves* equilibrium determination may also be difficult. Therefore, Guttman and Schnytzer (1992) focus on socially optimal matching rates and show that for the case of two players these matching rates may constitute an equilibrium.[29] The intuition is obvious in the case of symmetric countries where $m_1^* = m_2^* = 1$. If a country were to reduce its matching rate, then emissions in both countries would increase, which implies a welfare loss both to the deviator and to the country sticking to its commitment alike. By the same token, an increase in the matching rate would imply that the opportunity costs of abatement increase more than damage decreases. Thus, Guttman and Schnytzer's finding that socially optimal matching rate equilibria may exist in positive or negative externality games is not particularly surprising.

A more general problem with the strategic matching approach (which the authors claim to be capable of solving the deadlock inherent in externality games while staying within the realm of *non-cooperative game* theory) is

that its 'positive' result crucially depends on the assumption that matching commitments are fulfilled. For instance, if in the above-mentioned case of symmetric countries matching rates are kept at the socially optimal equilibrium rates but countries are allowed to alter their implied reduction level, which seems a natural conjecture within a non-cooperative setting, such an equilibrium is no longer stable. It is hard to see how such an equilibrium can be established without some external enforcement power. Of course, by the construction of the game any deviation is *instantaneously* matched by countermeasures which, in a 'true' dynamic setting, would be called punishments. Due to immediate reactions, the deviator cannot obtain any transitory gains as this is possible in a discrete time setting. Hence, the free-rider problem is trivially solved.[30]

In a strict sense, Guttman and Schnytzer's logic implies in an ordinary static prisoners' dilemma game that each country affirms the fellow player that it will match his/her action 'cooperation' by 'cooperation', and refrains from taking a free-ride because then both would end up in the non-cooperative and Pareto-inferior Nash equilibrium. Since we did not accept this logic previously, the question arises why we should accept it now, though, admittedly, the result comes about in a neatly disguised term called strategic matching.

Thus, we are led back to the question raised in the Introduction to this chapter: *Why not model a dynamic process as a dynamic game?* It seems to us that one does *not* gain any new insight by using such an approach compared to the traditional models discussed at some length in Chapters 4–8 in the context of discrete strategies (though Guttman and Schnytzer 1992, pp. 74ff., claim exactly this to be the case). In particular, in contrast to the previous models of this chapter, one can hardly argue that the advantage of the strategic matching approach over the 'conventional' dynamic approach lies in its greater simplicity.

What seems more convincing to us is to interpret strategic matching as a negotiation process leading parties to sign an IEA. Like a minimum participation clause which is part of many treaties (that is, the treaty only becomes binding if a minimum of the signatories have deposited their signature with the treaty secretariat; see Black *et al.* 1992), matching rates may be interpreted as the minimum commitment of reduction obligations by each party. Compliance would then not be defined in absolute terms but towards each party via matching rates. However, the treaty would have to be enforced by the conventional type of threat and punishment strategies within an infinitely repeated game framework.

10.7 THE THEORY OF RECIPROCITY

Sudgen's theory of reciprocity sets out by noting that a *principle of unconditional commitment*, that is, contributing to a public good, irrespective of whether other players make a contribution, is unrealistic.[31] Sudgen (1984, pp. 774–5) writes: 'Whatever the force of the principle of unconditional commitment at the level of moral theory, it is hard to see it taking root as a maxim of *practical* morality – as a maxim on which ordinary people are prepared to act.'

Therefore, he proposes his theory of reciprocity which he summarizes as follows:

> Let G be *any* group of people of which *i* is a member. Suppose that every member of G except *i* is making an effort of at least ξ in the production of some public good. Then let *i* choose the level of effort that he would most prefer that every member of G should make. If this most preferred level of effort is not less than ξ, then *i* is under the obligation to the members of G to make an effort of at least ξ.

Thus, no member of some group is expected and actually will not contribute more than others in that group, but also not less. Hence, there is no problem of free-riding within a group.

Sudgen sets out to analyze the case of the symmetric preferences of players and then discusses the more general case of asymmetric preferences. Since the symmetric case is sufficient to demonstrate our main concern with this theory, we restrict our attention to this case.

In the global emission context a contribution may be interpreted as emission reduction r_i, where $e_i = (1 - r_i) \cdot e_i^{\mathrm{I}}$ and $r = \Sigma r_i$ as in the previous section. In the symmetric country case, the equilibrium contribution is $r^* = N \cdot r_i^*$. Let r_i^{L} be that reduction level where $\pi_i'(r_i^{\mathrm{L}}, r_{-i} = 0) = 0$ and $\pi_i = \beta_i(\Sigma r_j) - \phi_i(r_i)$ be a strictly concave function in r_i. Now Sudgen (1984, p. 778) argues that for the equilibrium contribution $r^{\mathrm{L}} \leq r^* \leq r^{\mathrm{S}}$ must hold, where the subscripts L and S stand for lower bound and social optimum respectively. $r^{\mathrm{L}} < r^*$ is necessary to protect the self-interest of a country if it has no expectations that others reciprocate. $r^* < r^{\mathrm{S}}$ must hold because otherwise countries contribute more than they are obliged to do. In sum, he postulates that: (a) r^* must be somewhere between r^{L} and r^{S}; (b) there is more than one equilibrium; (c) one equilibrium is socially optimal; and (d) all other equilibria imply underprovision of the public good.

Though these conclusions are correct, at least if one chooses the 'right' interpretation of the equilibrium path, they are not terribly exciting. Following Sudgen's arguments, and taking a sequential point of view, one should expect $r^* > r^N$ since $\pi_i'(r_i^{\mathrm{L}}, r_{-i} = 0) > \pi_i'(r_i^{\mathrm{L}}, r_{-i}^N > 0)$. This implies that

in a first step each country chooses at least $r_i^{\mathrm{L}} > r_i^N$ which, according to the theory, it cannot adjust downward. Then, however, it is not ensured that $r_i^{\mathrm{L}} < r_i^{\mathrm{S}}$ holds, as Sudgen claims. That is, the overprovision of the public good is possible. If this sequential view is given up in favor of a simultaneous move game, then of course $r_i^{\mathrm{L}} = r_i^N$ by definition.

Sudgen notes that any contribution above r^{L} faces the assurance problem where each country likes to contribute if others will follow suit. Thus, it seems that Sudgen has the second interpretation in mind. If so, then his theory does not lead to more 'positive' results. If, however, the first interpretation was correct, then $r_i^{\mathrm{L}} > r_i^{\mathrm{S}}$ may lead to a welfare loss if the overprovision is pronounced enough. Moreover, the principle of reciprocity implies that players accept a minimum of obligations, which also requires players to be guided by some moral motives. Admittedly, the moral requirements are not as strong as assuming unconditional commitment which Sudgen (rightly) criticizes, but, as we think, his requirements are also not convincing in a game theoretical context.

NOTES

1. Of course, from previous chapters it is known that an extension to an infinite dynamic time horizon is capable of solving this putative deficiency. For the emission game this will be demonstrated in Chapters 12 and 14.
2. In oligopoly literature, assuming competition in quantities, if players move sequentially, this is referred to as Stackelberg competition (von Stackelberg 1934). If all players choose their quantities simultaneously, this is referred to as Cournot competition (Cournot 1938); see also Koutsoyiannis (1991) and Malinvaud (1985). Hence, in some environmental economics literature the NE is sometimes referred to as the Cournot–Nash equilibrium (for example, Endres 1993, 1997). The similarity to the emission model should be obvious: output is equivalent to emissions; the output of player i exhibits a negative externality on player j via a price drop. All players would be better off by forming an output cartel and by reducing supply.
3. To keep the discussion brief we consider only an interior Stackelberg equilibrium in the following. That is, we assume that the equilibrium is not restricted by the boundaries of the strategy space. In particular, for the Stackelberg leader $e_i^{\mathrm{ST}} < e_i^{\max}$ and for the follower $e_j(e_i^{\mathrm{ST}}) \geq 0$ are presumed to hold. Moreover, we assume that the second-order conditions in the Stackelberg equilibrium hold (which is not automatically ensured). Then it follows immediately that the Stackelberg equilibrium is unique. See Appendix VII.1 for details.
4. Recall $\partial\phi_i/\partial e_i = \partial\phi_i/\partial e_j = \partial\phi_i/\partial e$ due to the assumption of a pure public bad type of pollutant.
5. k_i may also be interpreted as a weighting factor of marginal damages.
6. A similar result can be obtained for transboundary pollutants as well.
7. For the example in (9.2), assuming $b_i = b_j = b$ and $c_i = c_j = c$, we find $e_i^{\mathrm{ST}} = [d(b^2 + bc + c^2)]/(b^2 + 3bc + c^2)$, $e_j^{\mathrm{ST}} = [d(b^2 + bc - c^2)]/(b^2 + 3bc + c^2)$ and $\Sigma e_k^{\mathrm{ST}} = [2d(b + c)]/(b^2 + 3bc + c^2)$. Comparing these emissions with those in the Nash equilibrium, as given in (9.6), confirms $\Sigma e_k^{\mathrm{ST}} \geq \Sigma e_k^N$. Due to the symmetry of this example $\beta_1'(e_1^N) = \beta_2'(e_2^N)$ and so one can immediately conclude that global welfare is lower in the Stackelberg equilibrium than in the Nash equilibrium.

8. See the general discussion in Section 2.3 and, in the context of the chicken game, Section 4.3.
9. $|\partial\phi_i/\partial e_i| > |\partial\phi_i/\partial e_j|$ is assumed in Proposition 10.2 to guarantee an interior Stackelberg equilibrium. Again, the second-order conditions in the Stackelberg equilibrium are assumed to hold, which may not generally be true. See Appendix VII.3 for details.
10. Of course, one could also specify the share in brackets as $e_i/\Sigma e_i$ and could derive the same qualitative results as presented below. However, in the later computations it turns out that the specification in (10.4) is more convenient to work with.
11. Let $\partial e_{-i}/\partial e_i = k_i$ as in Section 10.2, then:
$$\Theta = \frac{dk_i}{d(e_i/e_{-i})} \cdot \frac{e_i/e_{-i}}{k_i}.$$
12. The term outcome instead of equilibrium is used intentionally, since so far it has not been checked whether the conjectured outcome is stable. We shall turn to this issue below.
13. This follows from $\beta_i' - \phi_i' - \phi_i' \cdot (\partial e_{-i}/\partial e_{-i}) = 0$, substitution of (10.5), and rearrangement of terms.
14. For $\Theta = -1$, $\partial e_{-i}/\partial e_i = N-1$ and the first-order condition in the conjectured outcome (see previous note) becomes $\beta_i' - \phi_i' - \phi_i' \cdot (N-1) = 0 \Rightarrow \beta_i' - \Sigma\phi_i' = 0$.
15. This follows from the facts that (1) $0 < \partial e_{-i}/\partial e_i < N-1$ implies $e_i^S < e_i^C < e_i^N \forall i \in \mathrm{I}$ in the symmetric country case; (2) $e_i^S \forall i \in \mathrm{I}$ maximizes country specific and global net benefits; and (3) net benefit functions are strictly concave in emissions by A_1. For $\partial e_{-i}/\partial e_i$ *well above* $N-1$ (below $\Theta = -1$ in the example) welfare in the conjectured outcome may be *below* that in the Nash equilibrium since opportunity costs of abatement are far above optimal levels.
16. This is the second point of the claimed shortcomings of conjectural variation models which are listed in the Introduction to this chapter.
17. This follows from totally differentiating the first-order condition $\beta_i' - \phi_i'(1 + k_j) = 0$ in the conjectured optimum.
18. Cornes and Sandler (1985a, pp. 128ff.), claim that more positive results could be obtained in the case of impure public bads. To see this, assume for instance a transboundary pollutant and only two countries. Let payoff functions be given by $\pi_i = \beta_i(e_i) - \phi_i(a_{ii}e_i + a_{ij}e_j)$ where $0 \leq a_{ii} \leq 1$ and $0 \leq a_{ij} \leq 1$ are transportation coefficients, a_{ii} indicating that portion of emissions from country i which remains in this country and a_{ij} that portion of emissions from country j which is transported to country i. Then (10.7) reads $[\phi_i''(a_{ii}a_{ij} - a_{ij}^2 k_i)]/[\beta_i'' - \phi_i''(a_{ii}^2 + a_{ii}a_{ij}k_i)] = k_i$ which can be satisfied for positive conjectures provided $a_{ii}a_{ij} - a_{ij}^2 k_i < 0 \Leftrightarrow a_{ii}/a_{ij} < k_i$ (see also Cornes and Sandler 1984a).
19. For the example employed in Figure 10.3, $r_i^{C'} = -\dfrac{c(1+k)}{b + c(1+k)}$ ($k_i = k_j = k$ due to symmetry) and the rationality requirement implies $r_i^{C'} = k$ to which the solutions are $k = (2c + b \pm \sqrt{4bc + b^2})/2c$. For the parameter constellation of Figure 10.2 it turns out that $k = -0.146$ is the only stable conjecture. The second solution leads to emissions outside the domain of the strategy space.
20. When deriving a negative conjecture equilibrium one has to keep in mind that neither the first- nor the second-order conditions for an interior solution may generally hold. This is demonstrated in Appendix VII.4. See also Cornes and Sandler (1984b, p. 375) on this point.
21. Note that the conjectural variation approach allows us to link the Nash equilibrium, the social optimum and the Stackelberg equilibrium to each other in a *symmetric* global emission game in a simple manner (Varian 1984, Section 2.10). We have: Nash equilibrium: $k_i = 0 \ \forall \ i \in \mathrm{I}$; social optimum: $k_i = N-1 \ \forall \ i \in \mathrm{I}$ and Stackelberg equilibrium: $k_i = 0$, $k_j = r_j'$ where the FOC for an interior solution reads $\beta_i' - \phi_i'(1 + k_i) = 0$.
22. The basic idea of this section goes back to Nentjes (1994).
23. We show below in the proof of Proposition 10.3 that the SOC are always satisfied. See Appendix VII.5.
24. Aggregate emissions implied by a particular offer can simply be measured by drawing a

45° line through an emission tuple (not shown in Figure 10.4). The point where this line cuts the ordinate or the abscissa gives aggregate emissions.

25. From country 1's perspective $\mu = m_2^*$ would be the optimal exchange rate; from country 2's perspective this is $\mu = m_1^*$.
26. The reason is that from (10.9) such general conclusions cannot be drawn. Even for the specific functions (9.3), the equilibrium price and resulting emissions cannot be obtained analytically any more (except for symmetric countries) but only by simulations.
27. Of course, M^* could also be defined as an $N \times N$ matrix where the diagonal elements are 1.
28. This is also noted by Guttman (1987, pp. 9ff.). See the arguments when comparing (A10.6) with (A10.7) in Appendix VII.5 in deducing whether r_i, as the solution to the FOC, will be positive.
29. This equilibrium is characterized by matching rates $m_{ji}^* = [\partial \pi_j / \partial r_i] / [\partial \pi_j / \partial r_j]$. If actor i maximizes $\pi_i(r_i, r_j)$ with respect to $\bar{r}_i$, noting $r_i = \bar{r}_i + \Sigma m_{ij} \bar{r}_j$ and $r_j = \bar{r}_j + \Sigma m_{ji} \bar{r}_i$, the FOC read: $\partial \pi_i / \partial \bar{r}_i = (\partial \pi_i / \partial r_i)(\partial r_i / \partial \bar{r}_i) + (\partial \pi_i / \partial r_j)(\partial r_j / \partial \bar{r}_i) = 0$, from which $\partial \pi_i / \partial \bar{r}_i = (\partial \pi_i / \partial r_i) + (\partial \pi_i / \partial r_j) \cdot m_{ji} = 0$ follows. Substitution of m_{ji}^* from above delivers $(\partial \pi_i / \partial r_i) + (\partial \pi_j / \partial r_i) = 0$, which are the FOC in the social optimum.
30. This weakness is even more striking since the authors do not impose any rationality constraint of the sort described in the conjectural variation context. See Definition 10.1.
31. Sudgen's paper sets out to criticize Margolis's 'Theory of Unselfish Behavior' (Margolis 1982). Basically this 'theory' assumes that an agent is not only concerned about his/her own utility but also about that of the fellow players. Essentially, this theory postulates altruistic or any kind of moral motivation of individuals when contributing to the provision of a public good. Like Sudgen, we feel that this approach is too simple to explain voluntary contributions. If players were altruistic, free-riding in international cooperation would hardly be a problem worth analyzing.

11. Bargaining over a uniform emission reduction quota and a uniform emission tax

11.1 INTRODUCTION

In Section 10.5 we discussed how an auction of emission reductions could work. An auctioneer calls up different exchange rates which define the relation of emission reductions of countries and asks the participants for offers. The auction equilibrium was defined as that exchange rate where offers match. Thus, the bargaining market clears via an adjustment of 'prices'.

In reality, however, a different bargaining rule can frequently be observed at the pre-stage leading to an IEA. The *exchange rate is fixed* through an institutional framework and countries agree on the lowest bid, that is, on the *lowest common denominator*. For instance, potential signatories to an IEA frequently negotiate on a *uniform emission reduction quota,* which implies that countries have to reduce emissions by the same percentage compared to some base year. Typically that country which proposes the lowest reduction will be the 'bottleneck' in the negotiations and defines the terms of the agreement.

The list of examples of uniform *emission quotas* is long and includes the Montreal Protocol on Substances that Deplete the Ozone Layer, which specified an emission reduction of CFCs and halons by 20 percent based on 1986 emission levels to be accomplished by 1998.[1] Another example is the Helsinki Protocol, which suggested a reduction of sulfur dioxide from 1980 levels by 30 percent by 1993. Moreover, the Sofia Protocol Concerning the Control of Emissions of Nitrogen Oxides or Their Transboundary Fluxes signed in 1988 called on countries to *uniformly* freeze their emissions at 1987 levels by 1995 and the Geneva Protocol Concerning the Control of Emissions of Volatile Organic Compounds or Their Fluxes signed in 1991 required parties to reduce 1988 emissions by 30 percent by 1999.[2]

The question why *uniform* abatement obligations play such a prominent

role in international politics has not yet been answered, to our knowledge. Nevertheless, it has been suggested that uniform solutions are apparently perceived to be 'fair' and therefore find relatively easy acceptance. Uniform solutions thus constitute some kind of a focal point in the sense of Schelling (1960).[3] Moreover, to agree on differentiated solutions may take time and is therefore associated with higher transaction costs (Endres 1996a). However, in the sequel we do not model transaction costs or psychological motives. These arguments just serve to motivate the focus of this chapter on 'uniform solutions'.

The *lowest common denominator decision rule*, henceforth abbreviated to LCD decision rule, can frequently be observed in international politics (Endres 1996a, b; Unerdal 1998a, p. 6, and 1998b, p. 109). Typical examples comprise the voting procedure within the European Union or within the UNO on particular issues (Lenschow 1996). The reason for this decision rule in international politics is simple: due to the lack of an enforcement authority at the global level, the accession to IEAs and the agreement on abatement targets have to be voluntary. Moreover, if the LCD decision is applied an agreement can be reached with only limited information requirements (see Section 11.3). Applying the LCD decision rule also gives negotiators no incentive to misrepresent their preferences strategically (see Section 11.4).

From an economist's point of view it is surprising that uniform reduction quotas, which belong to the group of *command and control instruments*, are part of so many IEAs since they are generally inefficient. For a given agreed reduction level, marginal abatement costs will typically differ between countries. Therefore, economists have persistently argued that *market-based instruments*, such as effluent charges or emission permits, are better suited to achieve abatement targets *cost-efficiently*. However, this advice has not yet fallen on fertile soil. This chapter and the subsequent chapters present some evidence why this might be the case.

To achieve this task efficiently, we select only one market-based instrument for comparison, namely a *uniform emission tax,* and contrast this instrument with the quota. As we shall set out in Section 11.2, in the context of *global pollutants*, which we assume throughout this chapter, the assumption of a uniform application of a tax in all countries is in fact a necessary condition for cost-efficiency. As will become apparent below, a uniform tax also implies emission reductions at a *fixed exchange rate*, however, one which differs from that of a uniform quota. Similar to the quota regime, we consider a bargaining game in which countries settle for the lowest bid, that is, the lowest tax proposal.

The subsequent analysis of the quota and tax bargaining equilibrium focuses on two politically interesting questions:

1. Why are abatement targets within most IEAs rather low, though cost–benefit considerations would suggest lower emissions?
2. Why have effluent charges (as one representative of market-based instruments) not been applied in any IEA so far whereas quotas have found widespread application?

Initial answers to these questions will be given in this chapter and in Chapter 12; however, an extensive treatment of these questions will be left to Chapter 14 where this issue is investigated in the context of coalition formation. In the following, the cost-efficiency of the tax solution and the inefficiency of the quota solution is demonstrated in Section 11.2. In Section 11.3 the bargaining rule is discussed; in Section 11.4 the bargaining proposals are determined; and in Section 11.5 equilibrium emissions are derived. Section 11.6 compares equilibrium emissions and welfare levels; Section 11.7 discusses the possibility of strategic proposals; and Section 11.8 winds up the discussion with a short summary.

The subsequent analysis is confined to two countries for simplicity. The case of N countries will be treated in the context of coalition models (Chapter 14). It is assumed in this chapter that the general payoff functions ((9.1)) as well as the associated properties of these functions (assumptions A_1 in (9.2)) of Chapter 9 apply.

The bargaining model presented in this chapter is due to Endres (1996a, b). An excellent source of a less technical presentation of the model, where many institutional details are also discussed, is Endres (1995). An extension where emission permits are also covered can be found in Endres and Finus (1998a, 1999). The presentation of this chapter draws on Endres and Finus (1998b).

11.2 COST-EFFICIENCY OF THE SET OF INSTRUMENTS

As noted above, we want to contrast an inefficient regulatory instrument (quota) with a market-based instrument (tax) which develops its full cost-efficiency potential. Therefore, in the context of a pure public bad type of pollutant one cannot assume a tax which is differentiated across countries, but only a uniform tax rate t (see, for example, Barrett 1991a; Hoel 1992c).[4] To see this, note that efficiency requires that a given emission ceiling $\overline{\Sigma e_k}$ is achieved with maximal global benefits (minimal opportunity costs of abatement):

$$\max_{e_i,e_j} \beta_i(e_i) = \beta_j(e_j) \text{ s.t. } e_i + e_j \leq \overline{\Sigma e_k}. \tag{11.1}$$

From the Kuhn–Tucker conditions we receive (if the constraint is binding):

$$\text{CEA:} = \begin{cases} \beta_i'(e_i) = \beta_j'(e_j) \\ e_i = \overline{\Sigma e_k},\ e_j = 0 \Rightarrow \beta_i'(\overline{\Sigma e_k}) > \beta_j'(0) \\ e_i = e_j = \overline{\Sigma e_k} = 0 \Rightarrow \beta_i'(0) \text{ and } \beta_j'(0) \end{cases} \tag{11.2}$$

as an 'equilibrium condition' of a *cost-effectiveness analysis*, henceforth abbreviated CEA. Comparing (11.2) with the condition for social optimality (9.24) reveals that (11.2) is implied by (9.24); however, the opposite does not hold as long as $\overline{\Sigma e_k} \neq \Sigma_k^{\mathrm{S}}$.

Countries' industries, facing a uniform tax, solve:

$$\max_{e_i} \beta_i(e_i) - te_i \Rightarrow \begin{cases} t = \beta_i'(e_i^{\mathrm{T}}) \text{ if } t \leq \beta_i'(0) \\ e_i^{\mathrm{T}} = 0 \text{ if } t > \beta_i'(0) \\ e_i^{\mathrm{T}} = e_i^{\max} \text{ if } t = 0. \end{cases} \tag{11.3}$$

Consequently, comparing (11.2) and (11.3) it is evident that abatement *in the uniform tax regime is conducted cost-efficiently*. If t is chosen such that each country's industry adjusts to $e_i^{\mathrm{T}} = e_i^{\mathrm{S}}$ where the superscript T stands for tax regime, a uniform tax is not only cost-efficient but also socially optimal.

In contrast, a uniform quota implies:

$$\beta_1'(e_1^{\mathrm{I}}(1-r)) \neq (=)\beta_2'(e_2^{\mathrm{I}}(1-r)) \tag{11.4}$$

in equilibrium where we may recall that e_i^{I} denotes initial emissions in country i and $r \in [0, 1]$ the fraction of emission reduction. (Hence, $e_i^{\mathrm{Q}} = e_i^{\mathrm{I}}(1-r)$ where the superscript Q stands for quota regime.) Obviously, the equality holds only for very specific cases, for example, $\beta_1 = \beta_2$ and $e_1^{\mathrm{I}} = e_2^{\mathrm{I}}$ because then $\beta_1' = \beta_2'$ in equilibrium. Consequently, *an emission quota is generally not cost-efficient*.

11.3 THE BARGAINING SETTING

When negotiating on joint abatement, countries' representatives will base their proposals on payoff functions as given by (9.1). Basically, there are many possibilities for solving this bargaining problem. For instance, one could simply assume a socially optimal emission allocation (see, for example, Chander and Tulkens 1997), or apply a bargaining concept of cooperative game theory such as the Nash bargaining solution or the

Shapley value (see, for example, Barrett 1992b, 1997b; Botteon and Carraro 1997).[5] These concepts are well founded in the game theoretical literature, possess some interesting axiomatic properties and also satisfy some normative criteria. Among other factors, this explains their frequent applications in the environmental economics literature.[6]

A prerequisite of most cooperative bargaining concepts is that there is some *cooperative spirit* when it comes to choosing an abatement target and allocating abatement burdens. Additionally, a cooperative spirit is also required when the agreement is implemented since there is no supranational authority which could enforce compliance by all parties. Moreover, these concepts rest on the assumptions of *complete information* and *unlimited transfers*. Typically, they also determine a solution in terms of payoffs and abstract from the problem of how this is institutionally translated. That is, the choice of the policy instrument is not considered in these concepts. Since IEAs typically specify an abatement target below a Pareto-efficient level and employ inefficient instruments, this seems to suggest that some important features of international politics are not captured by those concepts.

Acknowledging these deficiencies, we propose to consider a two-stage game in which some of the above-mentioned assumptions are relaxed or changed. In the first stage – the *bargaining* stage – the two countries negotiate and agree either on the level of the emission quota or the emission tax. In the second stage – the *implementation* stage – the agreement is enforced. With respect to the particular assumptions the following modifications apply:

1. As mentioned in the Introduction to this Chapter, we assume that bargaining partners agree on the lowest common denominator to reflect the difficulties of agreeing on a joint abatement target.
2. Signatories have an incentive to take a free-ride due to the public good character of a joint abatement policy. Whereas the status quo is assumed to be the (unique) Nash equilibrium (NE),[7] an agreement implies a deviation from this equilibrium and therefore compliance cannot be assumed in an *ad hoc* manner. On the contrary, it has to be checked whether an agreement can be enforced by credible threat and punishment strategies. The details of this implementation stage will be discussed in Chapters 12 and 14. Thus, the bargaining equilibria derived in this chapter should be viewed as a method to pick some (plausible) stationary equilibria (such as the auction equilibrium; see the discussion in Section 10.5) from a larger set. Their dynamic stability properties will then be analyzed in subsequent chapters.[8]
3. The LCD decision rule makes it possible to relax the assumption of

complete information. Under the quota regime it suffices if each country knows initial emissions and, of course, its own payoff function in order to put forward a proposal. This does not seem a very restrictive assumption because for most pollutants data on current emissions are available.[9]

Under the *tax regime*, apart from the own payoff function, the opportunity costs of abatement in the neighboring country must also be known in order to compute how this country adjusts to a uniform tax rate. Thus, the information requirement is slightly higher than under a quota regime. However, under both regimes, it is not necessary; and in fact in most parts of this chapter (except Section 11.7) we do *not* assume that a country knows how environmental damages are evaluated in the neighboring country. It suffices to treat the game as one of *partial information*.[10] This assumption seems to approximate reality quite well. On the one hand, there are many empirical studies which have estimated abatement costs more or less reliably for various pollutants (for example, Fankhauser 1995; Fankhauser and Kverndokk 1996; Kaitala *et al*. 1991, 1992; Kverndokk 1993; Mäler 1989). This information is therefore accessible. On the other hand, estimates on damage costs are rare and entail many speculative elements.[11]

4. We rule out transfers. The lack of transfers in most IEAs has already been discussed in Chapter 8 in the context of issue linkage. Therefore, our reservation against bargaining solutions which are based on transfers does not have to be justified any further. Consequently, for the tax regime it is assumed that revenues remain in the country of origin. This seems to be suggestive because it is unlikely that countries will hand over their tax sovereignty to an IEA secretariat (Bohm 1994). For the quota regime we assume that quotas are not tradable because, up to now, no IEA is in force which allows for trade in reduction levels.

11.4 THE BARGAINING PROPOSALS

Under the *quota regime* a country's proposal, r_i, is derived from:

$$\max_{ri} \beta_i(e_i^{\mathrm{I}}(1-r_i)) - \phi_i(e^{\mathrm{I}}(1-r_i)). \tag{11.5}$$

That is, the exchange rate of the auction market is set to 1, that is, $\mu = 1$ (see (10.9)). According to the LCD decision rule, the agreed reduction level, denoted r^{A}, is equal to r_1 if $r_1 < r_2$ and is equal to r_2 if $r_1 > r_2$ where r_i denotes a country's proposal. With respect to Figure 10.3 depicting the offer curves in the auction market, this implies that one country (in the example

country 2) suggests an emission reduction below and one country (in the example country 1) above $r(\mu^{N*})$ ($r(\mu^{0*})$) at $\mu = 1$ if $e_i^{\mathrm{I}} = e_i^N$ ($e_i^{\mathrm{I}} = e_i^0$).

We call the country which makes the lowest bid the *bottleneck country* (which is in the following denoted country i). Further, as mentioned above, we assume that the status quo is represented by the NE, $e_i^{\mathrm{I}} = e_i^N \; \forall \; i \in \mathrm{I}$.[12]

Under the *tax regime* a country's proposal t_i follows from:

$$\max_{t_i} \beta_i(e_i(t_i)) - \phi_i(e_i(t_i) + e_j(t_i)). \tag{11.6}$$

where, again, according to the LCD decision rule, we have $t^{\mathrm{A}} = t_1$ if $t_1 < t_2$, $t^{\mathrm{A}} = t_2$ if $t_1 > t_2$ and $t^{\mathrm{A}} = t_1 = t_2$ if both proposals are identical.

The relation $e_i(t_i)$ and $e_j(t_i)$ is known from (11.3). Since $\partial e_i / \partial t_i = 1/\beta_i''$ (see Appendix VIII.1), the exchange rate defined in terms of emissions is given by $\partial e_i / \partial e_j = \beta_j''/\beta_i''$. For quadratic benefit functions this exchange rate is constant; for benefit functions of higher order the exchange rate is a function of emissions itself. Thus, each participant with information on $\partial e_i / \partial t_i$ and $\partial e_j / \partial t_i$ can compute an optimal uniform tax from its perspective.

The solution to (11.5) is unique since payoff functions are strictly concave with respect to a reduction level r_i. This has been shown in the proof of Proposition 10.4 (see Appendix VII.5) where strict concavity has been established for *any* positive exchange rate (and hence the case of $\mu = 1$ is covered as well). It turns out that a sufficient condition to establish strict concavity of the payoff function in the tax rate is to require that second-order effects dominate third-order effects (see for details Appendix VIII.1).[13] Therefore, we assume:

$$\mathrm{A}_2\text{: } \beta_i''' \text{ is sufficiently small.} \tag{11.7}$$

Consequently, the bargaining proposals under both regimes constitute a unique global maximum from each country's perspective and the LCD rule determines a unique bargaining outcome (which might differ from equilibrium emissions in some circumstances, as will be demonstrated in Section 11.5). We now can state our first result with respect to the LCD decision rule:

Proposition 11.1

Let $k_i \in \{r_i, t_i\}$ denote the bargaining proposals of country i derived from (11.5) and (11.6) under the assumption of partial information where each country knows that an agreement according to the LCD bargaining rule implies $k^{\mathrm{A}} = \min [k_i, k_j]$. Further, let $k_i' \neq k_i$ and $k_j' \neq k_j$ denote some arbitrary proposals of countries i and j and assume properties A_1

and A_2 to hold, then the lowest bid game has an equilibrium in dominant strategies, that is, $\pi_i(k_i, k_j') \geq \pi_i(k_i', k_j') \; \forall \; i \in I$.

Proof: Let k_i solve the problem $\max_{k_i} \pi_i(e_i, (k^A), e_j(k^A)), k^A = \min[k_i, k_j]$ and pick arbitrarily $k_j' \in K_j$.[14] Then if $k_i \leq k_j'$, this implies $\pi_i(k_i, k_j') \geq \pi_i(k_i', k_j')$ since k_i solves the maximization problem stated above. However, if $k_i > k_j'$, then if country i proposes $k_i' > k_j'$ instead, then bid $k_i' > k_j'$ leaves π_i unchanged, and a bid $k_i' \leq k_j'$ reduces π_i by strict concavity. QED

Proposition 11.1 implies that, assuming partial information, the LCD rule possesses the property that 'truth-telling' is a dominant strategy. Hence, it does not pay a country to make a strategically manipulated bid instead of its 'true' bid in order to bias the final agreement in its favor. This may reduce the transaction costs of bargaining. Other agreement procedures, such as agreeing on the arithmetic mean of the proposals or the auctioning equilibrium of Section 10.3, are generally more vulnerable to strategic behavior and may constitute neither an equilibrium in dominant strategies nor a Nash equilibrium. Countries may have an incentive to misrepresent their preferences in a first step, but also may have an incentive to alter their proposals once offers are disclosed. The result may serve as an additional explanation for the popularity of the (inefficient) LCD rule in international politics.

11.5 EQUILIBRIUM EMISSIONS

In the previous section we determined *bargaining outcomes* r^A and t^A according to the LCD rule. Now we are concerned with equilibrium emissions if the agreement is translated into policy. We define:

Definition 11.1: Compliance
Let e_j^k, $k \in \{Q, T\}$, denote equilibrium emissions in country j if the agreement is implemented, then the parties comply with the agreement if and only if $e_j^k \leq e_j(k^A)$. That is, countries may emit less than implied by the agreement but not more.

Generally, this definition imposes a restriction on countries' behavior since they have a free-rider incentive. As pointed out above, compliance will be checked in Chapters 12 and 14. However, as we show below, there may also be cases under the tax regime where there is an incentive for a party to do more than required by the agreement. This may happen if a country (non-bottleneck country) feels that the LCD decision rule leads to such a low

abatement target that it is in its interest to overcomply. Since such an action also exhibits a positive externality on the neighboring country ($\partial\pi_i/\partial e_j<0$), this kind of 'post-adjustment' is not regarded as a violation of the spirit of an agreement.

Definition 11.2: Incentive to free-ride and overcomply
Let $I_j := \beta_j'(e_j(k^A)) - \phi_j'(\Sigma e_k(k^A))$, then a country is said to have an incentive to free-ride if $I_j>0$. A country is expected to reduce emissions beyond its obligations if $I_j<0$.

The index I_j simply reflects a country's best reply as defined in (9.9) and where in the (interior) NE $I_j=0\ \forall\ j\in I$ holds. In both cases, whether a country adjusts or free-rides, we assume that a country's action is guided by its best reply function. In the case of an adjustment, *equilibrium* emissions are given by $e_i^k := e_i(k^A)$, $e_j^{k,a} := e_j(e_i(k^A)) < e_j(k^A)$ and $\Sigma e_k^{k,a} := e_i^k + e_j^{k,a}$ where the superscript *a* stands for adjustment. If no adjustment takes place and countries comply with their obligations, *equilibrium* emissions immediately follow from the agreement on k^A. That is, $e_i^k := e_i(k^A)$ and $e_j^k := e_j(k^A)$. Free-riding would imply (*out-of-equilibrium*) emissions $e_i^D := e_i(e_j(k^A)) > e_i(k^A)$ and $e_j^D := e_j(e_i(k^A)) > e_j(k^A)$ respectively where the superscript D stands for deviation.

Therefore, as a matter of terminology, we distinguish between a bargaining *outcome* and an *equilibrium*. Outcome refers to the agreement on k^A; equilibrium emissions or tax and quota equilibrium refer to emissions *after* possible adjustments have taken place.

With these definitions we can now state the following result:

Proposition 11.2
Let the bargaining outcome be determined by the LCD decision rule under partial information and allow countries to overfulfill but not to violate the terms of the agreement, then there is always a free-rider incentive under a quota agreement, that is, $I_j>0\ \forall\ j\in I$ and $r^A\in[0, 1]$, but never an incentive to overfulfill the terms of the agreement. Under the tax regime the non-bottleneck country may have an incentive to unilaterally reduce emissions below what is implied by the agreement.

Proof: Under the quota regime any agreement implies $r^A>0$. This follows from Appendix VII.5, where it has been shown for the auction market that offers will be positive for any exchange $\mu>0$ and under the quota regime $\mu=1$ holds. Since an (interior) NE would imply $r^A=0$ and $I_j=0$ (recall $e_j^I=e_j^N\ \forall\ j\in I$), $I_j=\beta_j'(e_j^N(1-r^A)) - \phi_j'(\Sigma e_k^N(1-r^A))>0$ for any $r^A>0$ must therefore hold.

In contrast, under the tax agreement $I_j = \beta_j'(e_j^T(t^A)) - \phi_j'(\Sigma e_k^T(t^A)) < 0$ (recalling that $e_j^{max} = e_j^T(0) \neq e_j^N$) is possible if t^A is sufficiently small. Consider payoff functions of countries 1 and 2 to be given by:

$$\pi_1 = b\left(de_1 - \frac{1}{2}e_1^2\right) - \frac{c}{2}(e_1+e_2)^2,\ \pi_2 = b\omega\left(de_2 - \frac{1}{2}e_2^2\right) - \frac{c}{2}\Theta(e_1+e_2)^2 \quad (11.8)$$

and hence the tax proposals of countries 1 and 2 are:

$$t_1 = \frac{2\omega bcd(\omega+1)}{b\omega^2 + \omega^2 c + 2c\omega + c};\ t_2 = \frac{2\omega bcd(\omega+1)}{b\frac{\omega}{\Theta} + \omega^2 c + 2c\omega + c}, \quad (11.9)$$

where $t_1 = t^A$ ($t_2 = t^A$) if $\Theta >= 1/\omega$ ($\Theta <= 1/\omega$) holds. Computing $I_i =: \beta_i'(e_i^T) - \phi_i'(e_i^T, e_j^T)$ reveals that $I_1 > 0$ for $1/\omega \leq \Theta < \infty$, $I_2 > 0$ for $1/\omega \leq \Theta < (\omega+1)/\omega$ and $I_2 < 0$ for $(\omega+1)/\omega < \Theta < \infty$ holds. Moreover, $I_1 < 0$ for $0 < \Theta < 1/(\omega+1)$, $I_1 > 0$ for $1/(\omega+1) < \Theta \leq 1/\omega$ and $I_2 > 0$ for $0 < \Theta \leq 1/\omega$. Consequently, for $(\omega+1)/\omega < \Theta < \infty$ country 2 adjusts, and for $0 < \Theta < 1/(\omega+1)$ country 1 adjusts. QED

Since we use the payoff functions in (11.8) in subsequent sections and also in Chapter 12, we briefly discuss their properties and the assumptions required in the following. Note that the functions in (11.8) imply $\beta_2 = \omega\beta_1$ and $\phi_2 = \Theta\phi_1$. Hence, the parameter ω allows us to model differences in opportunity costs of abatement in the two countries in a simple manner. The same is true with respect to the parameter Θ which allows us to model differences in the perception of environmental damages. Trivially, $\Theta = 1$ and $\omega = 1$ implies symmetric countries.

In Appendix VIII.2 we also derive the bargaining proposals for the quota regime and provide equilibrium emissions for both regimes and all parameter ranges. Also emissions in the Nash equilibrium, social optimum and minimax emission tuples are provided. In order to ensure interior solutions it is necessary to impose certain restrictions on the parameters which are summarized in assumption A_3 in (A11.5), Appendix VIII.2, which are assumed to hold in the remainder.

11.6 EQUILIBRIUM ANALYSIS

We are now prepared to conduct some interesting comparisons. In Subsection 11.6.1 we compare equilibrium emissions and welfare levels under the quota and tax regimes with those in the NE and in the social optimum.

Subsequently, in Sub-section 11.6.2 we compare emissions and welfare levels under the two bargaining regimes with each other. Section 11.7 considers the implications of strategic offers if countries can anticipate adjustment because of complete information.

11.6.1 Comparison of Emission and Welfare Levels under the Two Bargaining Regimes with the Nash Equilibrium and Social Optimum

The results obtained in Section 11.4 for the bargaining proposals together with the investigation of possible incentives to overfulfill an agreement in Section 11.5 allow us to derive some general results:

Proposition 11.3
Let the bargaining outcome be determined according to the LCD decision rule under partial information and allow countries to overfulfill but not to violate the terms of the agreement, then global emissions under the quota and tax regimes are always lower than in the Nash equilibrium, that is, $\Sigma e_i^k < \Sigma e_i^N$, $k \in \{Q, T\}$. Under the quota regime global emissions may be below socially optimal levels, that is, $\Sigma e_i^Q \gtreqless \Sigma e_i^S$, whereas under the tax regime global emissions are at least as high as in the social optimum, that is, $\Sigma e_i^T \geq \Sigma e_i^S$.

Proof: See Appendix VIII.3. QED

Proposition 11.3 implies that an agreement improves upon environmental quality compared to the status quo, but that, generally, aggregate emissions will be above those in the social optimum. The reason is simple: though the prospect of a joint environmental policy encourages both countries to agree on more ambitious abatement targets than in the NE, the institutional setting renders the bargaining solution suboptimal from a global point of view. Whereas the first institutional restriction *LCD decision rule* is binding under both regimes, the second institutional restriction *uniformity of the policy levels* is only binding under the quota regime. Hence, only in those specific cases where the proposals under the tax regime are equal, that is, $t_i = t_j$, are aggregate emissions under the tax regime socially optimal. Whereas identical proposals are a *sufficient condition* to obtain a socially optimal emission allocation in the two countries under the tax regime, it is only a necessary condition under the quota regime. Under a quota regime identical proposals imply only that global welfare is maximized *given* the constraint of a uniform solution ($r_i = r_j \Leftrightarrow r^A = r^*$, where r^* solves $\Sigma\pi_k$). However, as long as marginal abatement costs differ, such an

agreement is not socially optimal. Though $r^{A} \leq r^{*}$ is always true, $\Sigma e_k^N(1-r^*) < \Sigma e_k^S$ may hold, and therefore $\Sigma e_k^Q < \Sigma e_k^S$ is possible, as shown in Appendix VIII.3.

Proposition 11.3 has its analogy with respect to aggregate welfare. Since Proposition 11.4 is basically an immediate implication of Proposition 11.3, it is not commented on:

Proposition 11.4
In the uniform quota and uniform tax bargaining equilibrium aggregate welfare is always higher than in the Nash equilibrium but generally lower than in the social optimum. Under the quota regime equal proposals are a necessary condition to obtain socially optimal welfare levels whereas under the tax regime it is a sufficient condition.

Proof: See Appendix VIII.4. QED

Taken together, the results suggest that environmental quality and global welfare improve under both regimes compared to the NE, though global emissions are usually above and global welfare below the social optimum due to institutional restrictions.

From an individual government's perspective, however, what matters is not global gains to be realized in an agreement but the welfare implications for its 'own' country.

Proposition 11.5
Under a quota agreement each country gains compared to the Nash equilibrium, that is, $\pi_i^Q > \pi_i^N \; \forall \; i \in I$, regardless of which country determines the terms of the agreement. Under a tax agreement the non-bottleneck country always gains from an agreement, that is, $\pi_j^T > \pi_j^N$; the bottleneck country may gain or lose compared to the Nash equilibrium, that is, $\pi_i^T > \leq \pi_i^N$, though it receives at least an individual rational payoff, that is, $\pi_i^T \geq \pi_i^M$. If the non-bottleneck country j has an incentive to reduce emissions unilaterally below what is implied by the agreement on t^A, that is, $I_j(t^A) < 0$, then $\pi_i^T(e^{T,a}) < \pi_i^N(e^N)$ for the bottleneck country i.

Proof: See Appendix VIII.5. QED

Thus, even though countries may exhibit very asymmetric interests, both countries gain from a quota agreement in any case. This may not be true under a tax agreement. At first glance, the fact that the bottleneck country could lose under a tax regime may seem at odds with the rules of the game: that the bottleneck country determines the terms of the agreement.

However, at second glance it is evident that, because a uniform tax rate is not directly related to emissions in the NE, emission and welfare allocation might be very asymmetric under a tax regime. There might be no *uniform* tax rate for which $\pi_i^T \geq \pi_i^N$ is possible for the bottleneck country. Interestingly enough, this is even true if country *j* adjusts. In fact, whenever adjustment occurs, this implies $\pi_i^T \leq \pi_i^N$.

Of course, such possible asymmetric allocations of net benefits under the tax regime might be problematic with respect to the stability of an agreement. This issue is taken up in Chapters 12 and 14.

To summarize all information with respect to the functions in (11.8) compactly for subsequent chapters, we provide the following corollary:

Corollary 11.1

For the functions in (11.8), assumption A_3 in (VIII.4), Appendix VIII.2, partial information and the LCD decision rule, $\Theta = \omega = 1$ implies $r_1 = r_2 = r^A$ so that $\Sigma e_k^Q = \Sigma e_k^S$ and $\Sigma \pi_i^Q = \Sigma \pi_i^S$ hold. If $\Theta = \omega \neq 1$, then $r_1 = r_2 = r^A$ from which $\Sigma e_k^Q = \Sigma e_k^Q(r^*) > \Sigma e_k^S$ and $\Sigma \pi_k^Q = \Sigma \pi_k^Q(r^*) < \Sigma \pi_k^S$ follow. If $\Theta \neq \omega$, then $r_1 \neq r_2$, implying $\Sigma e_k^Q > \Sigma e_k^Q(r^*)$, Σe_k^S and $\Sigma \pi_k^Q < \Sigma \pi_k^Q(r^*) < \Sigma \pi_k^S$. $\Theta = \omega$ is a necessary and sufficient condition for a Pareto-efficient bargaining equilibrium under the quota regime.

Under the tax regime $\Theta = 1/\omega$ implies $t_1 = t_2 = r^A$ from which $\Sigma e_k^T = \Sigma e_k^S$ and $\Sigma \pi_k^T = \Sigma \pi_k^S$ follow. If $\Theta \neq 1/\omega$, then $t_1 \neq t_2$, implying $\Sigma e_k^T > \Sigma e_k^S$ and $\Sigma \pi_k^T < \Sigma \pi_k^S$. $\Theta = 1/\omega$ is a necessary and sufficient condition for a Pareto-efficient and socially optimal equilibrium under the tax regime.

For all parameter constellations $\pi_i^Q > \pi_i^N$ and $\pi_i^T > \pi_i^M \; \forall \; i \in I$.

Proof: Most parts of the proof are contained in the general proofs of Propositions 11.3–11.5 (see Appendices VIII.3, VIII.4 and VIII.5). The statement with respect to Pareto-efficiency follows from the auction market in Section 10.4, where it has been shown that at the equilibrium exchange rate (which implies in the present context equal proposals) the emission tuple is an element of the Pareto frontier. Since under the tax regime no adjustment takes place for $\Theta = 1/\omega$ (bargaining outcome equals tax equilibrium), Pareto-efficiency follows. QED

11.6.2 Comparison of Emission and Welfare Levels between the Quota and Tax Regimes

Unfortunately, it is not possible to conduct a comparison of equilibrium emissions and welfare between the two regimes at such a general level as in the previous section. However, with the help of the example in (11.8) it is easily checked that aggregate emissions and aggregate welfare in the quota

regime may be lower or higher than in the tax regime. Of course, whenever $\Sigma e_i^Q > \Sigma e_i^T$ holds, $\Sigma \pi_i^Q < \Sigma \pi_i^T$ must be true because then aggregate emissions under a tax regime come closer to those in the social optimum and abatement is conducted cost-efficiently. However, if $\Sigma e_i^Q < \Sigma e_i^T$ this general conclusion cannot be drawn any more because this may imply $\Sigma \pi_i^Q > \Sigma \pi_i^T$. In such cases, though reduction is conducted cost-efficiently under a tax regime (CEA condition (11.2) holds), aggregate emissions under a quota regime come closer to what is required from cost–benefit considerations (see CBA condition (9.24)). If $\Sigma e_i^Q << \Sigma e_i^T$, then the cost-efficiency effect (CEA effect) may be overcompensated by the cost–benefit effect (CBA effect) which renders global welfare higher under a quota regime than under a tax regime. The reason is that in these cases the bottleneck country finds the conditions under the quota regime more in line with its interests and therefore accepts higher abatement targets than under a tax regime.[15]

Proposition 11.6
Assuming that an agreement is reached according to the LCD decision rule, global welfare might be higher and global emissions lower in the quota than in the tax equilibrium.

Proof: The proof follows by example, such as using net benefit functions in (11.8) and Table VIII.1 in Appendix VIII.2. QED

To say a little more about the conditions when $\Sigma e_i^Q < \Sigma e_i^T$ or $\Sigma \Pi_i^Q > \Sigma \Pi_i^T$ hold, we consider two cases. Case 1 assumes for the example in (11.8) equal damage in both countries but allows for different benefit functions, that is, $\Theta = 1$ and $\omega \neq 1$. Case 2 considers equal benefits but different damages, that is, $\omega = 1$ and $\Theta \neq 1$.

For both examples we find $\Sigma e_i^Q < \Sigma e_i^T$ and $\Sigma \pi_i^Q > \Sigma \pi_i^T$ if $\gamma = b/c$ is large. The opposite holds for small values of γ.[16] That is, if abatement is relatively expensive compared to perceived environmental damage the quota regime performs better than the tax regime. Of course, the term 'relatively expensive' is unspecific and difficult to relate to real world problems. Therefore, conclusions can be only indicative and must be drawn with caution. With these restrictions in mind, however, the literature suggests that a problem where abatement costs are perceived to be rather high compared to possible damage costs is the greenhouse effect. In contrast, the depletion of the ozone layer could belong to the second category of low γ values because abatement costs are rather low compared to damage (Barrett 1991b, 1994b; Nordhaus 1993).

Taken together, the results suggest that in a second-best world with institutional restrictions a quota may perform better with respect to ecological

and welfare criteria than a tax regime. In these cases the bottleneck country finds the conditions under the quota regime more in line with its interests and therefore accepts higher abatement targets than under a tax regime. Of course, we have to remind ourselves that this result hinges on the assumption that bargaining equilibria are stable, which is still to be analyzed in Chapters 12 and 14.

11.7 STRATEGIC PROPOSALS

So far we have assumed that the bottleneck country does not consider the possible adjustment of the neighboring country when putting forward its proposal. As long as damage in the neighboring country is not known (assumption of partial information) the adjustment reaction cannot be computed. However, if we assume a game of complete information, then overfulfillment of a country might be anticipated and used strategically by its neighbor. Of course, a quota regime is immune to strategic offers because there is no proposal $r_i \geq 0$ for which $I_j(e^Q) < 0$ is true. Under a tax regime, however, a country performs:

$$\max\Big[\max_{t_i}[\beta_i(e_i(t_i)) - \phi_i(\Sigma e_k(t_i))],\ \max_{t_i}[\beta_i(e_i(t_i)) - \phi_i(e_i(t_i), e_j(e_i(t_i)))]\Big] \tag{11.10}$$

instead of (11.6). The second term in brackets states the optimization problem of a biased proposal. In this case the strategic mover behaves *de facto* as a *Stackelberg leader* and proposes a tax rate, t_i^{str}, which is lower than his/her 'true proposal', t_i, that is, $t_i^{str} < t_i$. The first term in the large square brackets is the familiar optimization problem of a non-biased proposal as known from (11.6). Thus, it is evident that it does not always pay to make a strategic proposal. Obviously, if a country already receives more without a strategic proposal than as a Stackelberg leader, a strategic offer would be irrational.[17] Moreover, since the adjustment is conducted according to a country's reaction function, this may imply an emission allocation which is less favorable to the strategic mover than the emission allocation under a tax regime which is based on the ratio of marginal benefits in both countries.

From Section 10.2 it is known that one can only solve for a Stackelberg equilibrium if either player 1 or 2 is in the lead; however, there is no equilibrium with two leaders. In the present context this implies that there is only a 'strategic equilibrium' if only one country uses information about possible overfulfillment strategically. Such an equilibrium has the following properties:

Proposition 11.7
A strategic proposal under a tax regime yields higher global emissions than without strategic considerations and compared to the stage game Nash equilibrium. Global welfare will be lower than without strategic considerations. The non-bottleneck country loses from the strategic proposal and its welfare is below that in the Nash equilibrium; however, it is above the minimax payoff. The bottleneck country will gain from the strategic bid and its welfare will be above that in the Nash equilibrium. Under a quota regime strategic proposals will never occur.

Proof: See Appendix VIII.6. QED

The result suggests that, in contrast to many other economic problems, complete information can have a negative effect in the present context. Complete information may put players in the position of anticipating possible overfulfillment and the resulting strategic offer increases emissions and reduces aggregate welfare; whereas the bargaining outcome under a tax regime is sensitive to strategic proposals, there will be no strategic moves under the quota regime.

11.8 SUMMARY

In a second-best world equilibrium emissions under a quota and a tax agreement were derived, considering possible incentives of countries to overfulfill treaty obligations. The bargaining setting reflected three institutional restrictions which represent typical features of most historical IEAs ratified so far:

1. uniform solutions;
2. no transfers; and
3. agreement on the lowest common denominator offer.

In such a setting it was shown that whereas a tax agreement is cost-efficient, this is generally not true for a quota regime. Nevertheless, the negotiation outcome under a quota may be superior to a tax regime in three respects:

1. Global emissions may be lower in the quota regime if the conditions for the bottleneck country are more favorable than under a tax regime, so that it agrees to a higher abatement target.
2. Global welfare may be higher in the quota regime if the effect described in 1 (cost–benefit effect) overcompensates the inherent inefficiency of the quota regime (cost-efficiency effect).

3. A quota regime is immune to strategic bids regardless of whether negotiators have complete or only partial information about their opponents' payoff functions. In contrast, in a tax regime proposals may be adjusted downward if negotiators have complete information.

The results may serve as a first indication why within many IEAs reduction targets are specified as uniform emission reduction quotas though more efficient instruments are available. In a next step the stability aspects of the quota and tax equilibrium have to be analyzed; this is done in Chapters 12 and 14.

NOTES

1. Uniform emission quotas are also part of the amendments signed in London (1990) and in Copenhagen (1992).
2. Also other IEAs which are not concerned with emission reductions often specify uniform duties for signatories. Examples include the 1972 London Convention on the Prevention of Marine Pollution by Dumping from Ships and Aircraft, the 1974 Helsinki Convention on the Protection of the Marine Environment of the Baltic Sea Area (HELCOM), and the 1973 Washington Convention on International Trade in Endangered Species of Wild Fauna and Flora. See also Barrett (1992d) and Hoel (1991, 1992a) on 'uniform regulations'.
3. The focal point as a device to select a particular equilibrium among a larger set has already been discussed in the context of the assurance game. See Section 3.4.
4. A uniform tax does not imply that taxes are increased *either* by the same absolute (ϵ) *or* by a relative (κ) amount compared to the Nash equilibrium. Let taxes which bring about a Nash equilibrium emission allocation in countries 1 and 2 be denoted by t_1^N and t_2^N ($\beta_i'(e_i^N) = t_i^N$), then, $t_1^N + \epsilon \neq t_2^N + \epsilon$ and $t_1^N(1+\kappa) \neq t_2^N(1+\kappa)$ will generally be true.
5. Since we do not use these concepts in the following we do not give a detailed description. Roughly speaking, the *Nash bargaining solution* allocates the gains from cooperation in proportion to the *threat points of players* (Nash 1950b). A frequently made assumption is that the threat points are the Nash equilibria or the minimax values of the one-shot game. See, for example, Binmore *et al.* (1986) and Owen (1982). The *Shapley value* distributes the gains from cooperation according to the *average marginal contribution* of a country to the overall welfare gain (see, for example, Moulin 1988; Schotter and Schwödiauer 1980). All concepts assume that the cooperating players maximize the aggregate welfare of the participants.
6. Bargaining models in the context of international pollution control are treated, for instance, in Barrett (1992a, d); Compte and Jehiel (1997); Eyckmans (1997); Kuhl (1987); Myerson (1997); Richer and Stranlund (1997); Rotillon and Tazdait (1996). General references on bargaining models are Binmore and Dasgupta (1987); Canning (1989); Osborne and Rubinstein (1990); Roth (1979, 1985); Rubinstein (1987); and Stähler (1998b).
7. To render the following analysis interesting, we assume an interior Nash equilibrium (which is unique by A_1) to represent the status quo so that there is an incentive for both countries to reduce emissions within an agreement. Since $e_i^N < e_i^{\max}\ \forall\ i \in \mathrm{I}$ some environmental policy must already be in place at the national level in the status quo.
8. This procedure seems to be suggestive since it turns out in Chapter 12 that, though the equilibrium set in a supergame framework may be substantially reduced by integrating

some 'real world restrictions' into the analysis, nevertheless in a global emission game a unique equilibrium can hardly be expected.

9. Within the Convention on Long-range Transboundary Air Pollution for the agreements on sulfur dioxide (SO_2), (Helsinki 1985, Oslo 1994), nitrogen oxide (NOx) (Sofia 1988) and VOC (volatile organic compounds) (Geneva 1991) emissions are gathered within EMEP (Environmental Monitoring and Evaluation Program) which was signed in 1984. In Article 4.1 the UN Framework Convention on Climate Change also calls on all parties to publish emission data which are accessible on the internet (http://www.unfccc.de). Of course, in some cases the reporting is incomplete (Benedick and Pronove 1992).
10. In a strict sense, we just assume that information on neighbors' damage costs are not used when deriving a proposal. We do not model incomplete information in a game theoretical sense, that is, putting priors on countries' expectations about their neighbors' damage costs. Since it will turn out that introducing the more restrictive assumption of 'complete information' may lead parties to put forward biased proposals, this possibility is also considered in Section 11.7.
11. This problem of insufficient information about damage costs has led in the national context to Baumol and Oates's standard-price approach as a pragmatic modification of a first-best Pigouvian tax (Baumol and Oates 1971). The problems associated with the measurement of damage are discussed in Endres and Holm-Müller (1998); Garrod and Willis (1999); Johansson (1990); Willis *et al.* (1999).
12. The possibility of $e_i^{\mathrm{I}} = e_i^{\max}$ is discussed in Endres and Finus (1999).
13. For an analogous assumption, see Marino (1988).
14. Thus, strictly speaking, in (11.5) and (11.6) the agreement procedure $k^{\mathrm{A}} = \min[k_i, k_j]$ should have already been incorporated for a complete description of the maximization problems. However, as Proposition 11.1 demonstrates, this would not change equilibrium behavior.
15. This result is confirmed by using other initial emission levels as the starting point and other functional forms of the net benefit functions. In particular, it also holds for the functions assumed in Chapter 14 on coalition formation.
16. A detailed derivation is available upon request.
17. It is also possible that *both* countries receive more than in a strategic equilibrium. For instance, if countries are symmetric and settle for a socially optimal tax, a strategic proposal would be to the disadvantage of both.

12. Infinite dynamic games with continuous strategy space

12.1 INTRODUCTION

In this chapter we analyze the global emission game characterized in Chapter 9 in the context of an infinite time horizon.[1] As argued in the introduction to Chapter 5, a supergame framework seems adequate for most IEAs due to the open-ended character of such agreements. In this chapter attention will be restricted to two countries only. Extensions to more than two countries are considered in Chapter 14.

Apart from characterizing the set of subgame-perfect equilibria (SPE), weakly and strongly renegotiation-proof equilibria (WRPE and SRPE) and strongly subgame-perfect equilibria (SSPE) in the global emission game, attention will be given to the stability of the quota and tax bargaining equilibria derived in Chapter 11 and the auction equilibrium derived in Chapter 10.[2] Thereby, we assume throughout this chapter that strategic proposals under the tax regime and during the auction process are absent if not explicitly mentioned.

The properties of subgame-perfect equilibria will only be discussed when a comparison to the other equilibrium concepts is immediately possible ($\delta_i \to 1 \; \forall \; i \in I$) due to the conceptual inferiority of this concept.

The structure of this chapter is similar to that of Chapter 7; that is, first discount factors close to 1 are assumed (Section 12.2) and then the analysis is extended to discount factors strictly smaller than 1 (Section 12.3). Particular attention will be paid to restricted punishment profiles due to various 'real world restrictions' (Sub-sections 12.2.4 and 12.3.4).

12.2 DISCOUNT FACTORS CLOSE TO 1

12.2.1 Subgame-perfect Equilibria

According to Folk Theorem IV (Theorem 5.4, Chapter 5) all payoff tuples which are *individually rational* can be backed as an SPE if discount factors of all players are close to 1. Thus, any check for stability is straightforward

since any payoff tuple has only to be measured against the benchmark *minimax payoff*. This allows us to draw rather general conclusions with respect to the stability of the quota and tax equilibrium, the auction equilibrium and the social optimum:

Proposition 12.1
Let the stage game of the global emission game be described by the general payoff functions (9.1), assumptions A_1 hold, and the strategy space be given by $E_i = [0, e_i^{\max}]$, $e_i^N < e_i^{\max}$ $\forall\, i \in I$, then the quota, tax and auction equilibrium (as derived in Chapters 10 and 11) can be sustained by subgame-perfect strategies provided discount factors of all players are sufficiently close to 1. Under the tax regime this holds regardless of whether adjustment is anticipated.

Proof: Follows from Propositions 10.4, 11.5, 11.7 and applying Theorem 5.4. QED

In contrast, payoffs in the social optimum may be very unevenly distributed if countries are asymmetric, and this may violate the individual rationality constraint (see Proposition 9.2). Hence, socially optimal emission tuples may even not be sustainable in an infinitely repeated game using SPE punishment strategies.

12.2.2 Weakly and Strongly Renegotiation-proof Equilibria

Preliminaries
As we know from Chapter 7, there is no simple benchmark to check for a WRPE or an SRPE. Only for particular payoff functions can the renegotiation-proof payoff space be determined. Therefore, the following analysis is based on the familiar payoff functions in Chapter 11, equation (11.8). Weakly renegotiation-proof conditions C_1 and C_2 in the *payoff space* are derived, proceeding as in Chapter 7. Additionally, these WRPE conditions are illustrated in the two-dimensional *strategy space*. It will be investigated whether the WRPE conditions are binding compared to the SPE conditions (as in the extended PD game V, Chapter 7) or not (as in the ordinary PD game, Chapter 7).

Derivation of the weakly renegotiation-proof conditions
The WRPE conditions (7.5) and (7.7) for discount factors close to 1 (but $\delta_i < 1$!) read in the context of the global emission game:

$$\pi_i^*(e_i, e_j) \geq \pi_i^C(e_i(e_j^i), e_j^i) \qquad (12.1)$$

$$\forall\, i \in N.$$

$$\pi_j^*(e_i, e_j) \leq \pi_j^R(e_i^i, e_j^i) \qquad (12.2)$$

Assume for a start that country 1 is the potential defector; then, using payoff functions (11.8), these conditions can be written as:

$$\pi_1^* \geq b(de_1(e_2^1) - \frac{1}{2}e_1(e_2^1)^2) - \frac{c}{2}(e_1(e_2^1) + e_2^1)^2 \tag{12.3}$$

$$\pi_2^* \leq b\omega(de_2^1 - \frac{1}{2}e_2^{1^2}) - \frac{c}{2}\Theta(e_1^1 + e_2^1)^2 \tag{12.4}$$

where $e_1(e_2^1)$ is country 1's best deviation according to its reaction function. For (11.8) $e_1(e_2^1) = (bd - ce_2^1)/(b + c)$ which may be substituted into (12.3). Then:

$$\pi_1^* \geq \frac{b(bd^2 - ce_2^{1^2} - 2cde_2^1}{2(b + c)}. \tag{12.5}$$

Since $\partial\pi_2^R/\partial e_1^1 < 0$, we set $e_1^1 = 0$ in (12.3) to have:

$$\pi_2^* \leq b\omega\left(de_2^1 - \frac{1}{2}e_2^{1^2}\right) - \frac{c}{2}\Theta e_2^{1^2}. \tag{12.6}$$

That is, (12.4) is 'most easily satisfied' for $e_1^1 = 0$ which allows us to determine the *outer boundaries* of the WRPE payoff space.[3] This assumption basically implies that country 1 accepts the harshest possible punishment in the game. Modification of this assumption will be treated in Subsection 12.2.4.

Solving (12.4) for e_2^1 gives:

$$e_2^1 \in \left\{\frac{db\omega - \sqrt{d^2b^2\omega^2 - 2\pi_2^*(b\omega + c\Theta)}}{b\omega + c\Theta}, \frac{db\omega + \sqrt{d^2b^2\omega^2 - 2\pi_2^*(b\omega + c\Theta)}}{b\omega + c\Theta}\right\}. \tag{12.7}$$

First note that the root is always positive if $\pi_2^* \leq \pi_2^U$ which obviously holds by the definition of the upper bound of the payoff space.[4] Second, note that e_2^1 should be chosen as big as possible because $\partial\pi_1^C/\partial e_2^1 < 0$. However, e_2^1 is restricted by the definition of the strategy space to $e_2^1 \in [0, d]$ (see A_3 in (VIII.4) in Appendix VIII.2). Whether the RHS expression in (12.7) or d is larger depends on the specific parameter values. Thus, we take e_2^1 to be:

$$e_2^1 = \min\left\{d, \frac{db\omega + \sqrt{d^2b^2\omega^2 - 2\pi_2^*(b\omega - c\Theta)}}{b\omega + c\Theta}\right\}. \tag{12.8}$$

Assume, first, that the second term in (12.8) is larger than the first term and hence $e_2^1 = d$. Substituting this into (12.5) gives the first condition for stability, L_1:

$$L_1 := \pi_1^* \geq \frac{bd^2(b-3c)}{2(b+c)} \Leftrightarrow \pi_1^* \geq \pi_1^M. \tag{12.9}$$

Of course, L_1 is nothing other than requiring payoffs to be above the minimax payoff which is the condition for subgame-perfection.[5]

Now consider the second possibility, that the second term in (12.8) is smaller than d. Inserting the second expression in (12.8) into (12.5) gives:

$$C_1 := \pi_1^* \geq \frac{b}{2(b+c)} \left(bd^2 - \frac{c(bd\omega + \sqrt{A})^2}{(b\omega + c\Theta)^2} - \frac{2cd(bd\omega + \sqrt{A})}{(b\omega + c\Theta)} \right)$$

$$A := b^2d^2\omega^2 - 2\pi_2^*(b\omega - c\Theta). \tag{12.10}$$

Obviously, it is not sufficient to show that country 2 can credibly punish country 1; it is also necessary to formulate the conditions in the opposite case. Since the computations are basically the same as outlined above, only the final result is presented:

$$L_2 := \pi_2^* \geq \frac{b\omega d^2(b\omega - 3c\Theta)}{2(b\omega + c\Theta)} \Leftrightarrow \pi_2^* > \pi_2^M \tag{12.11}$$

$$C_2 := \pi_2^* \geq \frac{b\omega}{2(b\omega + c\Theta)} \left(b\omega d^2 - \frac{c\Theta(bd + \sqrt{B})^2}{(b+c)^2} - \frac{2cd\Theta(bd\omega + \sqrt{B})}{(b+c)} \right)$$

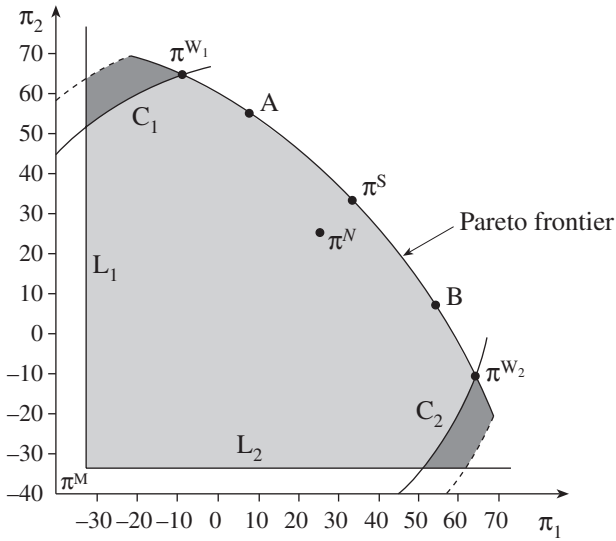
$$B := b^2d^2 - 2\pi_1^*(b-c) \tag{12.12}$$

Thus taken together, only if L_1, L_2, C_1 and C_2 hold simultaneously is a payoff tuple $\pi^* = (\pi_1^*, \pi_2^*)$ a WRPE.

Illustration of the weakly and strongly renegotiation-proof payoff space

Figure 12.1 depicts all four WRPE conditions. Whereas the SPE payoff space (for $\delta_i \to 1 \ \forall \ i \in I$), $\overline{\Pi}^{SPE}$, comprises the entire area lying within the boundaries L_1 and L_2 and below the Pareto frontier (light and dark shaded areas), the WRPE payoff space, $\overline{\Pi}^{WRPE}$, only includes the area below the C_1 and above the C_2 curve (light shaded area). The reason is the following. Payoff tuples lying in the dark shaded regions imply for one country a relatively high payoff and for the other country a relatively low payoff. Consequently, the low payoff country has a high free-rider incentive. To deter such a country from deviating requires a severe punishment in the form of high emissions. Since the punisher already receives a high payoff in the cooperative phase, it is difficult (or impossible in the dark shaded areas) to secure at least his/her cooperative payoff during the punishment.

Thus, for the example depicted in Figure 12.1 the set of WRPE payoffs



Note: Payoff functions (11.8) apply; $b = 2$, $c = 1$, $d = 10$ and $\Theta = \omega = 1$ are assumed.

Figure 12.1 Subgame-perfect, weakly and strongly renegotiation-proof payoff space

is a true subset of the set of SPE payoffs, that is, $\overline{\Pi}^{\text{WRPE}} \supset \overline{\Pi}^{\text{SPE}}$. However, it is easily checked that for other parameter values this must not be true. That is, the C_1 curve may run above and the C_2 curve below the boundaries defined by L_1, L_2 and the Pareto frontier. For instance, increasing the parameter value of b from $b = 2$ to, say, $b = 5$ and keeping constant the other parameter values assumed in Figure 12.1 would be such a case.

All payoff tuples lying on the Pareto frontier between π^{W_1} and π^{W_2} in Figure 12.1 are SRPE by Theorem 7.8. Depending on the parameter values, the set of SRPE payoffs may or may not be a *true* subset of the Pareto frontier, that is, $\overline{\Pi}^{\text{SRPE}} \supset \text{P}(\Pi^{\text{IR}})$ or $\overline{\Pi}^{\text{SRPE}} \supseteq \text{P}(\Pi^{\text{IR}})$.

Type A and B games

Though for the emission game conditions C_1 and C_2 are sometimes more restrictive than the minimax conditions L_1 and L_2, they are *never* restrictive with respect to the *whole* domain of individually rational payoff space. We call this a type B game (see Figure 12.2b). In contrast, in type A games the C_1 and C_2 curves are completely contained in the individually rational payoff space (irrespective of the parameter values, see Figure 12.2a). The distinguishing feature between types A and B is that in a type A game a player minimaxing his/her opponent can only receive less or at best his/her

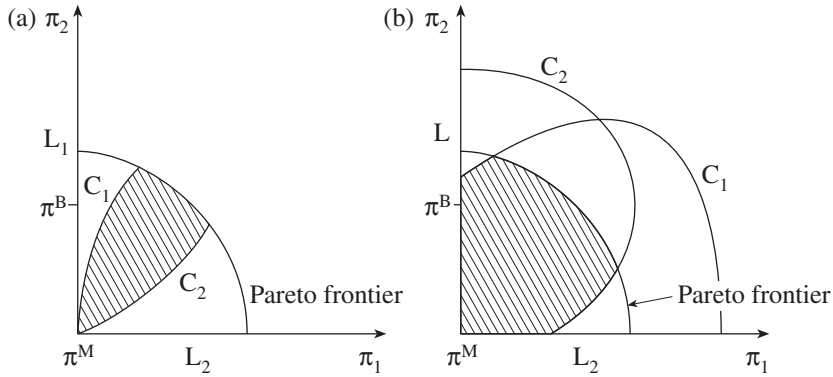


Figure 12.2 Weakly renegotiation-proof payoff space of type A and B games

own minimax payoff ($\pi_j^{M(i)} \leq \pi_j^{M(j)}$), whereas in a type B game s/he may receive more ($\pi_j^{M(i)} \leq, > \pi_j^{M(j)}$). Thus, in a game of type A punishment is always costly, but in a game of type B this depends on the specific circumstances (in the emission game on the particular parameter values; see also the discussion of Folk Theorem III, Section 5.2).

Proposition 12.1
In an infinitely repeated two-player game of type A, that is, in a game in which the renegotiation-proof conditions C_1 and C_2 are more restrictive than the minimax conditions with respect to the entire individually rational payoff space, minimaxing a fellow player implies that the punisher will receive less or at best his/her own minimax payoff.

Proof: Assume a type A game and a payoff tuple π^B in the payoff space for which $\pi_1^B = \pi_1^M$ and $\pi_2^B > \pi_2^M$ is true (see Figure 12.2a). In order for such a payoff tuple to be renegotiation-proof it must satisfy (12.1) and (12.2) $\forall\ i \in I$. Since $\pi_1^C \geq \pi_1^M$ must hold by individual rationality, $\pi_1^C = \pi_1^B = \pi_1^M$ follows. This implies that player 1 *must* be minimaxed during his/her punishment (even if s/he shows repentance and accepts the punishment). In order to satisfy condition (12.2), however, player 2 must receive at least his/her cooperative payoff, that is, $\pi_2^R \geq \pi_2^B$. If, however, the punishment of player 1 implies $\pi_2^R \leq \pi_2^M < \pi_2^B$ (by assumption of a type A game), payoff tuple π^B cannot be renegotiation-proof since condition (12.2) is violated. Since this holds for any point along line L_1 and, by symmetry, also along line L_2, *no* point along both lines can be

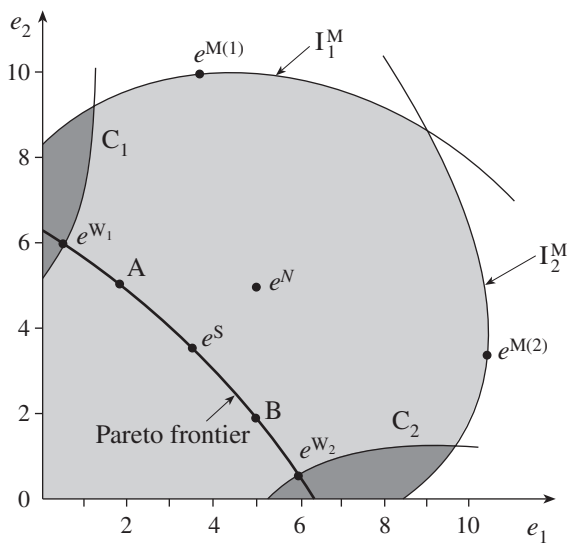
renegotiation-proof. In other words, C_1 and C_2 must lie completely within the individually rational payoff space. QED

Examples of type A games include the extended PD game V in Matrix 7.1 (see Figure 7.1). Other instances are Cournot and Bertrand duopoly, where the firm which is up to minimax its opponent must choose such a high (low) output (price) that its *own* payoff shrinks to zero, the minimax payoffs of these games.

In contrast, in a type B game punishment is not so costly. For instance, for the payoff functions (11.8) of the global emission game we have $e_1^l = 0$ and $e_2^l = d\ \pi_2^R = (d^2(b\omega - c\Theta))/2 \geq \pi_2^M$ if $\gamma \geq \Theta/3\omega$ which holds by assumptions A_3 in (VIII.4), Appendix VIII.2.

Illustration of the weakly and strongly renegotiation-proof emission space

Sometimes it is convenient to display conditions C_1, C_2, L_1 and L_2 in the strategy space instead of the payoff space. Conditions L_1 and L_2 then become the indifference curves I_1^M and I_2^M which are well-known from Chapter 9. To derive conditions C_1 and C_2 in the strategy space, one has to substitute the payoff functions of (11.8) for π_1^* and π_2^* in (12.10) and (12.12). Then C_1 and C_2 become implicit functions in the e_1–e_2 space. These are plotted in Figure 12.3, which assumes the same parameter values as in Figure 12.1.



Note: Payoff functions (11.8) apply; $b = 2$, $c = 1$, $d = 10$ and $\Theta = \omega = 1$ are assumed.

Figure 12.3 Subgame-perfect, weakly and strongly renegotiation-proof emission space

Again, the dark shaded regions comprise subgame-perfect emission tuples which are *not* weakly renegotiation-proof. Only emission tuples in the light shaded regions are WRRE. Payoff tuples π^{W_1} and π^{W_2} of Figure 12.1 have their analogy in the emission space which are the emission tuples denoted by e^{W_1} and e^{W_2} respectively. Thus, emission tuples along the segment between e^{W_1} and e^{W_2} on the Pareto frontier are SRPE.

Stability of the social optimum, the quota, tax and auction equilibrium

The example displayed in Figures 12.1 and 12.3 assumes symmetric parameter values (that is, $\Theta = \omega = 1$), implying that the quota and the tax bargaining equilibrium as well as the auction equilibrium coincide with the social optimum, that is, $\pi^S = \pi^Q = \pi^T = \pi^A$ and $e^S = e^Q = e^T = e^A$. Hence, trivially, for this *symmetric case* all payoff tuples are weakly and strongly renegotiation-proof. For the more general case, considering also the possibility of asymmetric payoff, we find:

Proposition 12.2

Let the stage game of the infinitely repeated emission game be described by payoff functions (11.8), let assumptions A_3 in (VIII.4) hold and let $\delta_i \to 1 \; \forall \; i \in I$, then the quota and tax equilibrium are weakly renegotiation-proof. Under the quota regime $\Theta = \omega$ and under the tax regime $\Theta = 1/\omega$ (implying equal proposals) are necessary and sufficient conditions that the bargaining equilibria are strongly renegotiation-proof. The auction equilibrium, where the exchange rate is defined in terms of emission reductions from the stage game Nash equilibrium, that is, $e_i^I = e_i^N \; \forall \; i \in I$, is always a strongly renegotiation-proof equilibrium.

Proof: The first part of Proposition 12.1 is proved by computing net benefits in the quota and tax equilibrium for all possible parameter constellations using emissions from Table VIII.1, Appendix VIII.2. For each parameter range it has to be checked whether C_1, C_2, L_1 and L_2 derived above hold, using assumptions A_3. Since the proof requires a substantial amount of computation, it is not reproduced here; however, it may be obtained upon request.

The second part of Proposition 12.2 simply follows from the fact that the quota and tax equilibria must lie on the Pareto frontier to qualify as potential SRPE candidates. This is the case for $\Theta = \omega$ and $\Theta = 1/\omega$ respectively (see Corollary 11.1). Since the above-mentioned proof holds for all parameter constellations, it therefore also covers $\Theta = \omega$ and $\Theta = 1/\omega$.

The third part of the statement is an implication of Proposition 12.4 below, which claims that the auction equilibrium is a strongly

subgame-perfect equilibrium if initial emissions are those in the Nash equilibrium (NE) and $\overline{\Pi}^{\text{SRPE}} \supseteq \overline{\Pi}^{\text{SSPE}}$ by definition. QED

The statement with respect to the auction equilibrium cannot be extended to cover the general case of $e_i^{\text{I}} \geq e_i^N$ since an explicit derivation of π_i^{A} is only possible assuming specific parameter values, but not at a general level (that is, for payoff function (11.8)). Though $\pi_i^{\text{A}} \geq \pi_i^{\text{M}}\ \forall\ i \in \text{I}$ is known from Proposition 10.4 and hence conditions L_1 and L_2 are satisfied, information about whether conditions C_1 and C_2 hold as well cannot be obtained at a general level. In contrast, as the proof of Proposition 12.4 shows, C_1 and C_2 can be satisfied for $\pi_i^* \geq \pi_i^N\ \forall\ i \in \text{I}$ *and* strongly subgame-perfect punishments. Therefore, since for $e^{\text{I}} = e^N\ \ \pi_i^{\text{A}} \geq \pi_i^N\ \ \forall\ \ i \in \text{I}$ by Proposition 10.4 and since it is known that the auction equilibrium is located on the Pareto frontier, the auction equilibrium is an SSPE and consequently an SRPE as well.

Since the social optimum may fail to be an SPE, it is clear that it cannot always be a WRPE. Of course, whenever it is a WRPE it is automatically an SRPE as well since the socially optimal payoff tuple is always an element on the Pareto frontier. From the previous discussion it should be evident that the more symmetric payoffs are, the higher is the probability that the social optimum is an SRPE.

12.2.3 Strongly Subgame-perfect Equilibria

From Chapters 6 and 7 it is known that a distinguishing feature of strongly subgame-perfect equilibria (SSPE) are Pareto-efficient punishments (apart from Pareto-efficient strategies during the cooperative phase). Hence, it has to be clarified what this means for the derivation of conditions C_1 and C_2 above. Basically, there are an infinite number of emission tuples which are Pareto-efficient. However, from Sub-section 12.2.2 it is known that to determine the outer boundaries of the payoff space, one should choose the punishment emission level as high as possible (subject of course to inequality (12.2)) and the repentance emission level as small as possible. Assuming country 1 to be the potential defector, this implies, as previously, $e_1^{\text{I}} = 0$ and an efficient punishment e_2^{I} follows immediately from $\max \pi_2(0, e_2^{\text{I}})$ which delivers the maximax payoff to country 2, π_2^{U}. Hence, $\pi_2^{\text{U}} = \pi_2^{\text{R(I)}}$ where the subscript in brackets indicates the punished player.[6] Then, inserting e_2^{I} into the best reply of country 1, $e_1(e_2^{\text{I}})$, allows us to derive π_1^{C}. For the functions in (11.8) we have:[7]

$$\pi_2^{\text{R(1)}} = \frac{b^2 d^2 \omega^2}{2(b\omega + c\Theta)},\ \pi_1^{\text{C}} = \frac{b^2 d^2 (b^2\omega^2 + 2bc\omega\Theta - 3bc\omega^2 - 2c^2\omega\Theta + c^2\Theta^2)}{2(b+c)(b\omega + c\Theta)^2}. \tag{12.13}$$

A similar procedure delivers:

$$\pi_1^{R(2)} = \frac{b^2 d^2}{(b+c)},\ \pi_2^C = \frac{b^2 d^2 \omega (b^2\omega + 2bc\omega - 3bc\Theta - 2c^2\Theta + c^2\omega)}{2(b+c)^2(b\omega + c\Theta)}. \tag{12.14}$$

Thus, we have determined all components of the inequality system (12.1) and (12.2) and may summarize this as follows:

$$\text{SSPE: } S_1\text{: } \pi_1^C \le \pi_1^* \le \pi_1^{R(2)} = \pi_1^U, S_2\text{: } \pi_2^C \le \pi_2^* \le \pi_2^{R(1)} = \pi_2^U \text{ and } S_3\text{: } \pi_i^* \in P(\Pi^{IR}) \tag{12.15}$$

where the respective payoffs are those given in (12.13) and (12.14) and where condition S_3 ensures that the payoff tuple in the cooperative phase is Pareto-efficient. That is, all payoff tuples which *simultaneously* satisfy conditions S_1 and S_2 are SSPE. Of course, in the present game the upper bound is not a binding constraint, so one may write only S_1: $\pi_1^C \le \pi_1^*$ and S_2: $\pi_2^C \le \pi_2^*$ for short.

In Figures 12.1 and 12.3 all points between points A and B satisfy the SSPE conditions. From the graphs it is evident that in the global emission game the set of SSPE is a *true* subset of the SRPE. Recall that for the examples in Chapter 7 this was different, namely the set of SSPE payoff tuples coincided with the SRPE payoff set. This finding for the emission game holds not only for the specific parameter values underlying Figures 12.1 and 12.3 but is in general true for the functions in (11.8).

Proposition 12.3
In the infinitely repeated global emission game where the stage game is defined by the functions in (11.8), and conditions A_3 in (VIII.4), Appendix VIII.2, hold, $\overline{\Pi}^{SRPE} \supset \overline{\Pi}^{SSPE}$.

Proof: The proof is obvious. Comparing the punishment emission levels, e_2^1, in (12.8) with e_2^1 as given in note 7 we find that the latter is smaller than the former. Recalling that $\partial \pi_1^C / \partial e_2^1 < 0$ and that π_1^C is country 1's lower bound for stable equilibria, then the claim is proved from country 1's perspective. A similar procedure would show that the same holds from country 2's perspective. QED

Note also the following interesting properties:

Proposition 12.4
Let the stage game payoff function of the infinitely repeated emission game be given by (11.8), assumptions A_3 in (VIII.4) hold and let $\delta_i \to 1$

$\forall\, i \in \mathrm{I}$, then equal proposals are a necessary and sufficient condition that the bargaining equilibria under the quota and tax regimes are strongly subgame-perfect equilibria. The auction equilibrium, where the exchange rate is based on emission reductions from the stage game Nash equilibrium, is always a strongly subgame-perfect equilibrium.

Proof: The first part of the statement is proved by computing net benefits under the quota and tax regimes, using emission levels of Table VIII.1, Appendix VIII.2. Then it is shown that $\pi_i^{\mathrm{U}} > \pi_i^{\mathrm{Q}} > \pi_i^{\mathrm{C}}$ and $\pi_i^{\mathrm{U}} > \pi_i^{\mathrm{T}} > \pi_i^{\mathrm{C}}\ \forall\, i \in \mathrm{I}\{1, 2\}$ holds for $\Theta = \omega$ (quota regime) and $\Theta = 1/\omega$ (tax regime) which are the conditions for Pareto efficiency (and which imply equal proposals by both countries; see corollary 11.1) by using A_3 in (VIII.4), Appendix VIII.2, and where π_i^{C} and π_i^{U} are those bounds given in (12.13) and (12.14).

The second part of the statement is proved by noting that for the payoff functions in (11.8) $\pi_1^{\mathrm{C}} \leq \pi_1^N$ and $\pi_2^{\mathrm{C}} \leq \pi_2^N$ hold for all parameter values and that the auction equilibrium constitutes a Pareto optimum for which $\pi_i^{\mathrm{A}} \geq \pi_i^N\ \forall\, i \in \mathrm{I}$ is true provided $e^{\mathrm{I}} = e^N$ holds by Proposition 10.4. QED

Thus, Proposition 12.4 strengthens the second part of Proposition 12.3, namely that equal proposals under the quota and tax regimes are not only a sufficient condition for an SRPE but also for an SSPE. Also the auction equilibrium is not only an SRPE but at the same time an SSPE as well. However, since equal proposals under quota and tax regimes are the exception rather than the rule, the quota and tax equilibria are generally only a WRPE and not an SRPE/SSPE.[8]

12.2.4 Restrictions of the Punishment Space

From the previous sub-section it became evident that requiring punishments to be efficient reduces the payoff space of stable equilibria. More generally, any restriction of the punishment options will reduce the possibility of backing a payoff tuple by credible strategies. This is particularly true for asymmetric payoff tuples. In reality one might expect that agents would be less concerned about efficient punishments, though other restrictions may apply.

Restrictions on the side of the punisher

For instance, the *upper bound of the punishment* may be restricted. Reasons for this could be that once countries have reduced emissions it is difficult to increase them beyond a certain threshold. Abatement technology may have

been implemented during the cooperative phase which is associated with sunk costs, so that punishment is limited.

For example, restricting the punishment to $e_j^{\max} = e_j^N$, then $\pi_i^M = \pi_i^N$ (see Section 9.5) and hence $\overline{\Pi}^{SPE} = \{\pi_i^* \in \overline{\Pi}^{SPE} \mid \pi_i^* \geq \pi_i^N \; \forall \; i \in I\}$. Similarly for a WRPE a *necessary* condition is $\pi_i^*(e_i, e_j) \geq \pi_i^C(e_i(e_j^i), e_j^i) \geq \pi_i^C(e_i(e_j^N), e_j^N) = \pi_i^C(e_i^N, e_j^N) = \pi_i^N$ if $e_j^{\max} = e_j^N$. That is, only emission tuples within the Nash lens qualify as potential WRPE. This requires a relatively symmetric distribution of abatement burdens. Apart from the social optimum, we know from Chapter 11 that the tax equilibrium may also not satisfy this condition, that is, $\pi_i^T < \pi_i^N$ may hold. In contrast, the quota equilibrium always takes this hurdle and therefore has a higher chance of being realized within an IEA. More concretely:

Proposition 12.5
Restricting the strategy space of the stage game defined by the payoff functions in (11.8) and assumptions A_3 in (VIII.4) to $E_i = [0, e_i^N] \; \forall \; i \in I$, the quota equilibrium can always be sustained by weakly renegotiation-proof strategies in an infinitely repeated game provided $\delta_i \to 1 \; \forall \; i \in I$, whereas the tax equilibrium and the social optimum may fail to be stable. If post-adjustment takes place after countries have agreed on a tax rate according to the LCD decision rule, the tax equilibrium is not stable, regardless of whether a country makes a strategic proposal.

Proof: The first part of the proposition with respect to the quota regime is proved in the spirit of the proof of Proposition 12.2. The second part of the proposition follows from Propositions 9.2, 11.5 and 11.7 and the necessary condition $\pi_i^* \geq \pi_i^N$ as noted above. QED[9]

Restricting the upper bound of punishment further makes it increasingly difficult to back agreements that call for substantial emission reductions compared to the NE. In particular, viewing the negotiation process of an agreement and subsequent modifications as a dynamic process (for example, in the spirit of sequential strategic matching; see Section 10.6), one can imagine that governments would gradually agree to increase their abatement efforts over time. Thus, aggregate and country-specific emissions would steadily decline along the emission path.

Now, consider again the possibility that emissions can only be increased by some percentages compared to the cooperative phase; then severe punishments would become increasingly difficult along the emission path. Since, as a tendency, lower (cooperative) emission levels imply a higher free-rider incentive to the parties involved, an abatement level may be reached above which no further reductions are possible. Put differently, one may

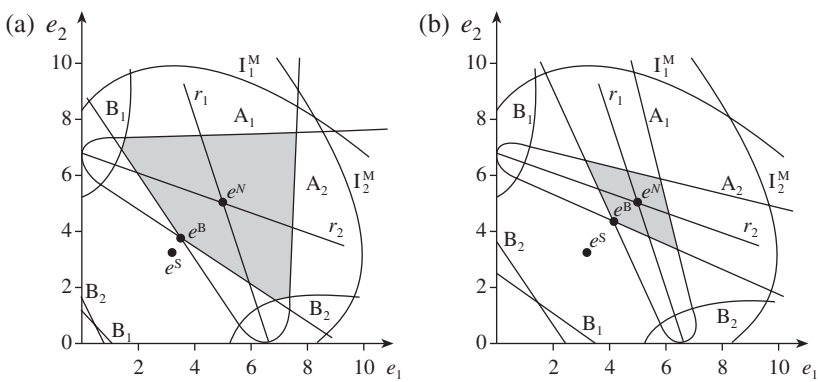
view this adjustment process as a path leading to a *steady-state WRPE*. Of course, strictly speaking, this (gradual adjustment) interpretation is no longer covered by our simple model, but the features described may very well apply in reality.[10]

Another reason why punishment might be restricted is *international law*. According to international law any misconduct by a country should only be punished in relation to the severity of the deviation. Though *relative proportional punishment*, or *reciprocal punishment* as it might be called, is a vague term and difficult to operationalize, it implies (provided governments accept this rule of conduct) that small deviations must be punished less severely than greater deviations.[11] This requires the formulation of sophisticated punishment profiles which are not simple in the sense of Abreu any longer.

In (12.16) an attempt has been made to operationalize the idea of relative punishment (see rule 1 in (12.16)) and the steady-state argument (see rule 2 in (12.16)) presented above in a simple manner, which is illustrated with the help of Figure 12.4. For the punishment the following relation has been assumed:

$$(1)\ e_j^i - e_j \leq |\chi(e_i(e_j) - e_i)| \Leftrightarrow e_j^i \leq |\chi\,(e_i(e_j) - e_i)| + e_j \ \text{ and } (2)\ e_j^i \leq (1+\xi)e_j. \tag{12.16}$$

Rule 1 The punisher j is allowed to increase his/her emissions proportionally to the deviation by player i where $\chi \geq 0$ is a parameter. Taking into account that, for emission tuples exceeding emissions in the NE, countries have an incentive to reduce emissions below e_i, and hence $e_i(e_j) - e_i < 0$, the



Note: Payoff functions (11.8) apply; $b = 2$, $c = 1$, $d = 10$, $\Theta = \omega = 1$ are assumed. The left graph assumes $\chi = 1$ and $\xi = 0.2$, the right graph $\chi = 0.8$ and $\xi = 0.05$.

Figure 12.4 Weakly renegotiation-proof emission space for reciprocal and restricted punishments

absolute deviation in the formula above is considered.[12] The higher χ, the easier it is to punish a country by choosing a high emission level. $\chi = 1$ implies that the punisher is allowed to increase its emission by the *same* amount as the deviation and therefore is compatible with the notion of reciprocal punishment. The weak inequality sign reflects the expectation that if it is in the interest of the punisher *not* to use the harshest permissible punishment (there is slack of enforcement power), s/he may choose a weaker punishment.

An interesting property of the punishment rule in (1) is apparent from the alternative formulation on the RHS of the arrow in (12.16), indicating that the punishment level, e_j^i, is a function of the cooperative phase emission level, e_j. Thus parts of the idea which have led to rule (2) are already reflected in rule (1).

Rule 2 The second restriction reflects the assumption that the punishment cannot exceed a certain level of the cooperative emission level where $\xi \geq 0$ is a parameter. For instance, $\xi = 0.1$ implies that emissions during the punishment cannot be increased by more than 10 percent.

General remarks In contrast to Sub-section 12.2.2 it is not possible to solve inequality (12.2) for the punishment emission level e_j^i and to substitute this in the inequality (12.1) of country i which gets punished to derive conditions C_i and L_i. Now inequalities (12.1) and (12.2) have to be expressed separately. Due to the inequality sign in (12.16), one cannot plot these inequalities only for a particular χ and ξ but must take into account that these values *only* define the *highest* permissible punishment; however, lower punishments are always possible. Thus, to satisfy condition (1) in (12.16) one searches for the set of emission tuples which satisfy (12.1) and (12.2) for any $e_j^i \in [0, \chi(e_i(e_j) - e_i)]$. By the same token, to satisfy condition (2) the envelope of the restrictions (12.1) and (12.2) for any $e_j^i \in [0, (1+\xi) \cdot e_j]$ has to be determined. Technically, this requires an algorithm searching sequentially for all stable emission tuples in the above-mentioned domain of the punishment space.

Results Apart from the same (and symmetric) parameters of the previous figures ($b = 2$, $c = 1$, $d = 10$, $\Theta = \omega = 1$), $\chi = 1$ and $\xi = 0.2$ have been assumed in Figure 12.4(a) and $\chi = 0.8$ and $\xi = 0.05$ in Figure 12.4(b). Condition (2) is reflected in Figure 12.4 by curve A_1 for country 1 and by curve A_2 for country 2 where, as pointed out above, A_1 is the envelope of inequalities (12.1) and (12.2) if country 1 is the potential defector and A_2 if this is country 2. Hence, emission tuples in the interior of A_1 and A_2 satisfy condition (2) in (12.16).

From the figures it is evident that only emissions tuples which are in the

vicinity of the reaction functions, r_1 and r_2, can be sustained by WRPE strategies. The more distant emission tuples are from the reaction function, the higher is the free-rider incentive, which is difficult to neutralize if the upper bound of the punishment is restricted. In particular, for emission tuples which imply relatively small emissions, e_j, the ceiling $e_j^i \leq (1+\xi)\cdot e_j$ becomes a limiting factor. In the examples e^B is that emission tuple with the lowest aggregate emissions which is still a WRPE. Obviously, in both cases $\Sigma e_k^B > \Sigma e_k^S$, so that e^S would not be stable.

Condition (1) is reflected by curves B_1 and B_2. All emission tuples above the B_1 and B_2 lines in the lower left-hand corner of the graph and all emission tuples to the right of the B_1 and to the left of the B_2 curves in the other two corners of the graph satisfy inequalities (12.1) and (12.2) subject to condition (1). Obviously, reciprocal punishment does not allow the backing of asymmetric emission tuples and emission tuples implying low emissions.

Thus, taken together, curves B_1 and B_2 further restrict the area enclosed by A_1 and A_2. The shaded areas remain as the WRPE emission space subject to the two restrictions in (12.16).

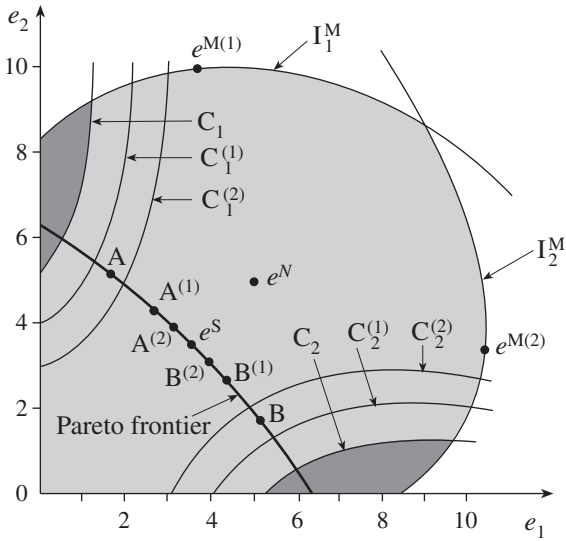
Since in Figure 12.4(b) more restrictive assumptions have been made regarding the punishment options compared to Figure 12.4(a), the WRPE emission tuple with the lowest aggregate emissions, e^B, moves closer to the NE, e^N, and the entire WRPE emission space shrinks.

Though we have not drawn the Pareto frontier in Figure 12.4 in order to avoid confusion, it is easily checked that in Figures 12.4(a) and (b) the set of SRPE and SSPE is empty.

Restrictions on the side of the punished player

Another important restriction which may apply in reality concerns the lower bound of the repentance emission level. When deriving the WRPE conditions we assumed $e_i^i = 0$, that is, the country which gets punished accepts the harshest possible punishment in order to resume cooperation. Though from a technical point of view this assumption allows us to determine the outer boundaries of the WRPE space with respect to the entire domain of the emission space, that is, $e_i \in [0, d]$, in reality a country may not be able to reduce its emissions so drastically. Though a country may have good intentions (and an incentive) to show repentance, it may be expected that at a given time and for a given abatement technology only modest emission reductions below the cooperative level are technically feasible. In particular, it should be expected that the lower the cooperative emission level is already, the more difficult it gets to show repentance by reducing emissions further. Similar to the argument presented in the context of the upper bound of the punishment level, a steady-state equilibrium may be reached where no further reductions can be stabilized.

In Figure 12.5 the effect of such a restriction on the lower bound of the



Note: Payoff functions (11.8) apply; $b = 2$, $c = 1$, $d = 10$ and $\Theta = \omega = 1$ are assumed.

Figure 12.5 Weakly renegotiation-proof emission space for restricted repentance strategies

repentance emission level is illustrated. To keep things simple, the lower bound has been fixed, independent of the cooperative emission level. C_1 and C_2 are the WRPE conditions without restriction, that is, $e_i \in [0, d]$, $C_1^{(1)}$ and $C_2^{(1)}$ assume $e_i \in [1, d]$ and $C_1^{(2)}$ and $C_2^{(2)}$ are drawn based on $e_i \in [2, d]$. From Figure 12.5 it is evident that the WRPE emission space decreases with an increase of the lower bound. In particular, asymmetric emission tuples, which also imply an asymmetric distribution of payoffs, are increasingly difficult to sustain as a WRPE.[13] The higher the lower bound of the repentance emission level, the more difficult it gets for the punisher to receive at least his/her cooperative payoff during the punishment (inequality (12.2)).

From Figure 12.5 it is also evident that because the C_1 and C_2 curves move inward with an increase of the lower bound of the repentance emission level, the set of SRPE (all points between A and B) also becomes smaller. The same holds true for SSPE where points A and B move inward with the new boundaries $A^{(1)}$ and $B^{(1)}$ if $e_i \in [1, d]$ and $A^{(2)}$ and $B^{(2)}$ if $e_i \in [2, d]$.

Summing up, in reality many restrictions on the strategy space may apply which render the solution space in which stable agreements must lie much smaller than 'pure' theory would suggest. As a tendency, neither asymmetric solutions nor solutions which call for a substantial emission reduction compared to the NE are then sustainable, even in a supergame framework.

12.3 DISCOUNT FACTORS SMALLER THAN 1

12.3.1 Derivation of the Minimum Discount Factor Requirement for Weakly and Strongly Renegotiation-proof Equilibria

From Chapter 7 it is known that for discount factors smaller than 1 three inequalities must be satisfied simultaneously, which are reproduced in the context of the global emission game:

$$\pi_i^{\mathrm{P}}(e_i^i, e_j^i, e_i, e_j) \geq \pi_i^{\mathrm{C}}(e_i(e_j^i), e_j^i) \tag{12.17}$$

$$\pi_i^*(e_i, e_j) \geq (1-\delta_i)\pi_i^{\mathrm{D}}(e_i(e_j), e_j) + \delta_i \pi_i^{\mathrm{P}}(e_i^i, e_j^i, e_i, e_j) \tag{12.18}$$

$$\pi_j^*(e_i, e_j) \leq \pi_j^{\mathrm{R}}(e_i^i, e_j^i) \tag{12.19}$$

$\forall\, i \in \mathrm{I}$ where:

$$\pi_i^{\mathrm{P}} = (1-\delta_i^{t_i^{\mathrm{P}}})\pi_i^{\mathrm{R}}(e_i^i, e_j^i) + \delta_i^{t_i^{\mathrm{P}}} \pi_i^*(e_i, e_j). \tag{12.20}$$

Our objective is to determine the smallest δ_i which is able to sustain the emission tuple $e = (e_i, e_j)$ in the cooperative phase. In a first step assume that $e^i = (e_i^i, e_j^i)$ is given. It follows from (12.18) that π_i^{P} should be minimized, which is equivalent to minimizing $\delta_i^{t_i^{\mathrm{P}}}$ according to (12.20). Solving (12.17), using (12.20), gives:

$$(1-\delta_i^{t_i^{\mathrm{P}}})\pi_i^{\mathrm{R}} + \delta_i^{t_i^{\mathrm{P}}} \pi_i^* \geq \pi_i^{\mathrm{C}} \Leftrightarrow \delta_i^{t_i^{\mathrm{P}}} \geq \frac{\pi_i^{\mathrm{C}} - \pi_i^{\mathrm{R}}}{\pi_i^* - \pi_i^{\mathrm{R}}} \text{ if } \pi_i^* \neq \pi_i^{\mathrm{R}}. \tag{12.21}$$

Hence, the optimal punishment time, $t_i^{\mathrm{P}*}$, is:

$$t_i^{\mathrm{P}*} = \left[\frac{\log\left(\frac{\pi_i^{\mathrm{C}} - \pi_i^{\mathrm{R}}}{\pi_i^* - \pi_i^{\mathrm{R}}}\right)}{\log \delta_i} \right] \text{ approx.: } t_i^{\mathrm{P}*} = \frac{\log\left(\frac{\pi_i^{\mathrm{C}} - \pi_i^{\mathrm{R}}}{\pi_i^* - \pi_i^{\mathrm{R}}}\right)}{\log \delta_i}. \tag{12.22}$$

To simplify things in the computation below, (12.22) is considered without the integer requirement.

Since $\delta_i^{t_i^{\mathrm{P}}} \leq \delta_i$ holds because of $\delta_i < 1$ and $t_i^{\mathrm{P}} \geq 1$:

$$\mathrm{C}_i^{\mathrm{A}}: = \delta_i \geq \frac{\pi_i^{\mathrm{C}} - \pi_i^{\mathrm{R}}}{\pi_i^* - \pi_i^{\mathrm{R}}} \tag{12.23}$$

is a necessary condition for δ_i.

With $t_i^{\mathrm{P}} = t_i^{\mathrm{P}*}$(approx.) it follows from (12.17) and (12.20) that $\pi_i^{\mathrm{P}} = \pi_i^{\mathrm{C}}$. Substituting this into (12.18) leads to the second renegotiation-proof requirement, $\mathrm{C}_i^{\mathrm{B}}$:

$$(1-\delta_i)\pi_i^D + \delta_i \pi_i^C \leq \pi_i^* \Leftrightarrow C_i^B: = \delta_i \geq \frac{\pi_i^D - \pi_i^*}{\pi_i^D - \pi_i^C} \text{ if } \pi_i^* \neq \pi_i^C. \quad (12.24)$$

Hence, for a given $e^i = (e_i^i, e_j^i)$:

$$\delta_i \geq \max \{C_i^A, C_i^B\} \quad (12.25)$$

follows. In a second step we vary $e^i = (e_i^i, e_j^i)$ to minimize max $\{C_i^A, C_i^B\}$ subject to the condition that the punisher should not be worse off than in the cooperative phase ((12.19)). For a given equilibrium emission tuple $e = (e_i, e_j)$, the minimum discount factor requirement of a country i is therefore given by:

$$\delta_i \geq \delta_i^{\min} = \min_{e^i} [\max \{C_i^A, C_i^B\}] \text{ s.t. (12.19).} \quad (12.26)$$

Technically, for a given e_i^i the largest e_j^i is chosen which still satisfies (12.19) (recall $e_j^i \leq d$); this is substituted into C_i^A and C_i^B and the maximum of both is minimized. From Sub-section 12.2.2 we know that e_j^i is given by (12.8) where $i = 1$ and $j = 2$. As argued there, such an e_j^i satisfying (12.19) (which is equivalent to (12.2)) exists for any payoff of country j in the domain of rational payoffs, that is, $\pi_j^* \in [\pi_j^M, \pi_j^U]$.

Observing (12.23) and (12.24), it is clear that, for δ_i close to 1, $C_i^A = C_i^B$. Since we chose $e_i^i = 0$ (harshest possible punishment) in Sub-section 12.2.2, this implies that we minimized C_i^A, and therefore automatically C_i^B as well. Hence, C_1 and C_2 in Figures 12.1 and 12.3 are C_i^A and C_j^A (and/or C_i^B and C_j^B) drawn for $\delta_i = 1$. Since we defined δ_i to be strictly less than 1, all emission tuples lying on the boundary of C_1 or C_2 do *not* belong to the renegotiation-proof emission space. Hence, also all boundary emission or payoff tuples which have been mentioned in the context of SRPE (for example, π^{W_1} and π^{W_2} in Figure 12.1 and e^{W_1} and e^{W_2} in Figure 12.3) are *excluded*.

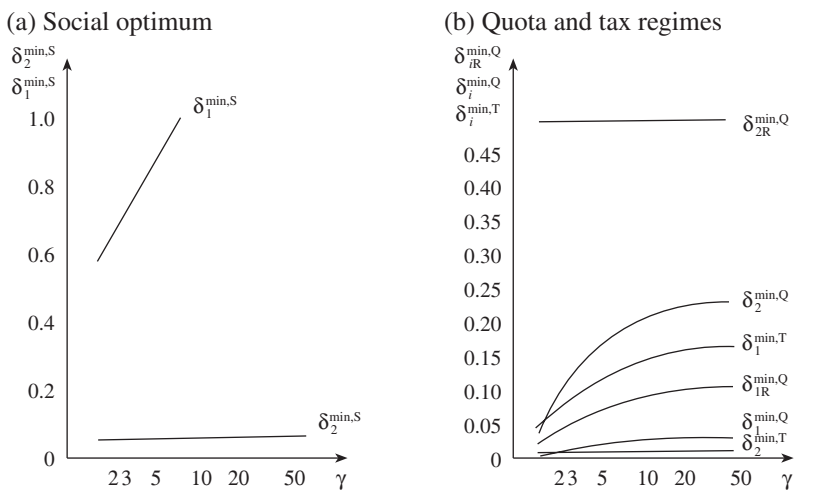
Now, for a discount factor smaller than 1, $e_i^i = 0$ is not always optimal when solving (12.26). Instead, it might be conducive to the stability of an agreement (lower discount factor requirement) to be not so harsh on country i in the punishment phase ($e_i^i > 0$), but to choose a longer punishment duration. For instance, if δ_i is small and $\pi_i^C >> \pi_i^R$, then $t_i^P = 1$ and $e_i^i = 0$ could already be too long to be acceptable to the defector. In other words, there is an *optimal mix* of punishment duration and punishment level.

12.3.2 Illustration of the Minimum Discount Factor Requirement for Weakly and Strongly Renegotiation-proof Equilibria

We are now in the position to determine the minimum discount factor requirement $\delta_i^{\min}$ for given emission tuples $e=(e_i, e_j)$. This is demonstrated for the quota and the tax equilibrium and the social optimum for selected parameter values, based on the payoff functions in (11.8).

First, note that the $\delta_i^{\min}$ requirements are independent of the parameter d. Second, note that only the relation $\gamma = b/c$ is decisive, not the absolute values of b and c.

Let us start by contemplating the effect of γ on stability. The discussion is illustrated in Figures 12.6(a) and 12.6(b) where the $\delta_i^{\min,k}$ requirements, $i \in \{1, 2\}$, $k \in \{Q, T, S\}$, are displayed as a function of γ.[14]



Note: Payoff functions (11.8) and $\Theta = 1$, $\omega = 5$ are assumed. The γ-axis is logarithmic scaled.

Figure 12.6 The discount factor requirement as a function of γ

This particular example assumes that country 2 exhibits opportunity costs of abatement five times higher than country 1, but equal damage. It is evident that the discount factor requirements for both bargaining equilibria and the social optimum increase with an increase of γ. This implies that the more costly abatement is compared to perceived environmental damage, the higher is the discount factor requirement. This result can be traced back to the following factors:

1. An increase in γ implies that equilibrium emissions of all three agreements go up because abatement becomes more costly compared to perceived environmental damage. That is, $\partial e_i^k/\partial\gamma > 0$ holds.
2. It can be shown that for γ going to infinity all emission levels approach $e_i^{\max} = e_i^0 = d$ and net benefits increase under all regimes, converging to the same value. Thus, the higher γ is, the less of a deterrent becomes a punishment threat (which is confined to $e_i^j \leq d \; \forall \; i \in \mathrm{I}$) and at the same time the more difficult it becomes for the punisher to receive a payoff during the punishment exceeding his/her cooperative phase payoff.

From Figure 12.6(a) it is also evident that, in the social optimum, country 1's discount factor requirement is higher than that of country 2. This is due to the fact that country 1 has to contribute more to the socially optimal solution because it can abate emissions at less cost. Hence, this country has a higher free-ride incentive than country 2. In the example the discount factor requirement surpasses the upper boundary of 1 for $\gamma > 8.8$, implying that under no circumstances would country 1 comply with socially optimal abatement duties.

From Figures 12.6(a) and (b), it is apparent that for country 1 the discount factor in the tax equilibrium is lower than in the social optimum. As in the social optimum, the 'abatement conditions' in the tax regime are less favorable for country 1 than for country 2. However, there is a distinct difference: under the tax regime, country 1 is the bottleneck in the negotiations (for these parameter constellations) according to the LCD decision rule and it will propose an abatement target below the socially optimal level.

In the quota regime ($\delta_i^{\min,\mathrm{Q}}$ curves in Figure 12.6(b)) the situation is reversed. For the given parameter constellation, country 2 is the bottleneck. Therefore, this country's discount factor requirement is higher than that of country 1.

We leave discussion of the discount factor requirements of the various agreements at this preliminary stage since this politically interesting issue will be taken up again in the context of coalition formation.[15] For the moment it will be sufficient to note how the minimum WRPE discount factor requirements can be determined and what are the driving forces which determine these requirements.

Though there is no definition in the literature of what a strongly renegotiation-proof equilibrium means in the context of discount factors smaller than 1, it seems obvious to define it as a stage game strategy tuple which is Pareto-efficient; that is, its average payoff tuple in the cooperative phase must be an element of the Pareto frontier and satisfy conditions (12.19)–(12.21) $\forall \; i \in N$ for a particular set of discount factors. Thus, in the

example ($\Theta = 1$ and $\omega = 5$), only the social optimum would be an SRPE if discount factors of all players are close to 1 and if $1.6 \leq \gamma \leq 8.8$ holds, as demonstrated in Figure 12.6(a).

So far we have checked for the stability of a particular cooperative emission tuple. However, according to the procedure for $\delta_i \to 1$ in the previous sections, it would seem 'natural' to determine instead the entire set of WRPE or SRPE emission tuples for given discount factors, similar to what was done for δ_i close to 1 in Figure 12.3. However, in the case of $\delta_i << 1$, even though assuming a particular punishment time and a particular discount factor for all players, the optimal punishment strategy tuple is not uniquely determined and depends on the cooperative emission tuples. In other words, the WRPE inequality system is 'overspecified' and further restrictive assumptions are needed in the case of $\delta_i << 1$ to draw such general graphs as in Figure 12.3 in the case of $\delta_i \to 1$. Such restrictions are discussed in the next two sub-sections.

12.3.3 Renegotiation-proof Trigger Strategies

Another extreme assumption comparable to $e_i^i = 0$ in the case of δ_i close to 1 is $t_i^P = \infty \; \forall \; i \in I$ in the present context of $\delta_i << 1$. Such an infinite punishment constitutes *de facto* a WRPE trigger strategy (see Chapters 4 and 5 for SPE trigger strategies). Such a trigger strategy implies $t_i^P = \infty$ from which $\delta_i^{t_i^P} = 0$ follows since $\delta_i < 1$ by assumption. Then, (12.17), using (12.19), becomes:

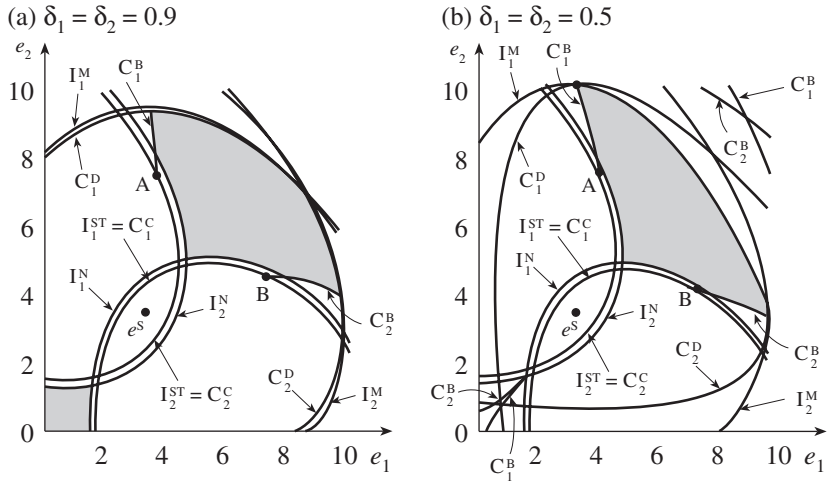
$$\pi_i^P = \pi_i^R = \pi_i^C. \tag{12.27}$$

Equation (12.27) says that for a punishment of infinite duration it is necessary for the punishment emission tuple to lie on country *i*'s best reply function. This is so because otherwise country *i* would not accept such a harsh punishment.

For a given e_j^i, e_i^i is now *uniquely* determined by (12.27). Then (the largest possible) e_j^i is also determined by (12.19). Hence, (12.23) (the former condition C_i^A) can be dropped and we are left with condition C_i^B. In contrast to the previous sections, it is now not always possible to find an e_j^i solving (12.19) for all payoffs in the whole domain of rational payoffs. To ensure that such an e_j^i exists, requirement C_i^C is needed:

$$C_i^C := \pi_i^* \leq \pi_i^{ST} \tag{12.28}$$

where π_i^{ST} stands for Stackelberg payoff. The details of the derivation of C_i^C are laid out in Appendix IX.1. As for $\delta_i \to 1$, one must ensure that for



Note: Payoff functions (11.8) apply; $b = 2$, $c = 1$, $d = 10$ and $\Theta = \omega = 1$ are assumed.

Figure 12.7 Renegotiation-proof emission space for an infinite punishment duration

the punishment $e_j^i \leq d$ holds.[16] Hence, from (12.18) a third condition has to be derived:

$$C_i^D := \delta_i \geq \frac{\pi_i^D - \pi_i^*}{\pi_i^D - \pi_i^M} \tag{12.29}$$

which takes into account $e_j^i \leq d$.[17]

Thus for $t_i^P = \infty$, C_i^B, C_i^C and C_i^D $\forall\, i \in I$ must be simultaneously satisfied. These requirements are shown in Figure 12.7 for two examples. Note that the assumption of symmetric discount factors in the examples is chosen just for convenience; however, it is not important for the following arguments. Let us first concentrate on Figure 12.7(a) where there are ten curves. The minimax and Nash indifference curves of countries 1 and 2 are indicated by I_1^M, I_2^M, I_1^N and I_2^N and have already been discussed in Chapter 9. I_1^{ST} and I_1^{ST} are *Stackelberg indifference curves*, tracing out all emission tuples that give a country the same payoff as if it were a Stackelberg leader (see Section 10.2). These two indifference curves represent conditions C_1^C and C_2^C as given in (12.28).

Thus, all emission tuples lying to the right of the C_1^B curve, above the C_1^C curve and below the C_1^D curve are WRPE from country 1's perspective. By the same token, all emission tuples lying above the C_2^B curve, below the C_1^C curve and to the left of the C_2^D curve are WRPE from country 2's perspective. (The C_i^B curve stops at points A and B because there the C_i^C conditions

become binding.). Taken together, the gray area represents the WRPE emissions space for $t_i^P = \infty$.

It is evident that only emission tuples which generate relatively low payoffs to both countries are WRPE. On the one hand, for $t_i^P = \infty$ the problem is that if the punishment emission level of the defector, e_i^i, is too small, his/her payoff is so low that s/he is not prepared to accept the punishment, and in particular such a long punishment. On the other hand, for high levels of e_i^i it might no longer be possible to satisfy (12.19).

Turning now to Figure 12.7(b), it is obvious the more that actors discount time, for example, $\delta_i = 0.5$ instead of $\delta_i = 0.9$, the smaller becomes the WRPE emission space. This is a similar result as stated in Corollary 7.1, where, we may recall, for $\delta_i \to 0$ the WRPE emission space shrinks to the NE of the stage game. Put differently, the more that actors discount time, the more difficult it becomes to stabilize agreements, deviating substantially from the stage game NE emission tuple.

From Figures 12.7(a) and (b) it is evident that the equilibrium emission tuple of the social optimum and the two bargaining regimes (which all coincide due to the assumption of $\Theta = \omega = 1$) are not renegotiation-proof for $t_i^P = \infty$ (e^S lies outside the gray area). It can be shown that this finding carries over as a general result:

Proposition 12.6
Let the stage game of the infinitely repeated emission game be described by (11.8) and conditions A_3 in (VIII.4) hold, then irrespective of the parameter values, the punishment level and the discount factors of players, the social optimum, the quota and the tax equilibrium cannot be backed by renegotiation-proof trigger strategies.

Proof: Compute net benefits for the emission levels as given in Table VIII.1 in Appendix VIII.2 and show that for any parameter constellation at least for one country $\pi_i^k > \pi_i^{ST}$, $k \in \{S, Q, T\}$ holds and hence condition C_i^C (see (12.28)) is violated. QED[18]

The result nicely stresses the difference between SPE and WRPE. For instance, the quota equilibrium can *always* be backed as an SPE by using a trigger strategy which reverts to the stage game NE in case of defection, since $\pi_i^Q > \pi_i^N \; \forall \; i \in I$. For the tax equilibrium and the social optimum this may also be possible; however, this depends on the specific parameter values (since $\pi_i^S < \pi_i^N$ and $\pi_i^T < \pi_i^N$ may hold). However, Proposition 12.6 clearly denies this possibility for renegotiation-proof trigger strategies.

12.3.4 Restrictions of Punishment Space

Whereas in the context of discount factors close to 1 restrictions of the punishment space showed up in smaller WRPE, SRPE and SSPE payoff or emission spaces, in the context of discount factors smaller than 1 this implies higher minimum discount factor requirements. For instance, restricting the upper limit of the punishment to NE emissions, that is, $e_j^i \leq e_j^N$, then minimum discount factor requirements of the quota equilibrium in Figure 12.6 increase for both countries almost twice as much as without this restriction. The $\delta_{1R}^{\min,Q}$ and $\delta_{2R}^{\min,Q}$ curves represent the minimum discount factor requirement for this restriction where we may recall that the $\delta_1^{\min,Q}$ and $\delta_2^{\min,Q}$ curves are those already discussed for $e_j^i \leq e_j^0 = e_j^{\max} = d$.

Of course, all the other restrictions discussed in Sub-section 12.2.4 can also apply here and because of their obvious effect they are not discussed any further. A particular restriction would be to require an efficient punishment which may be regarded as a necessary condition for a strongly perfect equilibrium for a given set of discount factors. Similar to renegotiation-proof trigger strategies, where it was possible to determine the punishment strategy tuple uniquely (see also Appendix IX), the condition for Pareto efficiency (see (9.32)) also defines a one-to-one mapping of e_i^i and e_j^i. Hence, similar conditions as C_i^B, C_i^C and C_i^D in Figure 12.7 may be derived and plotted for different values of the discount factors in order to obtain the set of SSPE. Since the procedure is similar to that in the previous sub-section, we skip discussing the details.

NOTES

1. Parts of this chapter draw on Endres and Finus (1998a) and Finus and Rundshagen (1998b).
2. Applications of the WRPE and SRPE concept, assuming $\delta_i \to 1 \; \forall \; i \in I$, may be found in the context of a global emission game with N players in Barrett (1994a, b) and in the context of a Cournot duopoly in Driffill and Schultz (1995) and Schultz (1994).
3. In other words, we search for the lowest WRPE payoff which is sustainable in the emission game.
4. $\pi_2^U = \dfrac{d^2 b^2 \omega^2}{2(b\omega + c\Theta)}$ which follows from $\max\limits_{e_2} \pi_2(e_1 = 0, e_2)$ as explained in Section 9.5.
5. Note that condition (12.6) is satisfied for $e_2^1 = d$. This follows from $\partial^2 \pi_2^R / \partial e_2^{1\,2} < 0$ (π_2^R is the RHS term in (12.6)) and the fact that $e_2^1 = d$ is bigger than the LHS expression and smaller than the RHS expression in (12.7) by assumption.
6. This amounts to maximizing $\lambda\pi_1(e_1^1, e_2^1) + (1-\lambda)\pi_2(e_1^1, e_2^1)$ for $\lambda = 0$. See Chapter 9, note 26.
7. The associated punishment emissions are:

$$e_1^1 = 0,\; e_2^1 = \frac{bd\omega}{b\omega + c\Theta},\; e_1^2 = \frac{bd}{b+c} \text{ and } e_2^1 = 0.$$

8. Barrett (1994a) determines the parameter ranges for his models which allows one to stabilize a socially optimal IEA. That is, he computes π_i^U, π_i^C and π_i^S for symmetric games and then determines the parameter values for which $\pi_i^U > \pi_i^S > \pi_i^C \; \forall \; i \in I$ is possible.
9. Though in the auction equilibrium $\pi_i^A \geq \pi_i^N \; \forall \; i \in I$ holds if $e^I = e^N$, this does not suffice to establish stability in the sense of WRPE since $\pi_i^* \geq \pi_i^N$ is only a necessary WRPE condition but not a sufficient one. Additionally, one has to test whether conditions C_1 and C_2 can be satisfied subject to the restriction $e_i \in [0, e_i^N] \; \forall \; i \in I$. For this payoffs π_i^A must be computed which, as pointed out above, is only possible for specific parameter values but not at a general level. Of course, $\pi_i^A \geq \pi_i^N \; \forall \; i \in I$ is sufficient to qualify the auction equilibrium as an SPE.
10. For a similar interpretation in the context of a global emission game using the stability concept of the core, see Sub-section 13.3.4.
11. There is little scope for legal punishments in common international law. There are only two ways in which the injured party can legally punish the breach of a *bilateral* treaty. On the one hand, the injured party can suspend or terminate the violated treaty provisions, however, *only to the extent to which they are violated by the other party* (Simma 1970, pp. 20ff.). On the other hand, the injured party can take reprisal measures against the violator of the treaty. These measures may consist of reciprocal violations of the violated treaty provisions. However, reprisals must also be in proportion to the delict according to the generally recognized *principle of proportionality* in common international law (Kelsen and Tucker 1967, pp. 20ff.). It is worth mentioning that leading scholars of international law think that neither of the described reactions to a breach of a treaty constitutes legal options for parties of a multilateral treaty, that is $N > 2$ (Heister 1997, pp. 91, 135). Only if punishment rules are an explicit part of a multilateral environmental treaty are punishments covered by international law (Sand 1992, pp. 14ff.).
12. Of course, one could argue that $e_i(e_j) - e_i < 0$ should not be punished at all because a country would abate more than required. However, agreeing on a high emission level e_i ($e_i >> e_i^N$) implies also a sub-optimal choice from both countries' point of view and should therefore not be regarded as stable *per se*.
13. Drawing the curves $C_1^{(1)}$, $C_2^{(1)}$, $C_1^{(2)}$ and $C_2^{(2)}$ in the payoff space would reveal exactly the same pattern. That is, these curves lie more inward than the curves C_1 and C_2 in Figure 12.1.
14. All parameters have been chosen to ensure that assumptions A_3 in (VIII.4), Appendix VIII.2 hold. This explains, for instance, why the smallest γ value in Figure 12.6 is 1.6.
15. The interested reader may consult Finus and Rundshagen (1998b), where the discussion is continued based on the present setting.
16. Recall, for $\delta_i \to 1 \; \forall \; i \in I$ the restriction $e_j^i \leq d$ delivered conditions L_1 and L_2 in the payoff space and the minimax indifference curves in the emission space.
17. Thus, C_i^D is basically C_i^B if $e_j^i = d$ has been substituted into $\pi_i^C(e_i(e_j^i), e_j^i)$.
18. A detailed proof is available upon request. With respect to the auction equilibrium, a statement is not possible for the reasons given in note 9.

13. Coalition models: a first approach

13.1 INTRODUCTION

In the remaining chapters we are concerned with the coalition formation process in an N-country world. Though many of the previous results have either already been derived for more than two countries or could easily be extended to the case of $N>2$, the possibility that sub-groups of players may form strategic alliances has been neglected. On theoretical grounds, one would like to explain an entire coalition formation process endogenously. However, it turns out that this is an extremely complex undertaking. This is because the number of possible coalition formations is usually very large. For instance, there may be not only a group of signatories and non-signatories, but several coalitions may coexist. On the one hand, signatories may split up into sub-groups with homogeneous interests since within the group of 'environmentally conscious' countries no agreement on abatement obligations may be possible. On the other hand, there is a possibility that less environmentally concerned countries could form a coalition to defend their interests.

However, things are even more complex. The decision to join a particular coalition depends not only on the abatement obligations of its members but also on the abatement strategy of other coalitions or single players and the possible transfer scheme (for example, in order to compensate some players) within this particular coalition and within alternative coalitions. Thus the payoff of a country depends on the *strategy triple membership*, *emission allocation* and *transfer scheme* where the *strategy choices of players influence each other mutually*. Considering the fact that the number of possible emission allocations and transfer schemes is basically infinitely large, it is evident that if the number of players is also large (and hence the number of possible memberships is large) some simplifying assumptions are necessary to keep a model tractable.

The following models make different assumptions. For instance, the *conjectural variation models* (Section 13.2) consider only possible deviation from an equilibrium coalition by a single country and discard the possibility that a sub-group of countries may jointly deviate. Moreover, equilibrium strategies are based on the myopic behavior of players, who consider only the immediate reaction of fellow players to a change in their strategy

but not the subsequent moves (chain reactions) which may be triggered. Additionally, these models consider only one coalition (signatories), the members of which coordinate their strategies and assume that all other countries (non-signatories) behave as singletons (see Carraro and Siniscalco 1998). The *concept of the core* (Section 13.3) is defined either in a static world or in a dynamic setting. However, in the latter case only Markov (history-independent) strategies are considered. Moreover, the core abstracts from the individual strategies of players and reduces decisions to the payoff dimension (characteristic function). The *supergame coalition model* (Chapter 14), although it takes up some of the concerns raised with respect to the models presented in this chapter, also restricts the coalition formation process to signatories and non-signatories. Moreover, as with the previous models, the choice of abatement targets and transfer schemes within the coalition is basically exogenous to the model, though a large number of alternative abatement targets and transfer schemes are considered and their selection is thoroughly justified.

Due to these conceptual shortcomings, we shall investigate in Chapter 15 whether new developments in the game theoretical literature on coalition formation are suitable to be integrated into the analysis of the genesis of IEAs. From the discussion it will be apparent that there is still much work to be done in order to construct models which are capable of explaining the entire coalition formation process endogenously.

To the best of our knowledge, all theoretical coalition models on the formation of IEAs assume an externality of the pure public bad type. This is because for transboundary environmental problems, different spillovers cause an additional asymmetry between countries, the strategic implications of which would have to be considered.[1] Hence in the tradition of previous chapters, we again restrict attention to global environmental problems.

13.2 CONJECTURAL VARIATION MODELS

13.2.1 Preliminaries

The *conjectural variation coalition models* first made their appearance in the analysis of oligopolies (see, for example, d'Aspremont and Gabszewicz 1986; and d'Aspremont *et al.* 1983). In the context of the formation of IEAs, their origin can be traced back to Barrett (1991c, 1992e) and Carraro and Siniscalco (1991). Typical representatives of these models are Barrett (1994b); Bauer (1992); Carraro and Siniscalco (1991); and Hoel (1992a).

These models are typically set up as *stage games* comprising two or three

stages. In the first stage players decide whether to *participate* in an agreement. It is assumed that this is a *binary choice*: 'join' and 'do not join', that is, $P_i = \{j_i, nj_i\}$. (Hence, the coexistence of several coalitions is ruled out by assumption!) The equilibrium number of coalition members is denoted by N^* and therefore the number of non-signatories is $N - N^*$. For notational convenience it is assumed that countries 1 to N^* form a coalition and countries $N^* + 1$ to N remain outside the coalition. Consequently, the set of players is given by $I^J = \{1, ..., N^*\}$, $I^{NJ} = \{N^* + 1, ..., N\}$, and therefore $I = I^J \cup I^{NJ}$ and $I^J \cap I^{NJ} = \emptyset$ hold.

In the second stage players choose their *emission levels*. Usually, the following assumptions are made: *signatories* choose within their group emissions *cooperatively*, that is, they maximize aggregate welfare of the coalition members. Towards outsiders signatories behave *non-cooperatively*. *Non-signatories* continue to play as singletons and choose their non-cooperative emission levels. Therefore, the strategy set of the second stage comprises emissions of signatories, $e^{J*} = (e_1^{J*}, ..., e_{N^*}^{J*})$ and those of non-signatories, $e^{NJ*} = (e_{N^*+1}^{NJ*}, ..., e_N^{NJ*})$.

In the third stage the *allocation of the welfare gains* among the coalition members is decided. If countries are symmetric, this stage seems superfluous since, intuitively, one would expect that signatories would agree on a symmetric abatement allocation which implies equal payoffs and therefore that no transfer scheme is needed.[2] Probably, any morally motivated welfare allocation rule or a focus point type of argument would not alter this result. Hence, it is also not surprising that all the well-known bargaining concepts of cooperative game theory would not call for a reallocation of payoffs via transfers.

However, if countries are asymmetric the choice of the transfer scheme is less obvious. In particular, stability of a coalition may depend on the *sharing rule* of the gains of the coalition. Typical assumptions are: (a) no transfers; (b) Nash bargaining solution; and (c) Shapley value, where the latter two are welfare allocation rules of cooperative game theory.[3] Assumption (a), again, implies that the third stage is redundant and each signatory receives those payoffs which directly follow from the emission and participation strategies. That is, $\pi_i^J(p^*, e^*)$ where $p^* = (j_1^*, ..., j_{N^*}^*, nj_{N^*+1}^*, ..., nj_N^*)$, $e^* = (e^{J*}, e^{NJ*})$ and where the asterisk denotes an equilibrium strategy combination. Only for assumptions (b) and (c) is the third stage important where the result of the first two stages may be altered. In this case $\pi_i^J(p^*, e^*, t^*)$ where t^* denotes a matrix of transfers.

In the literature there are basically two assumptions regarding the *sequence of moves* in the first two stages: (a) players choose their strategies in both stages *simultaneously* (Carraro and Siniscalco 1991; Bauer 1992; Hoel 1992a); (b) players choose their participation strategy in stage one

simultaneously, but emission levels in the second stage *sequentially* (Barrett 1991c, 1992e). In the following, the first assumption is referred to as the *Nash–Cournot assumption* and the second as the *Stackelberg assumption*. Theoretically, the latter assumption could imply that either non-signatories or signatories behave as Stackelberg leaders. However, since in the following models non-signatories are assumed to act as singletons and since it is known from Chapter 10 that one cannot solve for more than one Stackelberg leader, only Stackelberg leadership of the signatories (which basically act as a single player) has been assumed in the literature so far. Thus, strictly speaking, the Stackelberg assumption implies that the second stage consists of two *sub-stages*.

As we pointed out in Chapter 10, the assumption of Stackelberg leadership is sometimes difficult to justify. In the present context it may be argued that signatories are better informed than non-signatories about emission levels in other countries since they coordinate their environmental policies within an IEA. Moreover, signatories cooperate by forming a 'political bloc' against outsiders and therefore assume a stronger position in international politics than non-signatories who 'only' pursue their self-interests. Nevertheless, some qualms about this assumption remain since these explanations are exogenous to the model and argue for an asymmetry of information, though, at least in Barrett (1994a, b), countries are modeled as being symmetric in all other respects.

There are three conditions which have to be met in a *conjectural variation coalition equilibrium*. The first condition, C_1, may be seen as a basic prerequisite for a stable coalition and concerns the *profitability* to each member. This condition implies that each coalition member must be better off than in the status quo where no coalition has formed. The second condition, C_2, requires that no signatory is better off by leaving the coalition (*internal stability*); and the third condition, C_3, requires that the situation of a non-signatory cannot be improved by joining the coalition (*external stability*).

To ease the following definitions, it will prove helpful to change the previous notation slightly. Therefore, we write $\pi_i^J(N^*, e^*, t^*)$ instead of $\pi_i^J(p^*, e^*, t^*)$ and $\pi_i^J(0, e^N)$ instead of $\pi_i^J(p^{NJ}, e^N)$ (where $p^{NJ} = \{nj_1, nj_2, ..., nj_N\}$) to emphasize the number of signatories.[4]

Definition 13.1: Profitability of a coalition
A coalition is profitable if $C_1 := \pi_i^J(N^*, e^*, t^*) - \pi_i^J(0, e^N) \geq 0 \ \forall \ i \in I^J$.

Definition 13.2: Stability of a coalition[5]
1. *Internal stability* There is no incentive for a signatory to leave the coalition. That is, $C_2 := \pi_i^J(N^*, e^*, t^*) - \pi_i^{NJ}(N^* - 1, e^{*\prime}, t^{*\prime}) \geq 0 \ \forall \ i \in I^J$.

2. *External stability* There is no incentive for a non-signatory to join the coalition. That is, $C_3: = \pi_j^J(N^* + 1, e^{*\prime}, t^{*\prime}) - \pi_j^{NJ}(N^*, e^*, t^*) \leq 0 \; \forall$ $j \in I^{NJ}$.

Definition 13.2 states typical conjectural variation equilibrium conditions as encountered in Chapter 10 (see in particular Definition 10.1, Section 10.4). Strategies N^*, e^*, t^* are the 'state variables' and $N^* - 1$, $N^* + 1$, $e^{*\prime}$ and $t^{*\prime}$ indicate a deviation from this state which, according to the definition, should not be beneficial to any player in equilibrium. The definition implies that a player who belongs to a particular group assumes that if s/he alters his/her participation decision, all other players will *remain* in their groups. This is a similar assumption as in a Nash equilibrium (NE) where a best strategy of a player is chosen, given the best replies of all other $N-1$ players. In the context of a global externality this assumption almost follows from the incentive structure of the game itself. If a signatory decides to deviate, s/he has *no* incentive to ask other signatories to follow suit since s/he benefits from the abatement activities of the coalition. Also a non-signatory will find it hard to convince other non-signatories to join the coalition since they may be better off remaining outside the coalition by benefiting from the increased abatement efforts of the enlarged coalition without having to contribute to its success. Of course, in some respects this assumption implies that players are myopic. One can easily imagine that if, for instance, a country leaves the coalition other signatories may follow suit (since the smaller coalition may not be profitable any more) or, alternatively, a new coalition forms which is not considered in the above definition.

From Definition 13.2 it should be evident that with respect to the emission and transfer strategies there is no similarity to the (static) NE. A potential defector takes into account that if s/he left the 'old' coalition non-signatories would alter their strategies immediately (reoptimization). By the same token, a non-signatory who is thinking about becoming a coalition member considers that the 'new' and enlarged coalition as well as the former non-signatories would instantly choose other strategies after his/her accession to the IEA.

With these definitions the equilibrium number of signatories can now be determined. Unfortunately, this is not possible at a general level and has to rely on specific payoff functions. It will turn out that, apart from the sequence of moves, the functional form of the payoff function is crucial to the equilibrium number of signatories. Three types of payoff functions will be considered:

$$\text{Type 1: } \pi_i = b_i\left(de_i - \frac{1}{2}e_i^2\right) - \frac{c_i}{N}\left(\sum_{k=1}^{N} e_k\right) \tag{13.1}$$

$$\text{Type 2: } \pi_i = b_i e_i - \frac{c_i}{2N}\left(\sum_{k=1}^{N} e_k\right)^2 \tag{13.2}$$

$$\text{Type 3: } \pi_i = b_i\left(de_i - \frac{1}{2}e_i^2\right) - \frac{c_i}{2N}\left(\sum_{k=1}^{N} e_k\right)^2. \tag{13.3}$$

However, it will prove helpful to restrict the analysis to *symmetric* countries for a start (Sub-sections 13.2.2–13.2.4). That is, $b_i = b_j = b$, $c_i = c_j = c\ \forall\ i$ and $j \in \mathrm{I}$. Hence, we can abstract from transfers and only have to consider two stages for the moment.

Payoff function of type 1 ((13.1)) implies linear marginal benefits and constant marginal damage. From Chapter 9 it is known that this implies *orthogonal* reaction functions. The second type of payoff function ((13.2)) implies constant marginal benefits and linear marginal damage costs. Hence, reaction functions have a slope of -1. The third type of payoff function ((13.3)) implies linear marginal benefits and costs and satisfies assumption A_1 in Chapter 9. Reaction functions are downward sloping in emission space with slope less than 1 in absolute terms.

The division of damages by N symbolizes that each country suffers $1/N$ of total damage. However, technically this is only a scaling factor: *all* subsequent results remained valid if one assumes $\phi_i = f_i(\Sigma e_k)$ instead of $\phi_i = (1/N) f_i(\Sigma e_k)$. Note that in the literature on coalition formation some papers use slightly different functional forms or specify the externality problem in terms of emission reductions (for example, Barrett 1994b; Bauer 1992; see also Chapter 9, note 5). However, the *only* point that matters for the qualitative results obtained below is whether reaction functions are orthogonal, negatively sloped with slope less than 1 or equal to 1 in absolute terms.

For the subsequent analysis, it will prove convenient to recall a result mentioned in Chapter 10 in the context of non-Nash behavior (Section 10.4). There it was shown that the greater the number of countries involved in the externality problem, the more important it is from a global point of view to reach a cooperative agreement, that is, $\partial(\Sigma e_k^N - \Sigma e_k^{\mathrm{S}})/\partial N > 0$. In the following it turns out that more general conclusions are possible if the 'degree of externality' is not defined in absolute, that is, $\Sigma e_k^N - \Sigma e_k^{\mathrm{S}}$, but in relative terms, that is, $(\Sigma e_k^N - \Sigma e_k^{\mathrm{S}})/\Sigma e_k^{\mathrm{S}}$.[6,7] For all three types of payoff functions one finds (assuming symmetric countries and $\gamma = b/c$):[8]

$$\mathrm{I}_1 := \frac{\Sigma e_k^N - \Sigma e_k^{\mathrm{S}}}{\Sigma e_k^{\mathrm{S}}};\ \frac{\partial \mathrm{I}_1}{\partial N} > 0 \text{ for (13.1)–(13.3) and } \frac{\partial \mathrm{I}_1}{\partial \gamma} < 0 \text{ for (13.1) and (13.3).} \tag{13.4}$$

With respect to the parameter $\gamma = b/c$ the relations hold for payoff functions of types 1 and 3; for type 2 the index I_1 does not contain b and c at all. $\partial I_1/\partial\gamma < 0$ implies that abatement is particularly attractive from a global point of view if environmental damage is high compared to the opportunity costs of abatement. A similar index can also be defined in terms of payoffs:[9]

$$I_2 := \frac{\Sigma\pi_k^S - \Sigma\pi_k^N}{\Sigma\pi_k^S};\ \frac{\partial I_2}{\partial N} > 0 \text{ for (13.1)–(13.3) and } \frac{\partial I_2}{\partial\gamma} < 0 \text{ for (13.1) and (13.3).} \tag{13.5}$$

In the following we shall first consider the Nash–Cournot assumption and subsequently turn to the Stackelberg assumption. All results are summarized in three propositions in Sub-section 13.2.4.

13.2.2 Nash–Cournot Assumption: Symmetric Countries

As we have shown in previous chapters, the two-stage game must be solved by backwards induction. Hence, the equilibrium emissions of the second stage have to be determined first, assuming the equilibrium number of countries, N^*, to be given. Thus, assuming the first type of payoff function (and symmetric countries, that is, $b_i = b_j = b$, and $c_i = c_j = c$), signatories perform:

$$\max_{e_i^J} N^* \left[\left(b\left(de_i^J - \frac{1}{2} e_i^{J^2} \right) - \frac{c}{N} (N^* \cdot e_i^J + \Sigma e_j^{NJ}) \right) \right] \tag{13.6}$$

which is easily solved since countries have a dominant strategy (towards non-signatories):[10]

$$e_i^{J^*}(N^*) = \frac{bdN - N^* \cdot c}{Nb}. \tag{13.7}$$

Thus, the more signatories, the higher are the abatement duties (the lower are emissions) of a coalition member. In contrast, non-signatories, which are supposed to behave as singletons, perform:

$$\max_{e_j^{NJ}} \left[\left(b\left(de_j^{NJ} - \frac{1}{2} e_j^{NJ^2} \right) - \frac{c}{N} (\Sigma e_i^J + e_j^{NJ} + \Sigma e_{-j}^{NJ}) \right) \right] \tag{13.8}$$

which delivers:

$$e_j^{NJ^*} = \frac{bdN - c}{Nb}. \tag{13.9}$$

It is obvious that non-signatories' emission levels are independent of the number of signatories and the emission level of signatories. That is, each non-signatory has a *dominant* strategy and $e_j^{NJ^*} = e_j^N$.

To check for the stability of a coalition, emissions of a signatory leaving the coalition (internal stability check) and a non-signatory joining the coalition (external stability check) also have to be determined. Due to the simple structure of this game, it follows directly from (13.7) that signatories' emissions are given in these cases by:

$$e_i^{J*'}(N^*-1)=\frac{bdN-(N^*-1)\cdot c}{Nb},\ e_i^{J*'}(N^*+1)=\frac{bdN-(N^*+1)\cdot c}{Nb} \tag{13.10}$$

which is sufficient to compute conditions C_1, C_2 and C_3:

$$C_1=\frac{c^2(N^*-1)^2}{2bN^2},\ C_2=-\frac{c^2(N^*-1)(N^*-3)}{2bN^2},\ C_3=-\frac{c^2N^*(N^*-2)}{2bN^2}. \tag{13.11}$$

From (13.11) it follows that $C_1\geq 0$ is always satisfied, $C_2\geq 0$ if $1\leq N^*\leq 3$ and $C_3\geq 0$ if $N^*\geq 2$. Hence, the equilibrium number of signatories is either $N^*=2$ or $N^*=3$.[11] In order to evaluate these equilibria, we compute:

$$\frac{\partial\mu^*}{\partial N}<0,\ \frac{\partial I_3}{\partial N}>0,\ \frac{\partial I_3}{\partial\gamma}<0,\ \frac{\partial I_4}{\partial N}>0,\ \frac{\partial I_4}{\partial\gamma}<0,$$

$$\mu^*=N^*/N,\ \gamma=\frac{b}{c},\ \Sigma e_k^*=\Sigma e_i^J+\Sigma e_j^{NJ}<\Sigma e_k^N,\ \Sigma\pi_k^*=\Sigma\pi_i^J+\Sigma\pi_j^{NJ}>\Sigma\pi_k^N, \tag{13.12}$$

$$I_3:=\frac{\Sigma e_k^*-\Sigma e_k^S}{\Sigma e_k^S},\ I_4:=\frac{\Sigma\pi_k^S-\Sigma\pi_k^*}{\Sigma\pi_k^S}.$$

That is, the greater the number of countries involved in the externality problem, the smaller will be the percentage of signatories. Though the coalition improves upon the status quo, whenever cooperation would be needed most from a global point of view (large N and small γ; see (13.4) and (13.5)), equilibrium emissions and welfare differ markedly from the social optimum.

For the second type of payoff function it is easy to show that, due to the slope of -1, no coalition exists. Any emission reduction by the coalition is offset by non-signatories. Hence, the profitability of a coalition always fails to hold and the equilibrium of the game is the NE. Accordingly, $I_1=I_3$ and $I_2=I_4$ and the comparative static results in (13.12) carry over to this type of payoff function as well (except that $\partial I_3/\partial\gamma$ and $\partial I_4/\partial\gamma$ are zero).

For the third type of payoff function the equilibrium number of signatories is determined in the same way as was laid out for the first type of payoff

function. However, since the damage cost functions are quadratic, no dominant strategy exists, which renders computation more cumbersome. It turns out that the equilibrium coalition size is $N^* = 2$ and that the profitability constraint C_1 may fail to hold, depending on the parameter values. The relations in (13.12) also hold for this type of payoff function.

Carraro and Siniscalco (1992, 1993) frequently stress that the reason for the greater coalition size in the case of payoff functions of type 1 compared to that of types 2 and 3 is orthogonal reaction functions. They point out that in the latter two cases there is a higher free-rider incentive on the side of the non-signatories which expand their emissions compared to the Nash equilibrium. Hence, the abatement efforts of the signatories are completely (type 2) or partially (type 3) compensated. In contrast, payoff functions of type 1 imply that non-signatories choose their emissions *irrespective* of the abatement efforts of the signatories. In other words, there is no leakage effect.[12] Though this argument is valid to explain the differences in the equilibrium coalition size for the different payoff functions under the Nash–Cournot assumption, it cannot be applied to the Stackelberg assumption.

13.2.3 Stackelberg Assumption: Symmetric Countries

Stackelberg leadership of signatories implies that, when choosing their cooperative emission levels, signatories will take the reaction of the non-signatories into account. As demonstrated in Chapter 10, this may secure the leader a strategic advantage. However, since non-signatories have a dominant strategy in the case of payoff functions of type 1, there is nothing to be gained by assuming the Stackelberg leader position. Hence $N^* = 2$ or $N^* = 3$ as derived for the Cournot assumption and the relation in (13.12) apply.

Conclusions with respect to payoff functions of type 2 are also straightforward. Assume that no coalition has formed in the initial stage and that the coalition gradually expands. Then two cases have to be considered. On the one hand (case 1), in contrast to Chapter 10 where the Stackelberg leader *expanded* emissions, the 'obvious' aim of a coalition is to *reduce* emissions. Of course, any preliminary coalition cannot achieve this aim if non-signatories completely offset signatories' abatement efforts. On the other hand (case 2), though this seems less intuitive, it must also be checked whether a stable coalition can be reached if the countries of the preliminary coalition *expand* emissions. This is in fact the case. Since assuming a Stackelberg position is beneficial to signatories, there is always an incentive for a non-signatory to become a signatory as long as $N^* < N$ and hence the coalition is expanded until $N^* = N$.[13] Since any coalition is assumed to maximize the aggregate welfare of its members, this implies that the grand

coalition realizes the social optimum. Since by the (questionable) assumption (see the discussion in Section 13.8) that a signatory leaving the coalition behaves as a Stackelberg follower, $N^* = N$ is also internally stable. Taken together, under the Stackelberg assumption the grand coalition is realized for payoff functions of type 2.[14] Hence, trivially, $I_3 = 0$, $I_4 = 0$ and $\mu^* = 1$ hold.

Payoff functions of type 3 exhibit a more interesting pattern. We start at the second stage of the game in which emission levels are determined. As pointed out above, this stage itself is divided into two sub-stages. Since signatories behave as Stackelberg leaders, we first have to determine the reaction functions of non-signatories, assuming N^* to be given. These are derived from:

$$\max_{e_j^{NJ}} b\left(de_j^{NJ} - \frac{1}{2}e_j^{NJ^2}\right) - \frac{c}{2N}(e_j^{NJ} + \Sigma e_{-j}^{NJ} + \Sigma e_i^{J})^2 \tag{13.13}$$

where Σe_{-j}^{NJ} denotes emissions of all non-signatories except j, and Σe_i^{J} are aggregate emissions of signatories. Using $\mu^* = N^*/N$, $\Sigma e_j^{NJ} = (1-\mu^*)\cdot N \cdot e_j^{NJ}$ and $\Sigma e_j^{NJ} = \Sigma e_{-j}^{NJ} + e_j^{NJ}$, the FOC deliver the *'aggregate' reaction function* of non-signatories:

$$\Sigma e_j^{NJ} = \frac{(bdN - c\Sigma e_i^{J})(1-\mu^*)}{b + c(1-\mu^*)} \tag{13.14}$$

which exhibits the familiar property of a negative slope of less than 1 in absolute terms. Thus, abatement efforts by signatories will be partially offset by an expansion of emissions by non-signatories.

Signatories, maximizing the welfare of the coalition, perform:

$$\max_{e_i^{J}} \mu^* \cdot N \cdot \left[b\left(de_i^{J} - \frac{1}{2}e_i^{J^2}\right) - \frac{c}{2N}(\mu^* \cdot N \cdot e_i^{J} + \Sigma e_j^{NJ}(\Sigma e_i^{J}))^2 \right] \tag{13.15}$$

where the information about $\Sigma e_j^{NJ}(\Sigma e_i^{J})$ in (13.14) is used. After some basic manipulations, the FOC deliver:

$$\Sigma \tilde{e}_i^{J} = \frac{N\mu^* d(\gamma^2 + 2\gamma - 2\gamma\mu^* + 1 - 2\mu^* + \mu^{*2} - \gamma N\mu^* + \gamma N\mu^{*2})}{\gamma^2 + 2\gamma - 2\gamma\mu^* + 1 - 2\mu^* + \mu^{*2} + \gamma N\mu^{*2}}. \tag{13.16}$$

Substituting (13.16) into (13.14), one derives non-signatories' emission levels:

$$\Sigma \tilde{e}_j^{NJ} = \frac{N(1-\mu^*)d(\gamma^2 - 2\gamma\mu^* + \gamma N\mu^{*2} + \gamma - \mu^* + \mu^{*2})}{\gamma^2 + 2\gamma - 2\gamma\mu^* + 1 - 2\mu^* + \mu^{*2} + \gamma N\mu^{*2}}. \tag{13.17}$$

A routine check reveals that neither (13.16) nor (13.17) need necessarily be positive. In other words, there may be boundary solutions for which either

signatories' or non-signatories' aggregate emissions have to be set to zero. Hence, we have:

$$\Sigma e_i^{\mathrm{J}*} = \begin{cases} 0 & \text{if } \Sigma\tilde{e}_j^{NJ} < 0 \\ \Sigma\tilde{e}_i^{\mathrm{J}} & \text{if } \Sigma\tilde{e}_j^{NJ} \in [0,\ \gamma dN] \\ \gamma dN & \text{if } \Sigma\tilde{e}_j^{NJ} > \gamma dN \end{cases} \tag{13.18}$$

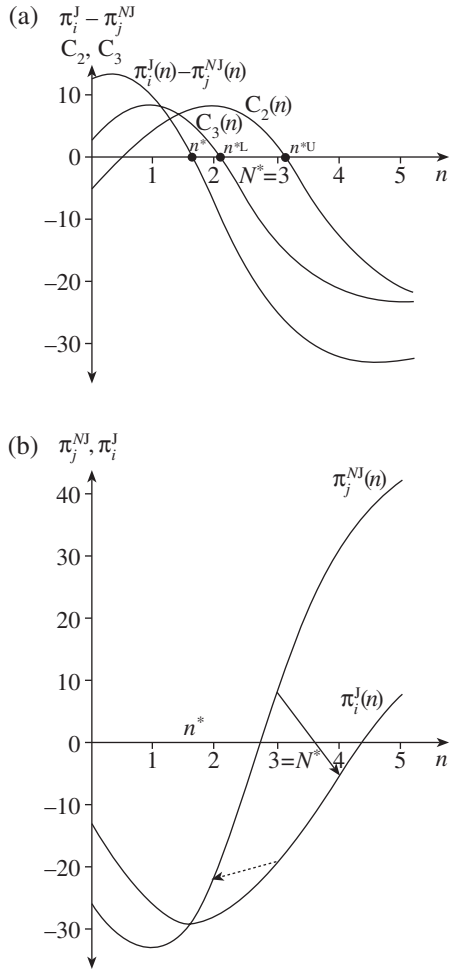
$$\Sigma e_j^{NJ*} = \begin{cases} 0 & \text{if } \Sigma\tilde{e}_i^{\mathrm{J}*} = \gamma dN \\ \Sigma\tilde{e}_j^{NJ} & \text{if } \Sigma\tilde{e}_i^{\mathrm{J}*} \in [0,\ \gamma dN] \\ (1-\mu)\dfrac{\gamma dN}{\gamma + (1-\mu)} & \text{if } \Sigma\tilde{e}_i^{\mathrm{J}*} = 0 \end{cases} \tag{13.19}$$

where an interior solution implies $\Sigma e_i^{\mathrm{J}*} = \Sigma\tilde{e}_i^{\mathrm{J}}$ and $\Sigma e_j^{NJ*} = \Sigma\tilde{e}_j^{NJ}$. Note that if $\Sigma e_i^{\mathrm{J}*} = 0$, then $\Sigma e_j^{NJ*} > 0$ and if $\Sigma e_j^{NJ*} = 0$ then $\Sigma e_i^{\mathrm{J}*} > 0$. That is, aggregate emissions are strictly positive. It is straightforward to show that it can never be a best choice for the signatories to choose emissions above $\Sigma e_i^{\mathrm{J}*} = \gamma dN$, implying $\Sigma e_j^{NJ*} = 0$, which therefore may be regarded as some kind of an upper bound in emission space in this coalition game.[15]

(13.18) and (13.19) are sufficient to compute C_1, C_2 and C_3 in order to determine the equilibrium number of signatories. In contrast to Barrett, who claims that this has to rely on simulations, we propose below an approach which allows us to obtain most information analytically. Nevertheless, it is helpful to work with an example, which is illustrated in Figure 13.1.

In Figure 13.1(b) the payoff function of a representative signatory, $\pi_i^{\mathrm{J}}(n)$, and of a representative non-signatory, $\pi_j^{NJ}(n)$, as a function of the number of coalition members, $n \in [0, N]$, are drawn. In Figure 13.1(a) the difference between these two functions, $\pi_i^{\mathrm{J}}(n) - \pi_j^{NJ}(n)$, and additionally conditions C_2 and C_3 are drawn. By definition, at points n^*, $n^{*\mathrm{L}}$ and $n^{*\mathrm{U}}$ $\pi_i^{\mathrm{J}}(n^*) - \pi_j^{NJ}(n^*) = 0$, $\mathrm{C}_3(n^{*\mathrm{L}}) = 0$ and $\mathrm{C}_2(n^{*\mathrm{U}}) = 0$. In the example $N = 5$ has been assumed and the equilibrium coalition size is $n = N^* = 3$ which is derived as follows.

On the one hand, a non-signatory compares his or her payoff for $N^* = 3$ with those when joining the coalition and hence $N^* + 1 = 4$. The welfare implication is indicated by the solid arrow in Figure 13.1(b) from which it is evident that a non-signatory would lose. On the other hand, a signatory does not want to leave the coalition since the payoffs as a non-signatory for $N^* - 1 = 2$ are lower than those received for $N^* = 3$. The welfare implication is indicated by the dotted arrow in Figure 13.1(b). Of course, this latter welfare comparison also implies that a non-signatory has an incentive to



Note: Assumption: N = 5, $b = 1$, $c = 1$ and $d = 10$.

Figure 13.1 Equilibrium number of signatories in Barrett's model

join the coalition for $N^* - 1 = 2$ so that $N^* = 3$. To illustrate this incentive the direction of the dotted arrow would have to be reversed (not drawn).

The equilibrium number of signatories can also be directly seen in Figure 13.1(a) since conditions C_2 and C_3 are simultaneously satisfied for $n^{*L} \leq N^* \leq n^{*U}$ where the superscript L stands for lower bound and U for upper bound. It also evident from Definition 13.2 that C_2 and C_3 are similar functions where $n^{*L} + 1 = n^{*U}$. Hence if n^{*L} and n^{*U} are integer values, the

equilibrium coalition size is not unique and $N^* = \{n^{*L}, n^{*U}\}$. However, since $N^* = n^{*U}$ delivers a higher payoff to signatories and non-signatories than $N^* = n^{*L}$ one would expect that in this case $N^* = n^{*U}$.[16] One can show that $n^* < n^{*L} < N^* \leq n^{*U}$ so that n^* is a lower bound for N^*. Whereas it is easy to compute n^*, n^{*L} and n^{*U} N^* can only be determined for a particular parameter set or for large N, that is, for $N \to \infty$ at the limit.

Irrespective of the number of countries, the following general conclusions can be drawn:

$$\Sigma e_k^N > \Sigma e_k^* \geq \Sigma e_k^S,\ e_i^N > e_i^{J*},\ e_j^N < e_j^{NJ*},\ \Sigma \pi_k^N < \Sigma \pi_k^* \leq \Sigma \pi_k^S,\ \pi_i^N < \pi_i^{J*} < \pi_j^{NJ*},$$
$$N^* \in [2, N],\ \frac{\partial \xi^*}{\partial N} < 0,\ \frac{\partial \xi^*}{\partial \gamma} < 0 \text{ where } \xi^* = \frac{n^*}{N}. \tag{13.20}$$

That is, in the coalition equilibrium global emissions are lower and global welfare is higher than in the NE. Each country is better off than in the status quo and therefore the profitability condition C_1 holds.[17] However, non-signatories' welfare is higher than that of signatories since they benefit from the increased abatement efforts of signatories (in the form of lower damage), but do not have to contribute to the success of the coalition. In fact, the abatement efforts of signatories are partially offset by non-signatories, which reflects the typical leakage effect (see the last paragraph in Sub-section 13.2.2 and note 11). The minimum number of signatories is 2 and the maximum number is N. The lower bound of signatories expressed as the fraction of participants, $\xi^* = n^*/N$, decreases with the number of countries suffering from the externality, N, and with the benefit–cost ratio, $\gamma = b/c$.

In the case of large N, the following additional results can be obtained:

$$\frac{\partial \mu^*}{\partial \gamma} < 0,\ \frac{\partial I_3}{\partial \gamma} < 0,\ \frac{\partial I_4}{\partial \gamma} < 0 \tag{13.21}$$

where μ^*, I_3 and I_4 have been defined in (13.12). That is, there is an inverse relationship between the degree of participation and the effectiveness of an IEA. Thus, paradoxically, a low value of γ implies that the externality problem is particularly pronounced (see (13.4) and (13.5)); however, the impact of the coalition on global emissions and welfare is only marginal, though relatively many countries sign an agreement. From Section 11.6 it is known that the problem of the depletion of the ozone layer can approximately be described by a low value of γ, whereas for the problem of global warming a higher value of γ can be expected. This may explain why many countries have signed the Montreal Protocol, whereas in the case of global warming no IEA has been signed that imposes binding abatement obligations on signatories.[18] However, the results also suggest that despite the

high participation rate in the Montreal Protocol, one cannot expect that much has been achieved from a global point of view compared to the full cooperative outcome (social optimum).

As mentioned above, in the case of small N the relations in (13.21) can no longer be established analytically. However, a large set of simulations suggest that (13.21) also holds in the case of small N.

13.2.4 Summary of the Results in the Case of Symmetric Countries

In this section we state the results obtained so far in the form of propositions and briefly comment on them.

Proposition 13.1
Under the Nash–Cournot and the Stackelberg assumptions the equilibrium number of signatories is $N^* = 2$ or $N^* = 3$ for payoff functions of type 1 ((13.1)). No stable coalition exists for type 2 ((13.2)) under the Nash–Cournot assumption and the grand coalition forms under the Stackelberg assumption, that is, $N^* = N$. For the payoff function of type 3 ((13.3)) the equilibrium number of signatories is $N^* = 2$ under the Nash–Cournot assumption and $N^* \in [2, N]$ under the Stackelberg assumption.

Proposition 13.2
Let $\mu^* = N^*/N$ be the degree of participation in an IEA, $\gamma = b/c$ the benefit–cost ratio from emissions and $I_3 := (\Sigma e_i^* - \Sigma e_i^S)/\Sigma e_i^S$ and $I_4 := (\Sigma \pi_i^S - \Sigma \pi_i^*)/\Sigma \pi_i^S$ indices measuring the degree of externality in the coalition equilibrium where Σe_i^* denotes aggregate emissions and $\Sigma \pi_i^*$ aggregate payoffs of signatories and non-signatories in equilibrium, then for the payoff functions (13.1)–(13.3), $\partial \mu^*/\partial N < 0$, $\partial I_3/\partial N > 0$, $\partial I_4/\partial N > 0$, $\partial I_3/\partial \gamma < 0$ and $\partial I_4/\partial \gamma < 0$ hold under the Nash–Cournot assumption, except $\partial I_3/\partial \gamma = 0$ and $\partial I_4/\partial \gamma = 0$ for payoff function (13.2). For payoff functions (13.1) and (13.3) these relations also hold under the Stackelberg assumption. For payoff function (13.2) the degree of externality is zero.

Proposition 13.3
For the payoff function of type 3 ((13.3)), assuming Stackelberg leadership of signatories, the percentage of participation in the stable coalition, μ^*, is a declining function of the benefit–cost ratio from emissions, γ, that is, $\partial \mu^*/\partial \gamma < 0$.

Propositions 13.1 and 13.3 suggest that assuming quadratic benefit and damage cost functions (satisfying assumptions A_1 in (9.2)) and signatories behaving as Stackelberg leaders allows us to model a broad spectrum of IEAs of different coalition sizes.[19] Another advantage of Barrett's model is that it allows us to relate the equilibrium number of signatories to the benefit–cost ratio from emissions and the number of countries suffering from the externality problem. Hence, the explanatory power of this model version is higher than that of the other versions discussed above. Of course, the assumption of Stackelberg leadership on the side of the signatories is a rather special assumption but, as argued above, there are arguments which can be put forward in defense.

To this end it is also evident that Carraro and Siniscalco's claim that orthogonal reaction functions would be conducive to cooperation applies only to their model version (Nash–Cournot assumption). As demonstrated, negatively sloped reaction functions are in fact a prerequisite for Stackelberg leadership on the side of the signatories to be used strategically so that a higher participation in an IEA than $N^* = 3$ may be obtained. In fact, for a payoff function of type 2 with slope of -1, $N^* = N$. This result also contrasts with that obtained in Chapters 10 and 11. In the coalition context, strategic Stackelberg leadership can be conducive to cooperation, increasing global welfare and reducing aggregate emissions compared to the NE. This is so since signatories as Stackelberg leaders do not reduce emissions in a 'naive' way but are aware of the leakage effects caused by non-signatories.

It is interesting that for all model versions – except for the Stackelberg assumption and payoff functions of type 2 – the same qualitative results could be derived, namely that whenever cooperation is needed most from a global point of view a coalition achieves only little (Proposition 13.2). In the case of Barrett's model, this part of the analysis had to rely partially on simulations (if N is small), though for the other model versions this was shown to hold in general. It turns out that this result carries over to the dynamic coalition model in Chapter 14. Roughly speaking, this is due to the fact that whenever the difference between NE and social optimum is large, the free-rider incentive is also very strong and hence a stable coalition can only achieve a moderate abatement target.

13.2.5 Heterogeneous Countries

In Sub-section 13.2.1 we argued that for *symmetric countries* the welfare distribution within the coalition is more or less trivially solved: each country receives the same payoff and no reallocation of payoffs is necessary. It is obvious that any asymmetric distribution would result in a smaller or

at best the same size of the coalition. For instance, recall that $N^* = 2$ or $N^* = 3$ for the payoff function of type 1. Suppose that $N^* = 3$ and that one country, say 1, receives more than an equal share of the gains from co-operation, which of course implies that another country, say 2, must receive less of the pie. Since, for $N^* = 3$, a signatory is indifferent between remaining in the coalition or leaving it, country 2 would leave the coalition if such a reallocation took place.

It also seems obvious that if countries are symmetric they choose that abatement level which maximizes joint welfare. Any other target would reduce welfare to a signatory and cannot secure a higher participation in an IEA.

In the case of asymmetric countries the implications of the welfare allocation rule and the choice of the abatement target(s) on the equilibrium coalition is less obvious. That is, both issues, which may be summarized as the design of an IEA, affect the participation rate in an IEA and the composition of its members and hence, ultimately, affect the success of an agreement. Of course, if unlimited transfers are possible, the highest participation and global welfare is obtained if signatories maximize the coalition's welfare and hence only the relation between the welfare allocation rule and the success of an agreement must be investigated. A similar simplifying assumption (for which a justification was given in Chapter 8) is to rule out any transfers, so that only the relation between abatement target(s) and the success of an IEA has to be analyzed.

So far the question of how the design of an agreement affects its success has received only limited attention. On the one hand, the importance of this issue has not been recognized in the literature to date; on the other hand, a systematic analysis of this issue is difficult. The main obstacle to obtaining more general results is, apart from the numerous possible designs of IEAs which one would have to investigate, that results crucially depend on the type and degree of asymmetry between countries.

In the following we briefly summarize the findings of four papers. It is important to stress that in these papers transfers are restricted to taking place *within* the group of signatories. The possibility that signatories 'bribe' non-signatories via transfers to participate in their agreement or the possibility of transfers from non-signatories to signatories to induce higher abatement efforts of signatories ('abatement leasing') are neglected. This issue will be taken up in Sub-section 13.2.6.

Bauer (1992) assumes the following payoff function:

$$\pi_i = i \sum_{j=1}^{N} r_j - \frac{r_i^2}{i} \tag{13.22}$$

where r_i denotes emission reductions and i is a country-specific index, $i = 1, 2, \ldots, N$. The first term in (13.22) represents (linear) benefits from emission

reductions, the second term (quadratic) costs from abatement. Thus, payoff functions (13.22) are of *type 1*, that is, reaction functions are orthogonal. Consequently, a distinction between the Nash–Cournot and the Stackelberg assumptions is not necessary.

The idea behind the specification in (13.22) is that the index i represents differences in the size of countries. Accordingly, large countries benefit proportionally more from emission reductions since a larger area is affected by environmental damages. There are economies of scale from abatement so that it is less costly for larger countries to reduce emissions than for smaller countries. Taken together, the function in (13.22) implies that a country that is double the size of a smaller country can reduce twice as much emissions at the same marginal costs as the smaller country.

Bauer finds that only coalitions of two countries of similar size are stable; that is, countries 1 and 2, 3 and 4, or 2 and 3, 4 and 5 and so on may form stable coalitions. Countries of different size with rank differences larger than 1, for example, countries 1 and 3, will never form stable coalitions. It is particularly interesting that in Bauer's model several coalitions may coexist at the same time. However, coalitions comprising more than two countries are only stable if coalitions of two countries act as single countries. This is, of course, an unrealistic assumption which is not in line with the notion of self-enforceability.

What is crucial in driving Bauer's result is the assumption that signatories choose that reduction level which maximizes aggregate welfare and that transfers within the coalition are ruled out. Thus, Bauer does not pursue the interesting question of whether there is a transfer scheme which could balance the asymmetries between countries so as to induce a higher participation rate. She also does not investigate whether the choice of another abatement target would lead to more optimistic results (given the assumption of no transfers). Since the abatement target which maximizes joint welfare imposes a particular strain on the smaller partner in the coalition, one should expect that either a lower abatement target or an asymmetric abatement allocation would achieve more.

Hoel (1992) also assumes a payoff function of *type 1* defined in terms of emissions. He assumes the same benefit function from emissions for all countries, but different damages. The constant marginal damage parameter is presumed to follow a uniform discrete distribution. Hoel considers several possibilities of how the emission level of a grand coalition is chosen assuming no transfers. Among those are: (a) socially optimal emission allocation (unconstrained optimization); (b) socially optimal emission allocation subject to the constraint that no country is worse off than in the status quo; and (c) uniform emission reduction according to the median country's proposal. Though the various solutions lead to different global welfare

implications when abstracting from stability considerations, Hoel reports that a stability check reveals that not more than two countries would form a stable coalition. Thus, as in Bauer's model, in the case of asymmetric countries, the maximum number of signatories may fall short of the maximum number in the case of symmetric countries (where $N^* = 3$). This is different in the next model.

Botteon and Carraro (1997) use estimates from Musgrave (1994) of five world regions for abatement and damage cost functions of *type 1* and conduct a stability analysis allowing for transfers. Among signatories payoffs are distributed either according to the Shapley value or the Nash bargaining solution.[20] They confirm the result of the symmetric country case, namely that the maximum number of signatories is $N^* = 3$. Moreover, they demonstrate that there is not only one single set of potential signatories but that various combinations of players could form a stable coalition. They find that membership in the coalition depends on the burden-sharing rule.

For their data set the Shapley value leads to a coalition generating a higher global welfare than a coalition in which the Nash bargaining solution is applied. This finding supports our conjecture mentioned above, that the design of an agreement has an impact on global efficiency.

Barrett (1997b) proceeds in two steps. In the first step he considers payoff functions of *type 1*, in the second step those of *type 3* where he defines these functions in reduction space. For both types he considers two model versions, one without and one with side-payments. He assumes that signatories maximize the aggregate payoff to the coalition. Barrett allows for different marginal benefit *and* damage cost parameters between countries (though the functional form of the net benefit functions is the same for all countries) and thus the possible degree of asymmetry is relatively large.

Payoff Functions of Type 1: Barrett shows that *without transfers* the maximum number of signatories is three, but also that no coalition may exist at all, depending on the degree of asymmetry. *With transfers*, employing the Shapley value, he finds that the minimum number of signatories is two, but the maximum number does not exceed $N^* = 3$. Though the results claim no general relevance, it seems very likely that the finding of the symmetric country case of a maximum coalition of three countries carries over to heterogeneous countries.

Payoff Functions of Type 3: Barrett assumes that countries belong to one of two groups of countries with homogeneous group characteristics. For both assumptions – no transfers and transfers – he basically finds his results of the symmetric country case confirmed (by using simulations). That is, $N^* \in [2, N]$, and whenever cooperation would be needed most, a coalition achieves only little.

Thus summarizing, the few attempts to model asymmetric countries suggest that the basic results obtained in the case of symmetric countries may carry over to the more general setting of heterogeneous countries. However, conclusions can only claim preliminary status since a systematic analysis is missing. In particular, the question of how the welfare allocation rule and/or the choice of the abatement target within the coalition affects the success of an IEA has not been given sufficient attention so far and awaits an intensive treatment in future research.

13.2.6 Model Extensions: Transfers between Insiders and Outsiders, Minimum Commitment, Fairness and Economies of Scale

In this section we want to discuss briefly some model extensions and aspects which have not been covered yet.

Transfers and minimum commitment

Though we have indicated our reservations about the assumption of transfers in Chapter 8, we should like to discuss briefly some results obtained by Botteon and Carraro (1997, 1998) and Carraro and Siniscalco (1993). With respect to symmetric countries, the following result can be derived according to Carraro and Siniscalco (1993, pp. 315ff.):

Proposition 13.4

If countries are symmetric it is not possible to enlarge an internally and externally stable coalition by self-financed transfers.

Proof: First note that *self-financed transfers* refer to transfers which are paid out of the welfare gain that signatories obtain by forming a larger coalition compared to the status quo. For signatories it is only attractive to pay transfers to a non-signatory to induce it to join if the additional gain is larger than the transfer. Furthermore, a non-signatory only joins the coalition if the transfer which it receives, t, is larger than the welfare loss it occurs from becoming a signatory. If we denote the status quo coalition size by N^*, then the enlarged coalition size is N^*+1 and the two conditions taken together are:

$$N^* \cdot [\pi_i^{\mathrm{J}}(N^*+1) - \pi_i^{\mathrm{J}}(N^*)] \geq t \geq \pi_j^{NJ}(N^*) - \pi_j^{\mathrm{J}}(N^*+1). \qquad (13.23)$$

However, from the definition of stability and the assumption that a transfer is needed to induce a non-signatory to join the coalition (the inequality sign of the external stability condition C_3 is strict!) it follows that $\pi_j^{NJ}(N^*) < \pi_j^{\mathrm{J}}(N^*+1)$ and hence compensating a non-signatory implies an overall gain of $[\pi_i^{\mathrm{J}}(N^*+1) - \pi_i^{\mathrm{J}}(N^*)] - t/N^*$ to a signatory. However,

alternatively a signatory could leave the coalition to gain $\pi_j^{NJ}(N^*) - \pi_j^{J}(N^* + 1)$ which by the definition of instability of $N^* + 1$ is preferable. QED

Due to this negative result, Carraro and Siniscalco (1993) go on to analyze whether the coalition can be expanded if there is a minimum commitment by some countries to cooperation (see also Petrakis and Xepapadeas 1996). There are two possibilities: first, there may be a commitment on the side of the signatories where some countries try to expand the coalition via transfers; second, there is the possibility that some outsiders are committed to pay transfers to signatories in order to benefit from increased abatement efforts and save abatement costs (abatement leasing). Carraro and Siniscalco show that in both cases more positive results may be obtained, depending on how many countries are committed.

Apart from the fact that this finding is hardly surprising, the basic problem with minimum commitment is that it is nothing other than assuming that some countries play an *unconditional cooperative* strategy. From Chapter 3 we know that such a strategy can *never* be an NE and hence, as Carraro and Siniscalco themselves note, such an assumption violates the notion of self-enforceability.[21]

A more interesting result is obtained by the study of Botteon and Carraro (1997) with *heterogeneous countries* and simultaneous moves of signatories and non-signatories. The authors show for their example that *without* commitment transfers can be used by signatories to expand their coalition. Of particular interest is the finding that for their data set such an expansion is possible if the gains of the coalition are distributed according to the Shapley value, but not if the Nash bargaining rule is applied. Once more, this stresses the importance of the design of an agreement for the stability and global efficiency of a coalition.

Fairness

There are two papers which have taken another route to generate more positive results with respect to coalition formation which are very much in the tradition of Sudgen's theory of reciprocity and Margolis's unselfish behavior (see Section 10.7). Jeppesen and Andersen (1998) introduce 'non-material' payoffs in Barrett's model which are derived from *fairness*. They argue that free-riding would be seen as unfair behavior by governments. Therefore, governments derive some payoff from the membership as such. Similarly, in Hoel and Schneider's (1997) model governments receive disutility from breaking an agreement (loss of reputation). Not surprisingly, both papers find that such moral aspects increase the number of participating countries in an IEA. However, it is evident that such arguments are not

compatible with the fundamental assumption of economic theory. What makes Hoel and Schneider's analysis interesting is their finding that transfers from signatories to outsiders may decrease the number of participants in an IEA. In particular, counterintuitively, the aggregate effect of transfers may be negative. Though non-signatories do more when receiving transfers, signatories do less ('abatement leasing') and overall, if N is large, emissions may increase.

Economies of scale

Heal (1994) views coalition formation as being similar to an assurance game (see Sections 3.4 and 3.6) where cooperation pays once enough players accede to an IEA. He argues that there must be a minimum number of countries in order for a coalition to be profitable and to come into force. He calls this a *minimum critical coalition*. Most of Heal's analysis is concerned with the derivation of the conditions of a minimum critical coalition. However, basically, transforming his approach into the context of the previous setting, it turns out that his definition of a minimum critical coalition is nothing other than the definition of the profitability of a coalition. Since it is known from above that profitability is only a necessary but by no means a sufficient condition for the stability of a coalition, Heal misses an important point. Hence, the really interesting aspect of Heal's analysis lies somewhere else: it is the idea that there might be economies of scale in forming a coalition. For instance, sharing the cost of R&D in developing new abatement technology may reduce costs substantially so as to induce more countries to join an IEA. Thus, integrating this idea into the conjectural variation models above would lead to more optimistic results.

13.2.7 Issue Linkage

Since the conjectural variation models are technically rather simple, they are ideally suited to analyzing the effects of linking an environmental agreement with another agreement. Since the instability of an IEA is caused by the *non-excludability* of the *public good* environment, the second agreement must deal with the provision of an *excludable good*, as for instance a club good (Carraro 1997).[22] For instance, Carraro and Siniscalco (1997), Botteon and Carraro (1998) and Katsoulacos (1997) consider the link between an IEA and cooperation on R&D. Signatories of an IEA share the costs of R&D and can exploit more-efficient technologies. Hence, the unit production costs of the signatories decrease through cooperation and they obtain a competitive advantage over non-signatories. Thus, if a signatory thinks about leaving the coalition it has to expect not only that abatement

of the remaining signatories will be lower after they have reoptimized their strategies, but also that it will have higher production costs. Thus, the forces inducing a country to remain in the coalition are stronger through issue linkage and therefore, not surprisingly, these papers find (using simulations) that the equilibrium number of signatories may be greater in the linked agreement than in the pure IEA.

Since Carraro and Siniscalco (1997) assume that there are decreasing economies of scale in cooperation on R&D, the equilibrium number of signatories may fall short of the grand coalition. Moreover, as the authors assume an oligopolistic market structure of the Cournot type, the gains from cooperation on R&D are limited. Since lower unit production costs imply a higher output by signatories and hence a lower market price, it might be in the interest of signatories not to have too many coalition members. In fact, for some parameters of the market demand and production function, it may happen that the linked agreement comprises fewer signatories than the original IEA (Carraro and Siniscalco 1997). Thus, issue linkage may also destabilize an IEA; this would be even more pronounced if account were taken of transaction costs from coordinating two agreements.

In a similar spirit Barrett (1997c) sets up a model in which an IEA is linked to a trade agreement. Trade of signatories with outsiders is completely banned, but outsiders may trade among themselves. Barrett shows (using simulations) that if minimum participation is ensured in the linked agreement, the agreement may be capable of securing a large number of signatories. An example of such a link is the Montreal Protocol on the depletion of the ozone layer where trade with non-signatories is partially restricted (Barratt-Brown 1991; Benedick 1991; Blackhurst and Subramanian 1992).

The problem with all the issue linkage models mentioned above is that they make rather special assumptions regarding the market structure. In particular, Barrett's (positive) result rests crucially on the assumption of an oligopolistic market structure. Assuming instead perfect competition, a 'trade cartel' would be of no interest to signatories. Moreover, the assumption in Barrett's model that imports of non-signatories to signatories' markets are completely cut off seems also a very restrictive assumption and is not compatible with the rules of GATT/WTO. Probably, it would be more realistic to assume only a tariff imposed on non-signatories (see, for example, Finus and Rundshagen 1999). In reality it is also hardly conceivable that all countries belonging to a large trade agreement would automatically be required to become members of an IEA. However, admittedly, allowing for mixed membership creates a lot of analytical problems. Finally, we should like to point out that Barrett implicitly assumes fixed

plant location, which makes it easier to mitigate free-riding. A more realistic and interesting approach would account for the possibility of endogenous plant location since firms can freely move their location if environmental standards become too costly for them (see Finus and Rundshagen 1999 and the literature cited therein).

13.2.8 Discussion

In this sub-section we want to discuss some *general* aspects of the conjectural variation models to the extent that they have not yet been mentioned under a particular heading. Most important is the *inconsistency of the assumptions* underlying the definition of internal and external stability as given in Definition 13.2. Whereas with respect to emissions and transfers an *instant reaction* is assumed, with respect to the number of signatories and non-signatories it is assumed that all players except the defector stick to their participation strategies. Whereas the first assumption is in the tradition of the conjectural variation models encountered in Chapter 10, the second assumption is in the tradition of the incentive structure of the stage game of a global emission game as laid out in Chapter 9. Thus, strictly speaking, all models mentioned above are not pure conjectural variation models. There are three possibilities to base the equilibrium concept on more consistent assumptions.

The first possibility would be to define internal stability and external stability in the tradition of a *static NE*. Hence, $C_2 := \pi_i^{\mathrm{J}}(N^*, e^*, t^*) - \pi_i^{NJ}(N^* - 1, e_i', e_{-i}^*, t_i', t_{-i}^*) \geq 0 \;\; \forall \;\; i \in \mathrm{I}^{\mathrm{J}}$ and $C_3 := \pi_j^{NJ}(N^*, e^*, t^*) - \pi_j^{\mathrm{J}}(N^* + 1, e_j', e_{-j}^*, t_j', t_{-j}^*) \geq 0 \; \forall \, j \in \mathrm{I}^{NJ}$. That is, only deviations by a *single player* are considered; all other strategies are assumed to remain constant. However, it is easily confirmed that for the payoff functions of types 1, 2 and 3 the condition for internal stability would always fail to hold. That is, there is always a free-rider incentive and therefore the incentive structure is similar to that in a *PD game*.

Hence, the second possibility would be to define stability within an infinitely repeated game framework specifying credible threats to punish defection as laid out in previous chapters (see in particular Chapter 12). This route will be taken in Chapter 14.

The third possibility would assume instant reactions with respect to *all* strategies: also with respect to the number of participants in an IEA. That is, a consistent conjectural variation equilibrium would be defined by $C_2 := \pi_i^{\mathrm{J}}(N^*, e^*, t^*) - \pi_i^{NJ}(N^{*\prime}, e^{*\prime}, t^{*\prime}) \geq 0 \;\; \forall \;\; i \in \mathrm{I}^{\mathrm{J}}$ and $C_3 := \pi_j^{NJ}(N^*, e^*, t^*) - \pi_j^{\mathrm{J}}(N^{*\prime}, e^{*\prime}, t^{*\prime}) \geq 0 \; \forall \, j \in \mathrm{I}^{NJ}$. Hence, if a signatory leaves the coalition it cannot exclusively assume that $N^* - 1$ signatories and $N - (N^* - 1)$ non-signatories will reoptimize their strategies but must reckon that an

alternative coalition could form. Thus the myopic view where a player considers only the immediate reaction to his/her deviation is given up in favor of a farsighted view in which all subsequent actions of players are accounted for (Carraro and Moriconi 1997).

In the context of *symmetric countries* this definition implies that the *former* equilibrium N^* according to Definitions 13.1 and 13.2 is no longer an equilibrium. Since it is always beneficial to be a non-signatory if a coalition exists, a signatory has an incentive to leave the coalition, knowing that it will be replaced by a former non-signatory.[23] Hence, the number of signatories would permanently remain at N^*; however, C_2 does not hold and membership permanently circulates. Consequently, in a strict sense, the former equilibrium N^* is no equilibrium any more (since C_2 and C_3 are not simultaneously satisfied), though it may be called a 'steady state'. Thus, also the threat to resolve a coalition of N^* once a signatory leaves is an empty threat since by the definition of a profitable coalition there is an incentive to form a coalition.[24] Only in the context of heterogeneous countries may a stable coalition at N^* exist. If the new coalition reoptimizes its strategies to a lower abatement level after a signatory has left, then the defector may lose. This is because if countries are heterogeneous the abatement target of the coalition depends on the membership and some countries may have a lower payoff as a non-signatory than as a signatory.

However, even in the case of symmetric countries an equilibrium coalition may exist, though not at (the former) N^*. We denote this 'new' equilibrium by N^{*C}. To see this, consider Matrix 13.1 which assumes payoff function of type 3 and Nash–Cournot behavior between signatories and non-signatories. The rows are ordered according to the number of signatories. The first entries are the payoffs of non-signatories, the subsequent entries those of signatories. From previous sections it is known that $N^* = 2$. As argued above, $N^* = 2$ is only a steady state according to the modified stability definition since a signatory leaving the coalition knows that it will be replaced by a non-signatory and hence can enjoy a payoff of $\pi_i^{NJ} = 0.19$. Now consider a coalition of four countries. If a signatory leaves, it must conjecture that in the new coalition of three countries, again, a country leaves. This leads to the steady-state coalition of two countries. Comparing the payoff as a signatory in a coalition of four countries ($\pi_i^J = 0.43$) with that as a non-signatory, if a coalition of two countries forms ($\pi_i^{NJ} = 0.19$) does not make it attractive to leave the coalition. Hence, according to the modified stability definition above, the coalition of four countries is an equilibrium coalition, that is, $N^{*C} = 4$. Consequently, the grand coalition cannot be stable since it pays a country to be a non-signatory in a coalition of four countries instead of being a signatory in the grand coalition ($\pi_i^J = 1.14 < 2.49 = \pi_i^{NJ}$).

Matrix 13.1 Consistent conjectural variation equilibrium[a]

Number of signatories	Payoffs				
1 (singletons)	−0.9	−0.9	−0.9	−0.9	−0.9
$N^* = 2$	0.19	0.19	0.19	−0.84	−0.84
3	1.5	1.5	−0.3	−0.3	−0.3
$N^{*C} = 4$	**2.49**	**0.43**	**0.43**	**0.43**	**0.43**
5	1.14	1.14	1.14	1.14	1.14

Note: [a]Payoff function (13.3), $b = 2$, $c = 1$, $d = 2$, $N = 5$ has been assumed.

It is interesting to note that in other examples (for example, where the number of countries is larger), several equilibrium coalitions may coexist. However, since the payoff vector of larger coalitions (strictly) Pareto-dominates smaller coalitions, it should be expected that the largest coalition will be the final equilibrium.

With the suggested modifications, reconsidering the problem and looking at it from another angle, one may very well argue that the merit of the conjectural variation models is, apart from their simplicity, that they explicitly limit punishment options to an (instant) reoptimization strategy. That is, these models take into account the actual limit punishment options in international politics (see Sub-section 12.2.4). The above-described reoptimization strategy would be rational and an SPE in the sense of a conjectural variation equilibrium since it takes full account of all subsequent moves (see Section 10.4) and conjectures are actually confirmed in equilibrium. Due to these advantages, we shall present in Chapter 15 a *farsighted coalition formation game* which is based on the modified conjectural variation approach presented above. This model will include two additional refinements: it allows for the possibility that *several coalitions coexist* and the possibility of *exclusive membership*.

A more general question with respect to the assumption of instant reactions is whether it is appropriate to account for the incentive structure in international pollution control (see the discussion in Chapter 10). It seems 'natural' to expect that, though a country may decide to become a signatory and agrees with abatement obligations of the coalition, it may later breach the agreement to obtain a *transitory gain* if this is not punished. However, this conjecture is not shared by Barrett, who argues that *once a*

country has decided to be a signatory it would comply with the obligations of the coalition.

There are two problems with this statement: first, on theoretical grounds, it is not satisfactory to assume commitment by signatories in an *ad hoc* manner; second, on empirical grounds the claim can hardly be sustained if confronted with reality. Though Barrett (1997a) tries hard to make the point that countries only sign treaties which they intend to fulfill by quoting the political and institutional science literature (for example, Chayes and Chayes 1993, 1995), there have been many instances where non-compliance has occurred.[25] It is particularly surprising that Barrett defends his approach by quoting such superficial work for two reasons. First, Chayes and Chayes interpret cases of non-compliance as 'capacity problems' (see, in particular, Chayes and Chayes 1995, pp. 13ff.). That is, countries do not free-ride because of self-interest but because of temporary lack of resources to comply. Thus, it is hardly surprising that they find evidence backing their hypothesis that countries have a propensity to fulfill treaties. Second, testing the hypothesis of compliance empirically, one has to take into account that the obligations of many treaties do not go beyond what would be implied by the NE anyway.[26] Thus from compliance as such one cannot infer the efficiency of a treaty.

Taken together, the point raised in the last paragraph comes down to the question already raised in Chapter 10: *whether a dynamic process would not be better modeled as a 'true' dynamic game, rather than as a single stage game* as in the conjectural variation framework. This question also seems relevant in the context of coalition formation.

Finally, note three minor points with respect to the models discussed in the previous sections. First, under the Nash–Cournot assumption all countries, and under the Stackelberg assumption non-signatories, maximize payoffs according to a zero conjecture with respect to emissions which, considering their reaction functions, is not confirmed if Definition 10.1 of a *rational* conjectural variation equilibrium is applied. Again, this stresses that Definition 13.2 of a stable coalition is a hybrid of a conjectural variation and a static or repeated game equilibrium.

Second, it is not compelling why a coalition equilibrium should be defined in terms of external stability and not only with respect to internal stability. Why cannot a coalition defend its coalition against non-signatories intending to join? This point is particularly relevant in issue linkage games where, above a certain number of signatories, further accession may decrease the 'old' signatories' welfare accruing from the club-good part of the agreement. Therefore, it may be better to restrict membership so as to stabilize the linked agreement. In this light, Carraro and Siniscalco's (1997) finding that a linked agreement may lead to a

smaller coalition size than in the isolated IEA would have to be reconsidered.

Third, under the Stackelberg assumption it is assumed that a signatory which possesses superior information as a Stackelberg leader maximizes its payoff according to Nash–Cournot once it leaves the coalition. This assumption obviously requires some form of irrationality by players and is difficult to justify. It is this very assumption which leads, for instance, to the grand coalition for a payoff function of type 2 under the Stackelberg assumption. Assuming instead for the preliminary coalition of $N^* = N$ that a signatory is as clever when leaving the coalition as when remaining in the coalition, no stable coalition would exist.

13.3 THE CORE

13.3.1 The Concept

Though the concept of the *core* belongs to *cooperative game theory*, it is discussed here since it represents an important branch of the literature on coalition formation. As laid out in Section 2.3, a typical feature of cooperative game theory is the assumption that legally binding contracts can be signed. Consequently, this concept seems inadequate in the context of international pollution control. However, rejecting this concept on this ground would be too simple. In fact, a closer look at the concept reveals that some form of implicit punishment is indeed part of the definition of the core.

There are basically two definitions of the core, one assuming *no transfers,* the other assuming *unlimited transfers* among the coalition members. Since the latter assumption has exclusively been made in all papers on international pollution control until now, we shall restrict our attention to this assumption as well (though we have raised objections to this assumption in Chapter 8). Before defining the core, we need two definitions (Luce and Raiffa 1957; Moulin 1995; Ordeshook 1992):

Definition 13.3: Characteristic function
The characteristic function provides for each possible coalition $\mathrm{I}^{\mathrm{J}} \subseteq \mathrm{I}$ the information about the highest obtainable aggregate payoff which is called the worth of a coalition, that is:

$$w(\mathrm{I}^{\mathrm{J}}) = \max_{e^{\mathrm{J}}} \sum_{i \in \mathrm{I}^{\mathrm{J}}} \pi_i$$

where e^{J} is the emission vector of the coalition and $\pi_i = \beta_i(e_i) - \phi_i(\Sigma e_k)$ the payoff function of a player.

Definition 13.4: Imputation
An imputation $\pi^{*\psi} = (\pi_1^{*\psi}, ..., \pi_N^{*\psi})$ is a payoff vector which allocates the payoffs of a socially optimal solution to each member of the coalition according to some particular rule.

With the help of Definitions 13.3 and 13.4, it is straightforward to define the core of an economy (see Foley 1970):

Definition 13.5: Core
An imputation is in the core with respect to the endowments of the economy if it cannot be blocked by any non-empty coalition. No other coalition can do better for all its members with its own resources. That is:

$$\sum_{i \in \mathrm{I^J}} \pi_i^{*\psi} \geq w(\mathrm{I^J}) \;\forall\; \mathrm{I^J} \subseteq \mathrm{I}.$$

From Definition 13.5 it is evident that the core depends on the *worth* of the coalition, $w(\mathrm{I^J})$, which in turn depends on the behavior of the players outside the coalition.[27] A frequently made assumption is that the non-coalition members put the coalition members down to their minimax or maximin payoffs. This leads to the α *and* β *characteristic function*:

$$w^{\alpha}(\mathrm{I^J}) = \min_{e^{NJ}} \max_{e^{\mathrm{J}}} \sum_{i \in \mathrm{I^J}} \pi_i(e^{\mathrm{J}}, e^{NJ}) \tag{13.24}$$

$$w^{\beta}(\mathrm{I^{IJ}}) = \max_{e^{\mathrm{J}}} \min_{e^{NJ}} \sum_{i \in \mathrm{I^J}} \pi_i(e^{\mathrm{J}}, e^{NJ}) \tag{13.25}$$

$w^{\alpha}(\mathrm{I^J})$ implies that the non-coalition members move first and choose their highest emission levels, $e_j^{\max}$, to which the coalition members react by choosing their best replies, $e_i(\Sigma e_j^{\max})$. $w^{\beta}(\mathrm{I^J})$ implies that coalition members move first and choose their best replies, assuming that the non-members will punish them as hard as possible. In the global emission game context both characteristic functions are identical, that is, $w^{\alpha}(\mathrm{I^J}) = w^{\beta}(\mathrm{I^J})$ (since in both cases $e_j = e_j^{\max}$ $\forall$ $j \in \mathrm{I}^{NJ}$).[28] Of course, if $\mathrm{I^J} = \mathrm{I}$, $w^{\alpha}(\{\mathrm{I}\}) = w^{\beta}(\{\mathrm{I}\}) = \Sigma\pi_i^{\mathrm{S}}$.

In the international environmental context the assumption about the behavior of non-members in the α and β characteristic functions seems not very realistic. It implies that the coalition members assume that players outside the coalition will carry out punishments which are injurious to their own welfare. Therefore, Chander and Tulkens (1995, 1997) propose the assumption that non-coalition members choose their strategy according to Nash behavior. If the coalition maximizes the aggregate welfare of its

members, this leads to a 'partial-agreement Nash equilibrium', abbreviated to PANE:[29]

Definition 13.6: Partial-agreement Nash equilibrium (PANE)
Given a coalition $I^J \subseteq I$, a partial-agreement equilibrium with respect to I^J is strategy vector $\tilde{e} = (\tilde{e}^J, \tilde{e}^{NJ})$ which is characterized by:

1. $\tilde{e}^J$ maximizes $\sum_{i \in I^J} \pi_i$ for $e^{NJ} = \tilde{e}^{NJ}$;
2. $\forall j \in I^{NJ}$: $\tilde{e}_j^{NJ}$ maximizes π_j for $e^J = \tilde{e}^J$ and $e_{-j}^{NJ} = \tilde{e}_{-j}^{NJ}$

where $\tilde{e}^J$ is the emission vector of the coalition with typical element $\tilde{e}_i^J$ and $\tilde{e}^{NJ}$ is the emission vector of the non-coalition members (behaving as singletons) with typical element $\tilde{e}_j^{NJ}$.

Definition 14.6 is very similar to the equilibrium in the emission space in the conjectural variation models under the Nash–Cournot assumption, assuming that the equilibrium coalition size is given.[30] Non-members, behaving as singletons, maximize their payoff non-cooperatively whereas the coalition members maximize their joint welfare cooperatively, acting as a single player. Given this assumption, the PANE constitutes an NE. Definition 14.6 allows us to specify the γ *characteristic function*:

$$w^{\gamma}(I^J) = \Sigma \pi_i(\tilde{e}^J, \tilde{e}^{NJ}) \tag{13.26}$$

where $\tilde{e} = (\tilde{e}^J, \tilde{e}^{NJ})$ is the PANE with respect to I^J.

Taken together, the α, β or γ core are defined by Definition 13.5 where the characteristic function is given by (13.24), (13.25) and (13.26) respectively.

13.3.2 Results

For most games it is difficult to characterize the entire set of imputations which lie in the core. Therefore, all papers analyze only whether an imputation can be constructed for the socially optimal emission vector which lies in the core. That is, one checks whether an imputation of the grand coalition $\pi^{*\psi} = (\pi_1^{*\psi}, ..., \pi_N^{*\psi})$ lies in the core where $\pi_i^{*\psi} = \pi_i^S(e^S) + t_i$ and t_i is a transfer which is implied by some transfer scheme. A first indication whether this may be the case is to check whether the so-called 'individual rationality constraint' is satisfied by the imputation. This necessary condition is derived from considering that $w^{\alpha}(\{i\}) = \pi_i^M$, $w^{\beta}(\{i\}) = \pi_i^{SC}$ (where $\pi_i^M = \pi_i^{SC}$ in the global emission game) and $w^{\gamma}(\{i\}) = \pi_i^N$ hold, implying

that a player in the coalition must receive at least an individual rational payoff or a payoff above the NE payoff respectively to sign the agreement.[31] A single-player coalition, $\{i\}$, simply forms if a player leaves the grand coalition which – according to the core concept – completely splits up into singletons and those players either minimax (α core), maximin (β core) or play Nash (γ core) against the defector.[32]

Moreover, considering the fact that a player is punished very harshly if s/he leaves the grand coalition according to the α and β characteristic functions, it is not surprising that some imputations lie in the α and β core and therefore the core of most games is not empty.[33] For instance, Mäler (1989) reports for his acid rain game that all Pareto optima lie in the α core and therefore he calls this concept useless in predicting the outcome of a game. Basically, any imputation in which welfare is redistributed such that each country receives at least its minimax payoff would lie in the core. In fact, if payoff functions are not too asymmetric transfers might not even be necessary for such an imputation to lie in the core.

With respect to the γ core it is more difficult to satisfy the individual rationality constraint and therefore it does not come as a surprise that the γ core (if it is non-empty) is contained in the α core.[34] In the case of heterogeneous countries, some sophisticated allocation rule may be needed to ensure that the social optimum lies in the γ core. Chander and Tulkens suggest in various papers consideration of the following transfer scheme:

$$t_i^* = [\beta_i(e_i^N) - \beta_i(e_i^S)] - \frac{\phi_i'(e^S)}{\Sigma\phi_k'(e^S)} \cdot [\Sigma\beta_k(e_k^N) - \Sigma\beta_k(e_k^S)] \qquad (13.27)$$

where $t_i > 0$ implies that a transfer is received and $t_i < 0$ that a transfer is paid by a country. Since the definition in (13.27) implies that t_i is not a vector but a single number, it is necessary from a technical point of view to assume that all transfers are administered by an international agency. This agency either receives a transfer from country i ($t_i < 0$) or pays a transfer to country i ($t_i > 0$).

The transfer scheme comprises two parts. The first part is a payment *to* each country which covers its decrease in benefits between the NE and the social optimum (first bracket). The second part is a payment *by* each country covering the total decrease of benefits in proportion to the fraction of marginal damage in each country to aggregate marginal damage in all countries (second bracket). That is, those countries which perceive damage as more severe (have a higher preference for a clean environment) pay a larger portion of their gains which they receive from a joint and socially optimal abatement policy.

For the transfer scheme implied by (13.27) it can be shown that the following result can be established (see Chander and Tulkens 1997):[35]

Proposition 13.5
Let the payoff function from global emissions be given by $\pi_i = \beta_i(e_i) - \phi_i(\Sigma e_k)$, assumptions A_1 in (9.2) hold and let $e^S = (e_1^S, ..., e_N^S)$ be the (unique) socially optimal emission vector. Then the imputation $\pi^{*\psi} = (\pi_1^{*\psi}, ..., \pi_N^{*\psi})$ defined by $\pi_i^{*\psi} = \pi_i^S(e^S) + t_i^*$ where t_i^* is given by (13.27) lies in the core provided one of the following assumptions is true: (a) damage functions are linear; (b) $\forall\, I^J \subset I, |I^J| \geq 2, \Sigma_{k \in I^J} \phi_k'(e^S) \geq \phi_i'(e^N)$ $\forall\, i \in I^J$; and (c) countries are symmetric.

Proof: See Appendix X.2. QED

Assumption (a) implies orthogonal reaction functions and thus non-signatories choose their emission level independently of the coalition. *Assumption (b)* basically implies that for coalition members marginal damages do not fall too much when moving from the social optimum to the NE.[36] The finding that $\pi_i^{*\psi} = \pi_i^S(e^S) + t_i^*$ lies in the core if *assumption (c)* holds is anything but surprising. Of course, for symmetric countries $t_i^* = 0$ and any country gains from a socially optimal solution compared to the NE and compared to any smaller coalition size than the grand coalition.

13.3.3 Discussion

In this sub-section we shall discuss the assumption and properties of the core concept as well as those of the models used by Chander and Tulkens in various papers:

1. In many papers, for example, Chander and Tulkens (1992, 1997), the authors use a far more complicated model than the simple global emission model that we introduced in Chapter 9. Since all main results of Chander and Tulkens, which have been summarized in Proposition 13.5, can also be established (without qualifications) in the simple framework, the motto of Rasmusen of *'no fat modeling' has been violated* by these scholars.[37]
2. Though this has not been explicitly mentioned above, the core concept assumes as the conjectural variation models *only two groups of countries*: those which form a coalition and those outside the coalition which either play as singletons (γ core) or jointly form a 'punishment coalition' (α and β core).
3. By its very nature, the core is a *static equilibrium concept*. However, from the definition of the characteristic function it is clear that some behavioral assumptions are made as to how players react if a country

leaves the coalition. Thus, the assumptions of the core are very much in the tradition of the conjectural variation models which may be interpreted as *static representations of dynamic games* (see Tulkens 1998). *Instant reactions* are assumed which do not allow a defector to net any transitory gain. Thus, the same critique as was raised with respect to conjectural variation models applies here (see Chapter 10 and Subsection 13.2.8), namely that the *core does not adequately account for the free-rider incentive* in international pollution control.

In view of this, defending the core concept by criticizing concepts like the strong Nash equilibrium (see Chapters 6, 7 and 12) or the coalition-proof Nash equilibrium (see Chapter 15), that they would not account for reactions by players to a deviation, seems a rather weak argument of Chander and Tulkens (1997, s. 3). Their critique is based on the fact that, in a *static* context, these equilibrium concepts determine (in the tradition of an NE) a best reply, *given* the best reply of the other players.[38] However, with a dynamic story in mind when analyzing a coalition formation process, it seems natural to evaluate the *dynamic* and not the static version of these alternative equilibrium concepts. A casual look at the definitions of the dynamic versions of these concepts (or recalling the definition of a strongly subgame-perfect equilibrium or renegotiation-proof equilibrium in a finite (Chapter 6) or an infinite (Chapters 7 and 12) time horizon in a two-player game) would immediately make clear that the critique is false. Moreover, after all, a consistent assumption in a *one-shot game* implies no reaction by the very meaning of this term (see the discussion in Section 10.1).

It is also not correct to claim that the core has an advantage over the conjectural variation models because the latter models would not account for reactions after a deviation of a (original) coalition member (Tulkens 1998, s. 4). As has been demonstrated above, conjectural variation models assume that if a signatory leaves the coalition, the reduced coalition reoptimizes its strategy.

4. If the core is given a *static interpretation*, then it is immediately evident that it is not an equilibrium in a non-cooperative game theoretical sense. If the core is given a *dynamic interpretation*, then the α *and* β *core versions* imply a *non-credible threat*. Neither in a finite nor in an infinite time horizon is minimaxing a player for the rest of the game a subgame-perfect trigger strategy. In contrast, the threat to play an NE once a player defects as assumed in the γ *core* concept is a *subgame-perfect equilibrium* in a supergame framework.[39] However, such a threat would neither be a weakly or strongly renegotiation-proof nor a strongly perfect equilibrium.
5. It seems *not very plausible that according to the* γ*-characteristic function*

a coalition splits up into singletons once a country leaves the coalition. This would only be convincing if the number of signatories were smaller than the usually considered grand coalition and if the coalition size corresponded to the minimum number of countries specified in some minimum participation clause of a treaty. A more natural assumption would be that the coalition without the defector j reoptimizes their strategies as laid out in Sub-section 13.2.8 in the context of the conjectural variation models.[40]

6. Though the *characteristic function* summarizes the strategies of the coalition in a compact way, *information about a single player's incentives gets lost*. As in the conjectural variation models, it is hardly realistic to assume that all coalition members unanimously maximize joint welfare[41] and that transfers are truly paid according to formula (13.27).
7. The transfer scheme in (13.27) is similar to the *ratio equilibrium* of Kaneko (1977) and Mas-Colell and Silvestre (1989) and therefore it has sound justification in public goods economics. Its advantage over the 'conventional' type of compensation payments is that it is explicitly linked to emission reductions in each country. Thus, it can be given an interpretation as a *cost-sharing rule* between countries pursuing a joint abatement policy.

   However, apart from the (open) question of how such a transfer scheme can be credibly enforced, *formula (13.27) provides an obvious incentive to countries to misrepresent their environmental preferences*. Knowing that a country's contribution to an international monetary fund crucially depends on the fraction $\phi_i'(e^S)/\Sigma\phi_k'(e^S)$, a country will bias its estimation of $\phi_i'(e^S)$ downward. By the same token, it must be expected that countries do not truly reveal their opportunity cost of abatement, so as to induce transfers which are more favorable to them. Taken together, this implies that it can hardly be expected that the 'true' social optimum can be correctly determined (see Chander and Tulkens 1992, pp. 396ff.).
8. The *explanatory power of the core concept* as applied in all papers on international environmental problems *is rather low*. The finding that the socially optimal emission vector lies in the core if associated with an appropriate transfer scheme is in stark contrast to the empirical findings that for some environmental problems no IEA exists, the participation rate of most IEAs is well below full participation, and abatement targets are far from being socially optimal. One reason for the optimistic results generated by the core concept is the assumption that the coalition resolves once a member leaves it. This is a very harsh punishment, which helps to sustain a first-best solution. Another important reason is the assumption of transfers. If transfers were *ruled out*,

then the social optimum would hardly lie in the core, which is particularly true for the γ core. Then smaller coalitions and abatement targets below socially optimal ones may only lie in the core. A reason why such a modification has not been considered so far is the more surprising since no plausible explanation is given as to how such a transfer scheme can be realistically enforced. The assumption of Chander and Tulkens that such a transfer scheme is simply monitored and administered by an international agency seems rather naive.

13.3.4 Dynamic Adjustment to a Socially Optimal Steady-state Equilibrium

In some papers the core concept has been integrated into an explicit *dynamic story*. The basic idea is to test whether it is possible for a coalition to reduce emissions gradually from NE to socially optimal levels (Tulkens 1979). Along such an emission path the local optima have to be in the core in order to view such a path as stable. The motivation for the assumption of a dynamic emission path is twofold: first, it is hardly realistic to assume that players move at once from the non-cooperative equilibrium to the full cooperative equilibrium; second, governments may only possess information about the damage and abatement cost functions around the present state but not over the full range of these functions. Thus, governments adjust their emissions only gradually towards the final state by using local information.

Whereas in Chander and Tulkens (1991, 1992) such a 'time iterative process' is analyzed in *continuous time*, in Germain *et al.* (1995, 1996), Kaitala *et al.* (1995) and Germain *et al.* (1998) this is done in a *discrete time* setting. Whereas some of the papers (such as Chander and Tulkens 1991, 1992, and Kaitala *et al.* 1995) assume a notion of the stability concept slightly different from the previously discussed core concepts, Germain *et al.* (1996, 1998) base their stability analysis on the γ core concept. All papers assume a flow pollutant, except the last-mentioned paper.[42] Since all previous models in this book have assumed a flow pollutant and since we have identified the γ core concept as the most suitable among the core concepts, the following discussion is based on Germain *et al.* (1995, 1996).

The basic idea of all the papers is not much different from that in the static context. One checks whether an imputation of the grand coalition lies at each time $t = 1, 2, ..., T$ in the γ core, assuming a locally optimal emission vector and some transfer scheme. After T iterations the final optimal state will have been reached. This state is the global optimum, or in the former terminology, the social optimum. The transfer scheme assumed to

achieve this is very much in the spirit of that used in the static context above (see (13.27)).

More concretely, assume payoff at time t to be given by:

$$\pi_{i,t} = \beta_i(e_{i,t}) - \phi_i(\Sigma e_{k,t}) \tag{13.28}$$

where $\beta_i(e_{i,t})$ is a *concave* benefit function and $\phi_i(\Sigma e_{k,t})$ is a *linear* damage cost function, that is, $\phi_i(\Sigma e_{k,t}) = d_i \Sigma e_k . \beta_i(e_{i,t})$ is assumed to be known for all $e_{i,t} \geq \bar{e}_{i,t} \geq 0$ where:

$$\bar{e}_{i,t+1} = \min\,[\bar{e}_{i,t},\, \eta_{i,t+1} e^*_{i,t+1}],\ 0 \leq \eta_{i,t+1} \leq 1. \tag{13.29}$$

It can now be shown that $\max_{e^*_t} \Sigma \pi_{i,t}$ subject to constraint (13.29) leads to the global optimum after a finite number of iterations where e^*_t denotes the locally optimal emission vector of the grand coalition at time t, that is, $e^*_{t=\mathrm{T}} = e^{\mathrm{S}}$ (see Germain *et al.* 1995, Theorem 1).

The transfer scheme assumed during the iterative process is given by:

$$t^*_{i,t} = [\beta_i(e^N_i) - \beta_i(e^*_{i,t})] - [\phi_i(e^N_i) - \phi_i(e^*_{i,t})] + \Sigma \psi_i [g_j(e^*_{j,t}) - g_j(e^N_j)] \tag{13.30}$$

where:

$$g_{i,t} = \beta_i(e_{i,t}) - d \cdot e_{i,t},\ d = \Sigma d_i,\ \psi_i = \phi'_i / \Sigma \phi'_k = d_i / d,\ 0 \leq \psi_i \leq 1,\ \Sigma \psi_i = 1. \tag{13.31}$$

Then, as in the static case, the following result can be established:

Proposition 13.6
Let the payoff function from global emissions at time t be given by (13.28) where the benefit function is strictly concave and damages are linear and let $e^*_t = (e^*_{1,t}, \ldots, e^*_{N,t})$ be the (unique) locally optimal emission vector. Assuming that countries have local information about payoffs for all emissions as defined by (13.29) and that they gradually approach the global optimum via local optima in several stages, then the imputation $\pi^{*\psi}_t = (\pi^{*\psi}_{1,t}, \ldots, \pi^{*\psi}_{N,t})$ defined by $\pi^{*\psi}_{i,t} = \pi^*_{i,t}(e^*_t) + t^*_{i,t}$ where $t^*_{i,t}$ is given by (13.30) lies in the core at each stage $t = 1, \ldots, \mathrm{T}$.

Proof: See Appendix X.3. QED

We would like to finish this section with two remarks. First, the assumption of a time-iterative process in which countries gradually approach an

optimal state is a major step in adapting models to reality (see also Sub-section 12.2.4). Many international environmental agreements have seen an evolution as described by this process (Caldwell 1984). Commonly, to begin with, a formal IEA is only a framework convention, stating good intentions. Then moderate abatement obligations are specified in a next step; these are successively revised and increased over time in successive protocols. Examples include the Montreal Protocol on the reduction of the depletion of the ozone layer and the Helsinki Protocol to reduce sulfur.

Second, though the adjustment of emissions is modeled in a dynamic framework, the threat and punishment strategies do not change compared to the static version of the core concept (see Appendix X.3). From a game theoretical point of view this is disturbing. As demonstrated for repeated games, it should be expected that the time horizon would have an influence on players' strategies. Basically, the core concept as applied in the above-mentioned papers implies that players follow a strategy which neither is based on the history of the game nor does it take the future into account: an assumption which can hardly be rationalized.

NOTES

1. An empirical paper which analyzes the stability of a grand coalition set up to fight acid rain in Europe is Finus and Tjøtta (1998). Since their setting resembles that of Chapter 14, their work will be discussed there.
2. This issue will be considered more thoroughly in Sub-section 13.2.5.
3. See the brief discussion of these concepts in Chapter 11, Section 11.3.
4. The definition implies that the status quo is given by the Nash equilibrium where no coalition has formed. The term coalition is reserved for the cooperation of at least two countries.
5. The '$\geq$' instead of the strict inequality sign is used for two reasons. First, it seems equally likely that a signatory remains in a coalition or leaves it if nothing is gained from leaving. Also, a non-signatory may remain outside or may join the coalition if this does not affect his/her payoff. Second, the strict inequality sign would imply that for the net benefit function of type 1 below no coalition equilibrium exists. (Though Carraro and Siniscalco 1993, 1998 use the strict inequality sign in their definition of stability, they report on small coalitions obtained in various model versions based on payoff functions of type 1.)
6. The advantage of this relative index is twofold. First, the index is independent of the unit of measurement. Second, the signs of the derivatives hold for all three payoff functions and for the Cournot and Stackelberg assumptions. Moreover, they would also hold if the functions were slightly modified, and therefore rather general conclusions are possible using this relative index. In contrast, Barrett (1994b) uses the absolute index and derives results which *only* hold for specification (13.3) but would *not* hold if damages were assumed to be given by $\phi_i = f_i(\Sigma e_i^J + e_j^{NJ})$ instead of $\phi_i = (1/N) f_i(\Sigma e_i^J + e_j^{NJ})$.
7. In Endres (1997) the parameter c is interpreted as the environmental awareness of the society of a country. He analyzes the effect of a change of this variable on the degree of externality in a static game with two countries. In Endres and Finus (1998c) the analysis is extended to N countries and a dynamic framework. Particular attention is given to the effect of a change in the environmental consciousness of society on the success and stability of an IEA.

8. We assume that the appropriate non-negativity constraints are invoked on the parameters such that $e_i^N \geq 0$ and $e_i^S \geq 0 \; \forall \; i \in \mathrm{I}$ hold.
9. $\pi_i^S > 0 \; \forall \; i \in \mathrm{I}$ holds for all payoff functions.
10. We assume in the following that the appropriate constraints on the parameters apply so that positive emissions are ensured.
11. A signatory in a coalition of $N^* = 3$ is indifferent about leaving the coalition, implying $N^* = 2$, and receiving a non-signatory payoff. Hence, $N^* = 3$ and $N^* = 2$ seem equally likely, though Carraro and Siniscalco (1991) claim that there is a unique equilibrium of $N^* = 3$.
12. Leakage effects have been studied in the energy market by Bohm (1993); Felder and Rutherford (1993); and Golombek *et al.* (1995). Energy conservation by 'green' countries leads to a reduction in demand for crude oil. Consequently, energy prices drop. This in turn triggers higher demand by less environmentally conscious countries and renders the environmental policy of green countries less effective. See also Barrett (1992c); Bohm and Larsen (1993); and Hoel (1994) on leakage effects and countervailing measures.
13. In a symmetric Nash equilibrium $e_i^N = b/c$ and $\pi_i^N = (b^2(2-N))/2c$. As long as $N^* < N$, the best strategy of signatories as Stackelberg leaders is to expand emissions up to $\Sigma e_i^{J*} = bN/c$ so that non-signatories emit nothing. Then, a signatory receives $\pi_i^{J*} = (b^2N(2-N^*))/2cN^*$ and a non-signatory $\pi_i^{NJ*} = -Nb^2/2c$. It is easily confirmed that $C_3 > 0$ for any $N^* < N$, and $C_2 > 0$ for $N^* = N$. (External stability is automatically ensured for $N^* = N$ by definition.)
14. This result contrasts to Barrett's (1994b) finding of no stable coalition for this payoff function.
15. Differentiating a signatory's payoff function with respect to e_i^J, assuming $\Sigma e_j^{NJ} = 0$, gives $\partial \pi_i^J / \partial e_i^J = bd - be_i^J - cN\mu^2 e_i^J$. Substituting $bd/\mu c$ for e_i^J (which follows from $\Sigma e_i^J = \gamma dN$) we get $\partial \pi_i^J / \partial e_i^J = bd - b2d/\mu c - N\mu bd < 0$ where the sign follows from $\mu \geq 1/n$. Since $\partial^2 \pi_i^J / \partial e_i^{J2} < 0$, it follows that $\partial \pi_i^J / \partial e_i^J < 0 \; \forall \; e_i^J \geq bd/\mu c$ which implies by symmetry $\Sigma e_i^J \geq \gamma dN$. QED
16. This claim and all subsequent assertions in this section are proved in Appendix X.1.
17. Barrett (1994a, b) does not consider the profitability constraint when determining the coalition equilibrium.
18. Only a framework convention has been signed stating 'good intentions' at zero costs, which explains why it has been signed by many countries.
19. Of course, a shortcoming is that the failure of some IEAs to come into force cannot be explained by this model version since $N^* \in [2, N]$. Then, payoff functions of the second type and Nash–Cournot behavior could be a possible explanation. See also the model in Chapter 14 which does not suffer from this deficiency and where $N^* \in [0, N]$.
20. Both concepts imply that aggregate welfare of the coalition is maximized.
21. Xepapadeas (1997, p. 205) tries to justify such assumptions of commitment. He writes: 'The internal stability requirements . . . can, however, be regarded as too restrictive in reality.' However, we feel that such arguments are misguided. Once a stability concept has been chosen, one should stick to the implied results and not reinterpret them because they are 'too strong' compared to the gray tones of reality.
22. On the characteristics of private, club and public goods, see Arnold (1992); Cornes and Sandler (1986).
23. Therefore, Carraro and Siniscalco (1992, 1993) frequently stress that the underlying incentive structure of international environmental problems is not that of a PD game but rather of a chicken game, where each country would like to be a non-signatory but would join the coalition if it did not form (see Section 3.6). However, as shown above for a continuous strategy space, in a consistent conjectural variation framework no coalition exists for symmetric countries and in the traditional stage game framework the incentive structure resembles that of a PD game more than of a chicken game. Only for the 'hybrid' Definition 13.2 of a coalition equilibrium is their assertion true. See also the discussion in Sub-section 9.7.2 on the incentive structure in a global emission game.
24. This issue is taken up again in the context of the core (see Section 13.3) where exactly this kind of inconsistency occurs.

25. As Keohane (1995, p. 217) puts it: 'compliance is not very adequate. I believe that every study that has looked hard to compliance has concluded . . . that compliance is spotty.' In their prominent empirical study on compliance of IEAs, Brown Weiss and Jacobson (1997, pp. 87ff.) find severe instances of non-compliance for all IEAs covered by their study. Sand (1997, p. 25) reports that no fewer than 300 infractions of the CITES treaty have been revealed per year. Also the whaling convention has been frequently breached by all important parties to the treaty (Heister 1997, p. 68). Moreover, many IEAs also have a very poor compliance record with respect to reporting requirements (GAO 1992, pp. 3ff.; Sand 1996, p. 55; Bothe 1996, pp. 22ff.). Since official compliance-monitoring in almost all IEAs relies exclusively on self-reporting by the states, some suspicion with respect to good official compliance records of some IEAs seems also to be justified (Ausubel and Victor 1992, pp. 23ff.).
26. See, for example, Finus and Tjøtta (1998) and Murdoch and Sandler (1997). The extensive IIASA study on the effectiveness of IEAs draws a similar conclusion: see Victor *et al.* (1998, pp. 661ff.).
27. Due to the assumption of transfers, only payoff vectors based on the socially optimal emission vector (imputations) can qualify as potential candidates to lie in the core. (Otherwise a Pareto improvement for all coalition members would be possible.) For the assumption of no transfers this would be different.
28. More generally, the α and β characteristic functions are equivalent for games with transferable utility (Ordeshook 1986, p. 331). For the definition of minimax and maximin payoffs, see Sub-section 4.3.2.
29. Chander and Tulkens define the PANE with respect to $I^J \subset I$. However, there is no reason why this definition should not also comprise the grand coalition, that is, $I^J \subseteq I$.
30. See Tulkens (1998) on the similarity of both approaches.
31. $w(\{i\})$ may also be called the *disagreement point*. This is the payoff a player can expect to receive when players cannot agree on a grand coalition.
32. Note that the definition of the characteristic function is puzzling since it implies that if a player leaves the grand coalition s/he forms a 'one-player coalition' and the former coalition members become non-members. Whereas in the conjectural variation models the coalition members are the 'good guys' and the non-members the 'bad guys', this is reversed in the core. In other words, a country which does object to the terms of a treaty proposes an alternative coalition (Tulkens 1998).
33. Shapley (1971) showed that the core of convex games is non-empty. A game is called convex if its characteristic function is convex. (Since the characteristic function already includes the behavioral assumption of the countries outside the coalition, the result is true for any core concept.) This function is convex if $w(I_1^J) + w(I_2^J) \leq w(I_1^J \cup I_2^J) + w(I_1^J \cap I_2^J)$ $\forall\ I_1^J, I_2^J \subseteq I \Leftrightarrow w(I_1^J \cup \{i\}) - w(I_1^J) \leq w(I_2^J \cup \{i\}) - w(I_2^J)\ \forall\ i \notin I_2^J$, $I_1^J \subseteq I_2^J$ hold. That is, there is a kind of increasing marginal utility for coalition membership. It is easily checked that the latter condition does not necessarily hold in the global environmental context. For instance, consider $N = 4$, then $w(\{1, 2, 3\}) + w(\{2, 3, 4\}) \leq w(\{1, 2, 3, 4\}) + w(\{2, 3\})$ must hold. However, for a payoff function of type 3 in (13.3), $b = 0.2$ and $c = 1$ and the γ characteristic function, this condition is not satisfied.
34. An intuitive reason for this result is the observation that $w^{\alpha}(\{i\}) \leq w^{\gamma}(\{i\})$ holds.
35. As a matter of notation: $I^J \subset I$ implies that I^J is a true sub-coalition of I.
36. It is surprising that Chander and Tulkens (1995, p. 18) claim that this assumption would hold for a broad class of quadratic utility functions since a counter-example is readily constructed. For instance, this assumption fails for payoff functions (14.1) in Chapter 14, assuming $N = 3$, $1 < \gamma < 4$ (which takes condition NNC_1 in (XI.3), Appendix XI.1 into account) and if countries 2 and 3 form a coalition, that is, $I^J = (2, 3)$.
37. Of course, in some papers a transboundary pollution problem and not a global environmental problem has been assumed. However, this does not affect the main conclusions.
38. A definition in an *N*-country context is given in Chapter 15.
39. The underlying assumption is the payoff function $\pi_i = \beta_i(e_i) - \phi_i(\Sigma e_k)$ as described in (9.1) which satisfies assumptions A_1 in (9.2) and hence by Theorem 9.2 there is a unique

NE. Therefore no SPE except playing the stage game NE repeatedly exists in a finite setting by Theorem 4.2.

40. The behavior of non-coalition members according to the α and β characteristic functions has already been criticized in Sub-section 13.3.1.
41. See the justification for the LCD decision rule as assumed in the bargaining process in Chapter 11.
42. Germain *et al.* (1998) may be regarded as the most advanced of these papers. Whereas Chander and Tulkens (1991, 1992) are purely theoretical papers, the others apply their framework to an acid rain game between seven northern European regions.

14. Coalition models: a second approach

14.1 INTRODUCTION

In this chapter the coalition formation process in a *global emission game* is analyzed in a *supergame framework* by applying the concept of a weakly renegotiation-proof equilibrium (WRPE).[1] We proceed in two steps. First, stability of a grand coalition is analyzed (Section 14.3), assuming discount factors close to 1 (Sub-section 14.3.2) and discount factors smaller than 1 (Sub-section 14.3.3). Stability will be investigated for seven different emission vectors:

1. a socially optimal solution;
2. a globally optimal uniform emission tax;
3. a globally optimal uniform emission reduction quota;
4. a uniform emission tax if the median country's proposal is applied;
5. a uniform emission reduction quota if the median country's proposal is applied;
6. a uniform emission tax if the lowest common denominator decision rule (LCD decision rule) is applied;
7. a uniform emission reduction quota if the LCD decision rule is applied.

Additionally, the maximum emission reduction under a uniform tax and uniform quota regime will be determined which can be sustained as a WRPE. This is done to evaluate the efficiency of the two policy regimes independently of the decision rule chosen among the negotiators.

Each emission vector implies a different abatement target and a different welfare allocation among countries. Transfers to alter the welfare allocation implied by an emission vector are not considered.[2] Since a uniform tax is a cost-efficient instrument to tackle a pure public bad (see Section 11.2), a globally optimal uniform tax leads to a socially optimal emission allocation and hence emission vectors 1 and 2 are identical. Emission vector 3 is generated if global payoffs are maximized, subject to the constraint that a

uniform emission reduction is applied. Emission vector 4 is derived if N countries make a proposal of a uniform tax; subsequently the proposals are ordered according to their level, for example, $t_1, t_2, ..., t_N$, and proposal $t_{(N+1)/2}$ is applied in all countries. The same applies to emission vector 5 under the quota regime. For the model specification in this chapter it will turn out that the median country's proposal is identical to the globally optimal level of that instrument. Thus, emission vector 4 is identical to 1 and 2 and emission vector 5 is identical to 3. Emission vectors 6 and 7 are determined according to the bargaining procedure outline in Chapter 11 (see in particular Section 11.4). That is, the smallest uniform tax (quota) proposal of the N countries, $t^A(r^A)$, is applied where $t^A = \min[t_1, t_2, ..., t_N]$ ($r^A = \min[r_1, r_2, ..., r_N]$).

The choice of the emission vectors is motivated as follows. Socially and globally optimal emission vectors 1 to 3 are chosen as typical benchmarks frequently encountered in welfare economics to evaluate second-best solutions. Moreover, these emission vectors imply that the (grand) coalition maximizes the aggregate welfare of all coalition members, which is a typical assumption of the conjectural variation models and the core model encountered in Chapter 13. The motivation to consider the median country's proposal (emission vectors 4 and 5) is threefold. First, in the public choice literature many models determine the election outcome according to the median voter (Mueller 1989). Second, it is straightforward to show that a bargaining game in which the 'median country decision rule' is applied has an equilibrium in dominant strategies as this has been demonstrated for the LCD decision rule. That is, no country has an incentive to misrepresent its preferences by putting forward a biased proposal. Third, the check on whether the median country proposal is a WRPE does *not* require the assumption of transferable utility (as does the LCD decision rule) and therefore is in the tradition of non-cooperative game theory.

The motivation to consider the LCD decision rule has been extensively laid out in Chapter 11 and therefore no further comment is needed.

Since it will turn out from the analysis in Section 14.3 that a grand coalition is very unlikely to form irrespective of the emission vector if N is large, the formation of sub-coalitions is investigated in a second step (Section 14.4). The formation process comprises the decision of *membership* and *coalition size*, the *choice of the abatement target* within the group of signatories and the *choice of the policy instrument* to achieve the envisaged abatement target (see Sub-section 14.4.1). The model allows us to derive the coalition size (Sub-section 14.4.2) and the choice of the policy instrument (Sub-section 14.4.3) *endogenously*, whereas for the membership and the abatement target some (plausible) *exogenous* assumptions will be

made. The results if all aspects of the formation process are considered simultaneously are reported in Sub-section 14.4.4.

The analysis of the first (grand coalition) and second step (sub-coalition) presents ample evidence to answer the two questions raised in the Introduction of Chapter 11 and one more:

1. Why are abatement targets within most IEAs rather low, even though cost–benefit considerations would suggest higher targets?
2. Why do only sub-groups of countries sign an IEA though more countries suffer from an environmental externality?
3. Why have effluent charges (as one representative of market-based instruments) not been applied in any IEA so far whereas quotas have found widespread application?

As in Chapter 13, the answers will be related to the model parameters benefit–cost ratio from emissions, $\gamma = b/c$, and the number of countries suffering from the externality, N.

14.2 PRELIMINARIES

The subsequent analysis is based on a standard type of payoff function similar to those encountered in Chapters 9 and 11 which satisfies assumptions A_1 in Chapter 9. In particular, consider the following payoff function of country i:

$$\pi_i = b\left(de_i - \frac{1}{2}e_i^2\right) - i\cdot\frac{c}{2N}\cdot\left(\sum_{k=1}^{N} e_k\right)^2,\ b>0,\ c>0,\ d>0,\ e_i \in [0, 2d],\ i \in \{1, ..., N\} \tag{14.1}$$

where b and d are benefit parameters and c is a cost parameter. This implies equal benefits from emissions in the N countries but different damages which are (discrete) uniformly distributed. Countries with a higher index i suffer more from pollution than do those with a lower index number.[3] Though equal benefits may be seen as a restrictive assumption in (14.1) this specification allows us to identify the interests of the N countries uniquely. Countries with a higher index have a greater interest in emission reduction than those with a lower index. This is evident from computing emissions in the static Nash equilibrium (NE) and the social optimum and where $\partial e_i^N/\partial i < 0$, $\partial(\pi_i^S - \pi_i^N)/\partial i > 0$ and $\pi_i^S - \pi_i^N <, >, = 0$ hold.[4] That is, countries with a higher index choose lower emissions in the non-cooperative equilibrium than those with a lower index. The high index countries also benefit

more from a socially optimal solution than low index countries due to their higher damage. In fact, low index countries may receive a lower payoff in the social optimum than in the NE (recall Proposition 9.2).

For specification (14.1) a familiar result from Chapter 13 also applies:

$$I_1 := \frac{\Sigma e_k^N - \Sigma e_k^S}{\Sigma e_k^S},\ I_2 := \frac{\Sigma \pi_k^S - \Sigma \pi_k^N}{\Sigma \pi_k^S},\ \frac{\partial I_1}{N} > 0,\ \frac{\partial I_1}{\gamma} < 0,\ \frac{\partial I_2}{N} > 0,\ \frac{\partial I_2}{\gamma} < 0. \qquad (14.2)$$

That is, the externality is distinct if many countries suffer from the global pollutant and if environmental damage is very high compared to the opportunity costs of abatement.

For payoff functions (14.1) it is straightforward to derive a globally optimal uniform tax, t^*, or expressed as relative emission reductions from the NE, $r^{T*} = \Sigma r_k^{T*}$ where $e_i^{T*} = (1 - r_i^{T*}) \cdot e_i^N$. As mentioned above, since a uniform tax is efficient in the context of global emissions, $t^* = t^S$ and $r_i^{T*} = r_i^S$ where $r^S = \Sigma r_k^S$, $e_i^S = (1 - r_i^S) \cdot e_i^N$ and t^S denotes a socially optimal or Pigouvian tax. Note that for (14.1) $e_i^S = e_j^S\ \forall\ i$ and $j \in I$. By the same token, a globally optimal uniform emission reduction quota, r^{Q*}, can be derived where $e_i^{Q*} = (1 - r^{Q*}) \cdot e_i^N$. Since $e_i^N \neq e_j^N$ and hence $e_i^{Q*} \neq e_j^{Q*}\ \forall\ i$ and $j \in I$, such a solution is not cost-efficient if benefit functions are identical in all countries as assumed in (14.1).

It is also straightforward to derive proposals t_i under a uniform tax regime and r_i under a uniform emission reduction quota regime. For (14.1) it turns out that $t_1 < t_2 < \ldots < t_{N-1} < t_N$ and $r_1 < r_2 < \ldots < r_{N-1} < r_N$. That is, under both regimes country 1 is the bottleneck which determines the terms of the agreement if the LCD decision rule is applied. This is not surprising since, given equal opportunity cost of abatement, country 1 evaluates environmental damages lower than its $N-1$ neighbors and is therefore interested in the lowest emission reduction.

As mentioned in the Introduction, $t_{(N+1)/2} = t^*$ and $r_{(N+1)/2} = r^{Q*}$ for (14.1). This implies that if in the subsequent analysis a globally optimal tax or emission reduction quota is not stable, this is also true for the median country proposal.

14.3 THE GRAND COALITION

14.3.1 Stability Concept

For the different emission vectors mentioned in the previous sections stability according to the WRPE concept can easily be checked. All relevant background information has already been derived in Chapters 7 and 12 so we can immediately proceed to the results. The only difference is that now the

WRPE inequality system may be very large if N is large. This is evident when recalling that in the case of discount factors close to one N^2 (see (12.1) and (12.2)) and in the case of discount factors smaller than $N \cdot (N+1)$ (see (12.17)–(12.19)) inequalities have to be satisfied simultaneously. Thus, in contrast to Chapter 12 where a stability check could be conducted analytically for $\delta \to 1$, in the case of more than two countries the analysis has to rely on simulations. However, this procedure seems to be justified because all simulation results presented below are quite robust and show a clear pattern. Moreover, due to the sheer size of the inequality system an algorithm is needed to check for stability. The details are provided in Appendix XI.2 for the case of $\delta \to 1$ and in Appendix XI.3 for the case of $\delta < 1$.

As in the case of two countries, only deviations by one country are considered. In the context of a global emission game this assumption seems to be justified (see also Sub-section 13.2.1). On the one hand, a country planning to take a free-ride has no incentive to do so jointly with other countries because the transitory gain is larger if other countries comply. On the other hand, on the side of the punishers there is no incentive to form a coalition when breaching the punishment obligations. If during the punishment phase a country among the punishers were to emit more (or less) than requested to punish the defector, it would not ask other countries to follow suit. The punishment constitutes a public good to control an agreement and therefore free-riding is most attractive to a country if conducted only by itself. Therefore, if a punisher does not meet its obligations, it is treated as a free-rider and will be punished. Of course, by the construction of the WRPE concept, such a punishment is regarded as unnecessary since the punishers receive more during the punishment phase than during the cooperative phase.

14.3.2 Stability Analysis for Discount Factors Close to 1

In Table 14.1 the results of the stability analysis are reported. To evaluate the different emission vectors a welfare indicator is provided which measures the 'degree of optimality'. The indicator is defined as follows:

$$DO^k = \frac{\Sigma\pi_i^k - \Sigma\pi_i^N}{\Sigma\pi_i^S - \Sigma\pi_i^N} \cdot 100 \tag{14.3}$$

where k stands for the policy regime and the particular abatement target, for example, $k = T_1$ or $h = Q_{max}$. Put differently, DO^k indicates how close a particular emission vector comes to the social optimum. By definition, the degree of optimality in the social optimum is 100 percent and since $r_i^{T^*} = r_i^S$, $DO^{T^*} = DO^S = 100$ percent holds. From Table 14.1 four main results may be derived:

1. Whenever the potential gains from cooperation would be great (N large and γ small, see (14.2)) stability poses a problem. This holds generally for 'optimal' and second-best solutions and is independent of the policy regime. This result is supported by noting that $\partial DO^{Qmax}/\partial N < 0$, $\partial DO^{Qmax}/\partial \gamma > 0$, $\partial DO^{Tmax}/\partial N < 0$ and $\partial DO^{Tmax}/\partial \gamma > 0$.
2. The socially optimal solution ($r_i^S = r_i^{T*}$) is only stable in a few cases where N is small and γ is large. These are the parameter constellations for which abatement targets are rather low. The same holds true for a globally optimal emission reduction quota (r^{Q*}). Hence, whenever a joint abatement policy is most attractive from a global point of view, 'optimal solutions' are not stable. Recalling that the median country proposal under the tax regime is identical to r_i^{T*} and under the quota regime to r^{Q*}, it is evident that such proposals scarcely have a chance of realization. Therefore, it does not seem sensible to consider the median country proposal any further.
3. Whenever the tax agreement is stable under the LCD decision rule, then this is also true for the quota agreement; however, the degree of optimality and global emission reductions are higher. Moreover, there are cases where the quota agreement is stable, though no tax agreement can be stabilized (for example, $N = 10$, $\gamma = 10$ and $\gamma = 30$).
4. The degree of optimality of the maximum backable emission reduction under the tax regime is higher than under the quota regime for the 'non-critical' parameter values (N small and γ large). In Table 14.1 this is true for $N = 5$, $\gamma \geq 50$ and $N = 10$, and $\gamma \geq 2000$. For all other parameter values the opposite holds.

The results summarized in 1 and 2 confirm a result established for the conjectural variation models in Chapter 13: namely, that whenever cooperation is needed most from a global point of view, a coalition achieves only little. In particular, for global environmental problems where N is large, a stable grand coalition seems very unlikely even if politicians were almost perfectly patient. But even if a grand coalition were stable, reduction targets agreed upon would have to be very low. This is a result which is confirmed by the historical evidence of IEAs signed so far.

The results summarized in 3 and 4 can be taken as a first indication that in a second-best world[5] the choice of a quota regime to tackle global emissions might be rational. Though governments cannot be expected to be concerned about global efficiency, it is evident that the typical efficiency argument in favor of the tax regime no longer holds in a second-best environment. From an institutional economics point of view this may explain why the quota and not the tax has emerged from the evolution of institutions.

Table 14.1 Stability analysis of the grand coalition for $\delta \to 1$[a]

		Quota					Tax				
N	γ	r^{Q*} (%)	r^{Q1} (%)	r^{Qmax} (%)	DO^{Q1} (%)	DO^{Qmax} (%)	r^{T*} (%)	r^{T1} (%)	r^{Tmax} (%)	DO^{T1} (%)	DO^{Tmax} (%)
5	10	*48.40*	19.60	42.00	64.0	97.4	*48.00*	13.30	23.30	49.6	74.4
5	30	26.80	10.10	26.80	60.3	98.9	*26.70*	5.70	19.30	39.5	92.6
5	50	18.50	6.70	18.50	58.6	98.8	*18.50*	3.60	16.70	36.7	99.1
5	100	10.50	3.70	10.50	56.9	98.7	10.40	1.90	10.40	34.2	100.0
5	500	2.30	0.80	2.30	55.3	98.7	2.30	0.40	2.30	32.1	100.0
5	1000	1.18	0.40	1.18	55.0	98.6	1.18	0.20	1.18	31.8	100.0
5	2000	0.60	0.20	0.60	54.9	98.6	0.60	0.10	0.60	31.7	100.0
5	3000	0.40	0.13	0.40	54.9	98.6	0.40	0.07	0.40	31.6	100.0
10	10	*76.50*	27.60	36.00	59.0	71.9	*76.20*	*22.50*	3.90	*51.4*	11.9
10	30	*58.40*	17.80	24.00	51.6	65.2	*58.20*	*11.30*	10.50	*35.5*	33.3
10	50	*47.20*	12.80	19.00	46.9	64.2	*47.10*	7.50	10.50	29.8	40.1
10	100	*32.00*	7.50	14.00	41.4	68.2	*31.90*	4.10	9.50	24.4	50.8
10	500	*8.90*	1.70	6.00	35.0	89.0	*8.90*	0.90	5.50	19.1	85.5
10	1000	*4.69*	0.88	4.20	34.0	98.6	*4.69*	0.45	3.73	18.4	95.8
10	2000	2.40	0.45	2.40	33.5	99.7	*2.41*	0.22	2.39	18.0	100.0
10	3000	1.60	0.30	1.60	33.3	99.7	1.62	0.15	1.62	17.9	100.0
20	10	*90.80*	*31.10*	31.00	56.8	56.6	*90.70*	*31.70*	−31.20	*58.4*	−77.4
20	30	*83.20*	*25.90*	22.00	52.6	45.9	*83.10*	*19.00*	−3.50	*40.9*	−7.9
20	50	*76.80*	*20.70*	18.00	46.7	41.4	*76.70*	*13.60*	1.50	*32.5*	4.3
20	100	*64.40*	*13.50*	13.00	37.6	36.3	*64.40*	*7.90*	4.50	*23.3*	13.8

20	500	*28.10*	3.50	6.00	23.5	38.1	*28.10*	1.80	4.40	12.7	28.9
20	1000	*16.49*	1.83	4.50	20.9	47.1	*16.49*	0.93	3.54	11.1	38.4
20	2000	*9.03*	0.93	3.20	19.6	58.3	*9.03*	0.47	2.72	10.2	51.3
20	3000	*6.22*	0.63	2.50	19.1	64.2	*6.21*	0.31	2.24	9.9	59.2
50	30	*95.80*	*32.30*	18.00	56.1	34.0	*95.70*	*30.60*	−43.10	*54.0*	−109.0
50	50	*94.30*	*30.10*	16.00	53.7	31.0	*94.30*	*24.50*	−23.50	*45.4*	−55.5
50	100	*90.90*	*23.80*	12.00	45.5	24.7	*90.90*	*16.30*	−8.80	*32.8*	−20.1
50	500	*70.40*	*8.20*	6.00	21.8	16.3	*70.40*	*4.50*	−1.40	*12.3*	4.0
50	1000	*54.93*	*4.45*	4.30	15.6	15.4	*54.92*	*2.33*	1.99	*8.3*	7.2
50	2000	*38.15*	2.33	3.10	11.9	15.6	*38.15*	1.20	1.92	5.0	9.8
50	3000	*29.23*	1.58	2.50	10.5	16.4	*29.23*	0.80	1.76	5.4	11.7
100	50	*98.00*	*32.90*	14.00	55.8	26.5	*98.00*	*33.00*	−64.82	*56.1*	−175.0
100	100	*97.10*	*30.40*	11.00	52.9	21.4	*97.10*	*24.80*	−30.60	*44.6*	−72.8
100	500	*90.10*	*14.10*	5.00	28.8	12.9	*90.10*	*8.30*	−3.30	*17.5*	−7.5
100	1000	*82.64*	*8.24*	4.30	18.9	10.1	*82.64*	*4.50*	−0.43	*10.6*	−1.0
100	2000	*70.92*	*4.50*	3.10	12.3	8.6	*70.91*	*2.36*	0.69	*6.6*	2.0
100	3000	*62.11*	*3.09*	2.50	9.7	7.9	*62.11*	*1.60*	0.93	*5.1*	3.0

Note: [a] r^{Q*} = globally optimal emission reduction quota; r^{Q1} = emission reduction quota according to the LCD decision rule; r^{Qmax} = maximum uniform emission reduction quota which can be backed as a WRPE; DO^{Qmax} = degree of optimality of r^{Qmax}; DO^{Q1} = degree of optimality of r^{Q1}; $r^{T*} = r^{S}$ = socially optimal emission reduction; r^{T1} = emission reduction under the tax regime according to the LCD decision rule; r^{Tmax} = maximum emission reduction under the tax regime which can be backed as a WRPE; DO^{Tmax} = degree of optimality of r^{Tmax}; DO^{T1} = degree of optimality of r^{T1}; italic figures indicate instability in the sense of WRPE.

14.3.3 Stability Analysis for Discount Factors Smaller than 1

In the previous sub-section the stability analysis was confined to a discount factor close to 1. In reality, however, political representatives are hardly (almost) perfectly patient since short-term success is important for politicians in democracies (Hahn 1987, p. 300). It will be interesting to investigate whether there are differences regarding the 'patience requirements' in the two policy regimes. Of course, such an analysis is only interesting for those parameter constellations of the previous sub-section for which stability is possible if $\delta \to 1$. In order to reduce the data set further, the analysis is restricted to $N \leq 20$.

For convenience, δ_i^{min} is expressed as maximum interest rate, ρ_i^{max}, where $\delta_i^{min} = 1/(1 + \rho_i^{max})$ holds.[6] Therefore, ρ_i^{max} is the maximum discount rate for which a country can be deterred from taking a free-ride by the other $N-1$ countries. Since it turns out that country 1 has always the lowest *maximum discount rate requirement*, attention has been restricted to ρ_i^{max} in Table 14.2 exclusively. In other words, $\rho_i^{max} \geq \rho_i$ (where ρ_i is the actual discount rate) is a necessary and sufficient condition for country 1 ($i = 1$) not to take a free-ride (provided all other countries comply with the agreement). However, it is a sufficient condition for all other countries, $i \neq 1$.

The following analysis is restricted to the abatement target resulting under the LCD decision rule. Two main results can be derived from Table 14.2:

Table 14.2 Discount rate requirement of country 1 under the LCD decision rule[a]

		γ							
		10	30	50	100	500	1000	2000	3000
$N = 5$	Quota	16	11	10	9	8	8	8	8
	Tax	7	6	5	5	5	5	5	5
$N = 10$	Quota	1.8	2.1	2.0	1.7	1.5	1.5	1.5	1.4
	Tax	—	—	1.0	1.2	1.1	1.1	1.1	1.1
$N = 20$	Quota	—	—	—	—	0.3	0.3	0.3	0.3
	Tax	—	—	—	—	0.2	0.2	0.2	0.2

Note: [a] The discount rate ρ_1^{max} has been calculated on the 1 percent significance level for $N = 5$ and on the 0.1 percent level for $N = 10$ and $N = 20$. — means not stable for $\rho_1 \to 0$.

1. Only if the number of countries suffering from the global externality is small can a stable agreement be expected. Already in the case of ten countries the discount rate must be less than 2.1 percent. For $N=20$, country 1 must be even more patient, that is, $\rho_i \leq 0.3$. The result backs the assertion above that a grand coalition is very unlikely to form if a pollution problem is of a real global character.
2. The discount rate requirement of country 1 is less restrictive under the quota than under the tax regime. If low index countries are not only less interested in emission reductions but also have a higher time preference rate than their neighbors, then these 'critical' countries find it easier to comply with a uniform emission reduction quota than with a uniform effluent charge.

The first result again suggests an investigation of the formation of subcoalitions. This will be done in Section 14.4 below. The second result can be interpreted as an additional explanation of the fact that the emission reduction quota has been so popular in international pollution control, though it is not a cost-efficient instrument.

14.3.4 Extensions

From the previous section it appeared that for global externalities a grand coalition is unlikely to be a WRPE if the number of countries is large. Considering that this result has been obtained assuming 'optimal' punishment profiles, it is clear that if punishment profiles are restricted for some reason, stability of a grand coalition would be even more unlikely. In particular, in the case of many countries, the optimal coordination of punishment may be difficult and associated with high transaction cost. Therefore, in reality one should expect that simple punishment profiles according to rules of thumb will be applied to allow immediate and deterrent punishment. The effect of such simple punishment profiles on the discount rate requirement is investigated in Finus and Rundshagen (1998a). It is not surprising that the authors find that the discount factor requirements are higher for restricted than for unrestricted punishment profiles (that is, $r_i^{\max}$ (unrestricted) $> r_i^{\max}$ (restricted); see Chapter 12). What is more interesting is their finding that, particularly for the critical parameter values (N large and γ small), the discount factor requirements are greatly increased if simple punishment profiles are applied. That is, countries face a dilemma: on the one hand, particularly if many countries participate in an IEA, simple strategy profiles will be employed to make coordination easier; on the other hand, in this case the discount rate requirements are already very high assuming optimal punishment strategy profiles. Moreover, Finus and

Rundshagen's results suggest that it is not only worthwhile thinking about the effect on the success of an IEA of the agreement procedures (for example, median country proposal versus LCD decision rule) and the instruments employed to achieve an abatement target agreed upon in an IEA (such as taxes versus quota), but also to think about which rule of thumb should be chosen to coordinate punishment. Different rules lead to different discount factor requirements and therefore the design of punishment obligations should be given a high priority when negotiating an agreement.

In the previous section it turned out that, particularly for the critical parameter values, a socially optimal emission reduction is not a WRPE. From Chapter 12 it is clear that if the social optimum is a WRPE, it is automatically a strongly perfect renegotiation-proof equilibrium (SRPE) too. The quota and tax agreements according to the LCD decision rule, although they might be a WRPE, are never an SRPE. Since for payoff function (14.1) the tax and quota proposals of countries differ, that is, $t_i \neq t_j$ and $r_i \neq r_j$, agreements based on the LCD decision rule are never Pareto-efficient. Consequently, only a socially optimal agreement is a potential candidate which could qualify as strongly subgame-perfect equilibrium (SSPE). Since an SSPE requires not only Pareto-efficient normal phase payoffs and a credible punishment profile but also an efficient one, the set of parameter values for which a socially optimal agreement is an SSPE is smaller than that for which such an agreement is an SRPE. By the same token, imposing only the restriction that an agreement is a subgame-perfect equilibrium (SPE) renders the set of parameter values larger than that for which it is an SRPE.

The first extension is considered by Barrett (1994a, b). In a similar model setting he investigates the parameter set for which a grand coalition – aiming at a socially optimal agreement – is an SSPE. His findings are very much in line with those obtained above. For those parameter constellations for which a grand coalition could substantially improve upon the status quo it is either not stable or achieves only little.

Results of the second extension are reported in Finus and Rundshagen (1997). It is shown that for many parameter values for which a socially optimal agreement (r^{T*} in Table 14.1) and an agreement on a globally optimal uniform quota (r^{Q*} in Table 14.1) cannot be sustained by WRPE strategies, this is possible by SPE strategies. Moreover, the quota and tax agreement according to the LCD decision rule (r^{Q1} and r^{T1} in Table 14.1) are stable for almost all parameter values. Thus, by applying the SPE concept the problems in international pollution control to form stable IEAs could not be explained. Moreover, conclusions about when a stable agreement could be expected and when such an agreement would be effective

would not be possible. This stresses that the WRPE concept is a useful extension of the SPE concept in order to explain politically interesting and relevant phenomena.

This last assertion is supported by an empirical study by Finus and Tjøtta (1998). The authors analyze the stability of a grand coalition which intends to reduce acid rain in Europe. They show that a socially optimal agreement would not be a WRPE. The reason is twofold. First, due to the wind patterns in Europe emission transportation coefficients (see Chapter 10, note 18) exhibit a large asymmetry. Second, environmental preferences and abatement costs vary substantially between countries. Both reasons imply that the gains from a socially optimal agreement are very unevenly distributed among signatories. Therefore, the authors go on to determine the maximum backable emission reduction (such as $r^{Q\max}$ in Table 14.1) by WRPE strategies. They find that the maximum emission reduction is only marginally above that in the NE. This result helps to explain why emission reductions as laid down in the Oslo agreement signed in 1994 are not a 'great leap forward' and basically freeze emissions at the NE level of the year 2000.

14.4 THE SUB-COALITION

14.4.1 The Formation Process

So far it has been assumed that if countries decide jointly to improve environmental quality they seek to form a coalition of all countries. It became apparent that there are many obstacles which render the probability of a stable grand coalition rather low. This result is perfectly in line with existing environmental conventions in which only some of N countries participate if they come into force at all. For example, the Convention for the Regulation of Whaling has been signed by 36 countries and the Antarctic Treaty by 26 countries, even though all states should have an interest in the problem.

Basically, environmentally conscious countries face a dilemma. On the one hand, they like to see many countries participating in an IEA to improve environmental quality and to share the abatement burden with them. On the other hand, if the LCD decision rule is applied, the abatement target which can be established within a grand coalition might be very low. Moreover, it has been demonstrated that less environmentally concerned countries are also the critical ones regarding stability. Therefore, one would suspect that those countries which have a high interest in an international environmental policy seek partners but only among those with similar interests. This

procedure seems even more likely when taking into account that coordination in international politics is associated with transaction costs.[7]

As mentioned in the Introduction to Chapter 13, the number of possible sub-coalitions is very large.[8] To narrow down the analysis, it is assumed that countries with higher environmental damages will form a coalition among themselves (see also Botteon and Carraro 1997; and Hoel 1992a). Therefore, in the context of the present model the set of non-signatories is given by $I^{NJ} = \{1, 2, ..., N_0\}$, and the set of signatories by $I^J = \{N_0 + 1, ..., N\}$. Within a potential coalition three issues have to be decided simultaneously: *abatement target*, *instrumental choice* and *coalition size*.

Abatement target

It is assumed among signatories that the LCD decision rule is applied to determine the abatement target. As in the case of a grand coalition, it turns out that for a given coalition size it is the country with the smallest index which makes the lowest proposal. Therefore, it is this country's government which determines the terms of the agreement (see Appendix XI.1). For instance, if $N = 10$ and $I^J = \{7, 8, 9, 10\}$, then country 7 is the bottleneck country and its proposal is laid down in the treaty.

For the non-signatories it is assumed that they choose their initial non-cooperative emission level, e_i^N. Since non-signatories have an incentive to increase their emissions above e_i^N as a result of the higher abatement efforts of signatories, it has to be checked whether signatories can control the 'free-riding' of non-signatories through suitable punishments.

From this it follows that the stability of an IEA comprises two components:

1. *External stability* It has to be checked whether it is possible for a coalition to control emissions of non-signatories using WRPE strategies. That is, non-signatories are not allowed to increase their emissions above the initial level, e_i^N.
2. *Internal stability* It has to be investigated whether signatories can control their agreement internally using WRPE strategies. That is, signatories are not allowed to increase their emissions above the level implied by the agreement on t_{N_0+1} and r_{N_0+1}.

Due to lack of space the following analysis is restricted to $\delta \to 1$. A description of the stability test is given in Appendix XI.4.

Instrumental choice

Potential signatories choose the instrument employed in the IEA according to the majority decision rule (MDR). For instance, if $N = 10$, $I^J = \{7, 8, 9, 10\}$ and preferences are given by 7: Q, 8: Q, 9: Q and 10: T, this coalition of

four countries would employ an emission reduction quota. The same applies for any preliminary coalition which is constructed during the coalition formation process.

Coalition size

It is assumed that the coalition formation process starts with country N (the most environmentally conscious country) approaching country $N-1$ in order to discuss a joint environmental policy.[9] In this first round each country makes an emission reduction proposal. According to the LCD decision rule, only the smaller of both suggestions will be realized. This is country $N-1$'s proposal. Since it can be shown that country N always prefers to cooperate with country $N-1$, even though country N's proposal is not accepted, the LCD assumption seems indeed plausible.[10]

For this preliminary sub-coalition the associated welfare levels of countries N and $N-1$ can be computed. They serve as a benchmark to decide whether a third country should be asked to join. In the second round of the negotiation process country $N-2$ is asked to submit its proposal if countries $N-2$, $N-1$ and N would form a coalition. For this new coalition, countries N and $N-1$ will also make an offer. Again, the parties have to agree on the smallest offer which is submitted by the countries with the lowest index number, that is, country $N-2$. Provided country N and $N-1$ benefit if $N-2$ joins (and determines the condition of the treaty), then these countries offer country $N-2$ participation in their preliminary agreement. If both countries were to lose if country $N-2$ joins, these countries would form a coalition only among themselves. If countries N and $N-1$ hold different views regarding an extension of its preliminary coalition, the coalition formation process stops. This is because otherwise the country which is against enlargement can threaten to leave the coalition. Hence, the coalition size is decided according to the unanimity decision rule (UDR) where the formation process stops once a country votes against an enlargement.[11] Thus, in contrast to the conjectural variation models, signatories can defend their coalition against potential accessors if this is not beneficial to them.

In order to keep track of the simultaneous decision process, we proceed in two steps. In the first step the equilibrium number of signatories is determined, assuming the instrumental choice to be exogenously given (Subsection 14.4.2). In the second step, this is done if the instrumental choice is endogenously determined within the negotiation process leading to an IEA (Sub-section 14.4.3). Results are discussed in Sub-section 14.4.4.

14.4.2 The Equilibrium Number of Signatories in the Two Policy Regimes

To gain a deeper understanding of the coalition formation process, a typical example is shown in Table 14.3. All relevant background information is provided in Appendix XI.1.

Table 14.3 The formation process of a sub-coalition[a]

Coalition size		1	2	3	4	5	6	7	8	9	10
Country N_0+1		10	9	8	7	6	5	4	3	2	1
	$r_{N0+1}(\%)$	0	29.3	45.0	51.3	51.9	49.1	**44.1**	**37.5**	29.1	17.8
	$r_{\max}(\%)$	0	32.5	53.5	65.0	69.9	67.0	**54.0**	**44.0**	34.0	24.0
	$r^{Q}(\%)$	0	5.1	11.9	18.5	23.8	27.5	**29.3**	**29.0**	25.8	17.8
	π_{10}	−2190	−2081	−1819	−1508	−1216	−981	**−820**	**−755**	−828	−1163
	π_9		−1730	−1514	−1244	−983	−766	**−614**	**−546**	−603	−895
	π_8			−1210	−981	−750	−552	**−409**	**−338**	−378	−627
	π_7				−718	−517	−339	**−204**	**−132**	−155	−362
Quota:	π_6					−285	−126	**0**	**74**	67	−98
	π_5						87	**203**	**279**	287	165
	π_4							**406**	**483**	507	425
	π_3								**685**	725	685
	π_2									942	942
	π_1										1198
	Enlargem. **yes:no**	1:0	2:0	3:0	4:0	5:0	6:0	**7:0**	**3:5**	0:9	—

	t_{N_0+1}	84.5	**144.4**	**178.3**	**191.8**	**191.5**	**182.2**	166.5	145.1	116.6	75.0
	t_{max}	84.5	**151.5**	**195.8**	**220.1**	**229.9**	**201.0**	165.0	134.0	105.0	73.0
	$r^T(\%)$	0	**5.1**	**12.1**	**18.9**	**24.4**	**28.1**	29.2	25.0	19.3	10.5
	π_{10}	−2190	**−2066**	**−1789**	**−1460**	**−1150**	**−898**	−742	−810	−1011	−1452
	π_9		**−1744**	**−1513**	**−1225**	**−946**	**−714**	−563	−609	−778	−1165
	π_8			**−1237**	**−991**	**−742**	**−529**	−385	−408	−545	−879
	π_7				**−756**	**−538**	**−345**	−206	−207	−313	−593
Tax:	π_6					**−335**	**−160**	−27	−6	−80	−306
	π_5						**24**	152	195	153	−20
	π_4							331	396	386	266
	π_3								597	618	552
	π_2									851	839
	π_1										1125
	Enlargem. **yes:no**	1:0	**2:0**	**3:0**	**4:0**	**5:0**	**6:0**	3:4	2:6	0:9	—
Countries in favor of a quota			9	8	7–8	6–7	5–7	**4–7**	**3–10**	2–10	1–10
Countries in favor of a tax			**10**	**9–10**	**9–10**	**8–10**	**8–10**	8–10			
Instrument chosen			T	T	T	T	T	Q	Q	Q	Q

Note: [a] The example assumes $N = 10$, $b = 30$, $c = 1$ and $d = 10$.

From the table it is evident that in the quota regime all potential signatories favor an enlargement of the coalition until a coalition size of eight countries is reached, including countries 3–10. If eight countries participate in the IEA, only three countries are in favor of an enlargement. Therefore, according to the UDR, the equilibrium coalition size comprises eight countries (shaded column).

Up to a coalition size of five countries the reduction level proposed by country $N_0 + 1$, r_{N_0+1}, increases. In this range ambitious abatement targets are proposed by country $N_0 + 1$ because the marginal benefits from a joint emission reduction are high. Then, for a coalition size above five countries, the proposal of country $N_0 + 1$ decreases. Nevertheless, all potential signatories benefit from an enlargement until a coalition of eight countries has formed. In this range, the gains to the potential signatories from an additional contributor are stronger than the (negative) effect of a lower joint abatement target. If more than eight countries were to join, then the reduction target would become too low for countries with high environmental damage and a majority of countries would be against an enlargement.

In contrast, in the tax regime the equilibrium coalition size is not eight but seven countries. In this example country 4's tax proposal, $t_4 = 165.5$ cannot be backed by WRPE strategies. Therefore, it is assumed that the maximum stable tax rate, $t_{\max} = 165$ will be applied in the agreement. Since countries 5–10 prefer a coalition of seven countries (which then includes country 4) to a coalition of only six countries, though only the low tax rate $t_{\max}$ is applied country 4 is asked to join.

For other parameter constellations Table 14.4 summarizes the relevant information which will be discussed in Sub-section 14.4.4.

14.4.3 The Endogenous Determination of Instrumental Choice

So far the policy regime has been assumed to be exogenously given and the equilibrium coalition size has been derived for each regime. Now, the choice of the policy regime within the negotiation process is also considered. This choice will affect the equilibrium coalition size and the abatement target agreed upon in an IEA as well.

Generally, countries may hold different views regarding the instrument which is applied in an IEA. There is a tendency for countries with low index numbers to prefer the emission quota because for a given reduction level they have to carry a lower reduction burden than under the tax regime.[12] In contrast, countries with high index numbers prefer a tax because this leads to an emission allocation which favors them.

However, since the instrumental issue is linked to the other two issues, countries with high index numbers might also favor a quota if this implies

that more countries participate in an IEA and/or the abatement target agreed upon is higher than in a tax regime. To gain more insight into the forces we again use Table 14.3 to illustrate a typical negotiation process.

From the previous discussion we know that for this example the equilibrium coalition size comprises seven countries in the case of the tax regime, and eight countries in the case of the quota regime. For each possible coalition size the countries which are in favor of a particular instrument are listed in the second and third last rows.

For example, if countries 9 and 10 form a coalition, country 9 prefers a quota regime ($-1730 > -1744$) and country 10 a tax regime ($-2081 < -2066$). In the case of such a stalemate, it is assumed that the preliminary coalition settles for the tax. This 'conservative' assumption seems appropriate because it turns out below that for most parameter constellations the quota regime will finally be chosen among the signatories. Generally, however, it is assumed that the majority decision rule (MDR) is applied with respect to this issue. The outcome of the decision process is indicated for each stage by the bold numbers and listed in the last row in Table 14.4.

However, in the preliminary coalition comprising countries 9 and 10 the instrumental issue is not really important because both countries would unanimously prefer one more country to join their 'club'. This is true, irrespective of the policy regime. Consequently, the coalition is enlarged. Only if six countries have joined the coalition does the instrumental choice become important: three countries prefer a tax and three a quota regime. If one more country joins, four signatories prefer a quota and three a tax regime. According to the MDR, signatories agree upon the quota. Then, the coalition is extended to eight countries of which all unanimously prefer the quota and the extension of the coalition ceases.

It is important to take account of the strategic linkages between issues. For instance, country 10 might think of blocking enlargement at an early stage of the formation process in order to influence the decision process in favor of the tax regime. In the example illustrated in Table 14.3, country 10 could veto an extension from six to seven countries in order to have a majority of countries favoring the tax. However, comparing payoffs to country 10 in a coalition of six countries applying a tax rate with a coalition of eight countries applying a quota, country 10 prefers a coalition of eight countries. Therefore, country 10 will not make use of its veto at the early stage of a coalition comprising six countries.

According to the procedure described above, the endogenous instrumental choice and coalition size has been determined for other parameter constellations. The results are presented in Table 14.4.

Table 14.4 Equilibrium size and instrumental choice of a sub-coalition for $\delta \rightarrow 1$[a]

N	γ		Coalition size	r^{SC} (%)	r^{GC} (%)	DO^{SC} (%)	DO^{GC} (%)	DO^{GC}_{max} (%)	Instrumental choice
	10	Q	4	22.9	19.6	65.1	64.0	97.4	T
		T	**4**	22.0	13.3	64.2	49.6	74.4	
	30	Q	**4**	11.9	10.1	61.4	60.3	98.9	Q
		T	4	9.9	5.7	54.9	39.5	92.6	
	50	Q	**4**	8.0	6.7	59.7	58.6	98.8	Q
		T	4	6.4	3.6	52.0	36.7	99.1	
	100	Q	**4**	4.4	3.7	58.0	56.9	98.7	Q
		T	4	3.4	1.9	49.4	34.2	100.0	
5	500	Q	**4**	0.9	0.8	56.4	55.3	98.7	Q
		T	4	0.7	0.4	47.0	32.1	100.0	
	1000	Q	**4**	0.5	0.4	56.1	55.0	98.6	Q
		T	4	0.4	0.2	46.6	31.8	100.0	
	2000	Q	**4**	0.2	0.2	56.0	54.9	98.6	Q
		T	4	0.2	0.1	46.5	31.7	100.0	
	3000	Q	**4**	0.2	0.1	56.0	54.9	98.6	Q
		T	4	0.1	0.1	46.4	31.6	100.0	
	30	Q	**8**	29.0	17.8	70.3	51.6	65.2	Q
		T	7	29.2	10.5	68.0	33.3	33.3	
	50	Q	**8**	21.7	12.8	66.2	46.9	64.2	Q
		T	8	20.2	7.5	63.4	29.8	40.1	
	100	Q	**8**	13.2	7.5	61.0	41.4	68.2	Q

		T	8	11.7	4.1	56.3	24.4	50.8	
10	500	Q	**8**	3.2	1.7	54.2	35.0	89.0	Q
		T	8	2.7	0.9	48.1	19.1	85.5	
	1000	Q	**8**	1.6	0.9	53.1	34.0	98.6	Q
		T	8	1.4	0.5	46.8	18.4	95.8	
	2000	Q	**8**	0.8	0.5	52.5	33.5	99.7	Q
		T	8	0.7	0.2	46.1	18.0	100.0	
	3000	Q	**8**	0.6	0.3	52.3	33.3	99.7	Q
		T	8	0.5	0.2	45.9	17.9	100.0	
	100	Q	**13**	34.4	13.0	71.4	36.3	36.3	Q
		T	12	23.7	4.5	55.5	13.8	13.8	
	500	Q	**15**	10.8	3.5	57.7	23.5	38.1	Q
		T	15	10.2	1.8	55.3	12.7	28.9	
20	1000	Q	**15**	5.8	1.8	53.8	20.9	47.1	Q
		T	15	5.4	0.9	51.1	11.1	38.4	
	2000	Q	**15**	3.0	0.9	51.5	19.6	58.3	Q
		T	15	2.8	0.5	48.6	10.2	51.3	
	3000	Q	**15**	2.0	0.6	50.7	19.1	64.2	Q
		T	15	1.9	0.3	47.8	9.9	59.2	

Note: [a] r^{SC} = global emission reduction of the sub-coalition; r^{GC} = global emission reduction of the grand coalition if country 1's proposal is applied, or, if this is not stable, the maximum stable emission reduction (see Table 14.1); DO^{SC} = degree of optimality of r^{SC}; DO^{GC} = degree of optimality of r^{GC}; DO^{GC}_{max} = degree of optimality of the grand coalition if r^{Q}_{max} or r^{T}_{max} is realized.

14.4.4 Results

In Table 14.4, the equilibrium number of signatories (if all three issues are decided simultaneously) has been marked bold in the column headed 'Coalition size'. The equilibrium instrumental choice resulting from the coalition formation process is listed in the last column.

From this table five main results can be derived:[13]

1. The relative coalition size ($\mu^* = N^*/N$) is not very sensitive to changes in the parameter values, but decreases in N and increases in γ. This implies that the number of signatories is lower for the 'critical' parameter values (N large and γ small).
2. The degree of optimality of the sub-coalition, DO^{SC}, decreases in γ.
3. The degree of optimality of the sub-coalition, DO^{SC}, is higher than that of the grand coalition, DO^{GC}, if the LCD decision rule is applied. The difference is particularly distinct for large N.
4. For the critical parameter values (N large and γ small) the degree of optimality of the sub-coalition, DO^{SC}, is higher than that of the grand coalition which chooses the maximal backable emission reduction target, DO^{GC}_{max} (for example, $N = 20$ and $\gamma = 100$).
5. In most cases the quota regime emerges from the decision process. Only if N is small may a tax regime be chosen among signatories (for example, $N = 5$ and $\gamma = 10$).

The first result confirms the findings of the conjectural variation models of Chapter 13 (see Propositions 13.2 and 13.3). The second, third and fourth results underline the rationale behind the formation of sub-coalitions. Even though the signatories are not concerned with global welfare but only with welfare in their countries, agreements based on compromises (LCD decision rule) might be more efficient if they were signed among a small group of relatively homogeneous countries instead of if a grand coalition is formed. In the light of possible high transaction costs, this argument receives even more momentum. In particular, for the critical parameter range of low values of γ and large N the formation of a sub-coalition seems particularly attractive. The rationale for the fifth result is the following: according to (14.1) a low value of N implies a low variance of damage between countries and therefore also a low variance of payoffs. Therefore, the allocation of payoffs from a joint abatement policy is also relatively even under the tax regime for small N. Consequently, a majority of countries votes for a tax in order to capture the efficiency gains associated with this instrument.

However for large N, the interests of countries are heterogeneous, which

leads to an uneven distribution of payoffs from emission reductions under a uniform tax. This in turn causes stability problems which can only partially be compensated by low reduction levels.

In contrast, under a quota regime abatement is based on initial emission levels (NE) which already reflect the preferences of countries to some extent. Hence, the abatement burden is allocated more in line with the interests of countries. This also allows higher abatement targets to be approved within an IEA and/or a higher stability.

Since there exists an incentive for the initiator governments (countries with a high index) that other countries contribute to a joint environmental policy and approve high abatement targets, a quota regime becomes particularly attractive if the global externality problem is distinct, which is the case if N is large.

14.5 SUMMARY AND DISCUSSION

Coalition formation has been studied in a supergame framework by applying the WRPE concept. The stability of a grand coalition and the formation process of a sub-coalition has been analyzed in order to answer the three questions posed in the Introduction:

1. *Abatement targets* of most IEAs are rather low since otherwise the free-rider incentive cannot be controlled by the signatories. It was shown that a grand coalition is very unlikely if signatories aim at a globally optimal abatement target or if the median country proposal of a uniform emission tax or a uniform emission reduction quota is applied within the agreement. However, even if more moderate abatement targets are laid down in the agreement, for example, the smallest tax or quota proposal among the N countries (LCD decision rule), stability is still a problem whenever an environmental problem is of a real global character (N large) and/or environmental damages are high compared to abatement costs (low γ).
2. Due to the above-mentioned stability problems, it is more likely that only some environmentally conscious countries would form a *sub-coalition*. For the initiator countries a smaller coalition may have an advantage in that more demanding abatement targets can be realized. Particularly for the critical parameter values (large N and small γ) it was shown that a sub-coalition may be more efficient than a grand coalition.
3. The analyses of the grand and the sub-coalitions suggested many explanations as to why the *quota* may be so popular in international

pollution control despite its postulated inefficiency in a first-best world: (a) except for some non-critical parameter values, the quota may lead to higher global welfare and lower emissions in a second-best world in which an agreement is reached according to the lowest common denominator. Some aspects of the rationale behind this result have been discussed already in Chapter 11 (cost–benefit versus cost-efficiency effect; see Section 11.6); (b) in this chapter it was shown additionally that it is usually easier for the grand coalition to sustain a quota than a tax agreement since the interests of those countries with the lowest interest in environmental protection are better accounted for and welfare among countries is more evenly distributed; (c) from almost all simulations the quota emerged as the final outcome of the endogenous decision process of a sub-group of environmentally conscious countries: only for environmental problems of regional character (N small) may signatories agree on a uniform emission tax.

The *advantage of the model* in this chapter is fourfold. First, internal stability of a grand or of a sub-coalition is checked within a consistent and plausible framework. Possible transitory gains from free-riding are accounted for and contrasted with the punishment options of the coalition members. Thereby, the difficulties of coordinating punishment and of restricted punishment options in international politics are integrated into the analysis. Thus, in contrast to the conjectural variation models, compliance is neither assumed in an *ad hoc* manner nor is it assumed that compliance is ensured through instant reactions by the punishers. The approach also contrasts with the core models which unrealistically assume that the coalition resolves once a country free-rides. Such threats are not credible in the sense of the WRPE concept.

Second, the definition of external stability is more convincing than that of the conjectural variation models. The initiator countries only allow other countries to join their (preliminary) IEA if all signatories agree.

Third, the present model departs from the (unrealistic) assumption of homogeneous countries and investigates stability for a variety of different abatement targets. In particular, the LCD decision rule accounts for the problems of agreeing on a joint abatement target at the pre-stage of the formation of an IEA. In contrast, almost all conjectural variation models (except Hoel 1992a) ignore this conflict and assume that signatories agree on abatement targets which maximizes aggregate welfare of the coalition members. Moreover, the present model accounts for the reservation of governments in the past to pay transfers whereas the conjectural variation (except Hoel 1992a) and the core models assume unlimited (though self-financed) transfers.

Fourth, the present model derives endogenously the choice of the policy instrument to achieve an abatement target. The model allows us to explain why a uniform emission quota has found such widespread application in many IEAs. In contrast, the conjectural variation and core models have nothing to say on this issue.

Nevertheless, the present model is far from perfect with respect to the theoretical requirements of an ideal coalition model as laid out in the Introduction of Chapter 13. First, in the tradition of all models assuming heterogeneous countries (except Bauer 1992), the analysis had to rely on simulations. However, this disadvantage was compensated by considering a large set of parameter values and various decision rules about how the abatement targets (seven different emission vectors), the coalition size (grand and sub-coalition) and the instrumental choice (UDR and MDR) are agreed upon in the coalition, which delivered very robust results.

Next, in the tradition of the conjectural variation models, only the existence of one but not of several sub-coalitions was considered.

Finally, the punishment options were restricted to the emission space. A more realistic assumption would be to extend the punishment strategy space to other policy fields in the tradition of issue linkage models. However, in the case of many signatories the question arises whether simultaneous membership of all signatories in two or more agreements is a realistic assumption (see Sub-section 13.2.7). Moreover, governments may also hesitate to punish a country using other policy issues since this may jeopardize the possibilities for cooperation on those issues too. Taking these qualifications into account the assumptions of the present model seem less restrictive and, in fact, may reflect the limited punishment options in international relations quite well. This is even more true when recalling that the present model allows us to take account of various limitations of punishment profiles (see Chapter 12, in particular Sub-section 12.2.4, and Subsection 14.3.4 of this chapter).

NOTES

1. This chapter draws heavily on Finus and Rundshagen (1998a).
2. See Chapter 8 for a justification of this assumption.
3. See Hoel (1992a) for a similar specification of the payoff function. Note that all qualitative results obtained below also hold if damages are not divided by N. Thus, $1/N$ is only a scaling factor. See Finus and Rundshagen (1997).
4. All results mentioned in this section are derived in Appendix XI.1.
5. The term second-best world with respect to result 4 is used because in a first-best world the stability problems of an agreement are usually neglected.
6. This definition implies that the probability that the game continues is assumed to be 1 (see Section 5.1). Note that $\delta_i \rightarrow 1$ is equivalent to $\rho_i \rightarrow 0$.
7. In the following, transaction costs are not modeled since their effects are rather obvious

(transaction costs increase as a function of the number of signatories); however, their impact should be kept in mind when interpreting the results below.

8. This is particularly true in a supergame in which the number of potential equilibria is large.
9. The described formation process over several stages makes it easier to grasp the intuition behind the determination of the equilibrium coalition size. However, strictly speaking, complete information would allow players to move directly to the final equilibrium in one step.
10. This result is apparent when observing Table 14.3. A general proof of this result may be obtained upon request.
11. Alternatively, one could also consider a majority decision rule (MDR) with respect to this issue. However, it turns out that the equilibrium coalition size is quite robust with respect to the decision rule. See Finus and Rundshagen (1997).
12. This result has been demonstrated in the case of the grand coalition. For the sub-coalition it is also true that the country with the lowest index, that is, $N_0 + 1$, prefers the quota rather than the tax regime. The proof is simple and follows by construction.
13. The results also hold if the coalition size is decided according to the MDR. See Finus and Rundshagen (1997).

15. Coalition models: a third approach

15.1 INTRODUCTION

A drawback of the coalition models discussed in the previous chapters is that they exogenously assume that there is a group of players which cooperate (signatories) and a second group of players which play as singletons (non-signatories). However, *a priori*, it is not clear whether equilibria in which several coalitions coexist can be excluded. Suppose such equilibria actually exist. Then the assumption of the previous models that an equilibrium coalition must only be immune to deviations by a single country may no longer be adequate. Then, a coalition structure should only be called an equilibrium if it is not challenged by any kind of deviation, regardless of whether a single country or a sub-group of countries deviates.

The aim of this chapter is to discuss some recent developments in the game theoretical literature on the formation of coalitions which allow for the coexistence of several coalitions. Most of these concepts have not yet been applied to the problem of international pollution control. Of those which have been applied to similar problems, only very general results have been derived, making it difficult to draw sound conclusions of practical relevance.[1] The subsequent discussion therefore pays particular attention to introducing the reader to the idea of these concepts and to evaluate them with respect to a possible application in future research. A simple example will illustrate the basic ideas of these concepts.

All subsequent concepts belong to the realm of *non-cooperative game theory*, though some of them have been developed from cooperative concepts. Hence, the classical approaches of cooperative game theory which are based on the characteristic function (or derivatives of it) have to be extended to account for the individual incentives of players in a coalition structure and to account for the *spillovers across coalitions* (that is, a best-reply strategy vector of a coalition depends on the strategies of the other coalitions). This extension is straightforward since all subsequent concepts assume the same *component game*, though they differ in how they model the coalition formation process. That is, all concepts are based on the assumption that once a particular coalition structure has formed (first stage), *coalition members cooperate among themselves and play a non-cooperative*

strategy against other coalitions (second stage). The idea corresponds to the Nash–Cournot assumption of the conjectural variation models encountered in Chapter 13. How these models differ is in the first stage, in which players decide on their membership. This issue will be discussed below.

In a non-cooperative game theoretical context a compact device to analyze the coalition formation process is the equilibrium valuation, also called a per-membership partition function. It is defined as follows (see also Bloch 1997):[2]

Definition 15.1: Equilibrium valuation or per-membership partition function

Let $c = \{c_1, ..., c_M\}$ denote a coalition structure with M coalitions, $c \in C$ where C denotes the set of coalition structures, $c_j \cap c_k = \varnothing$, $c_j \cup c_k \cup ... \cup c_M = I$; then an equilibrium valuation is a mapping which associates with each coalition structure $c \in C$ a vector of individual payoffs $\pi^*(c) = \{\pi_1^*(c_j, c), ..., \pi_N^*(c_k, c)\}$ where $\pi^*(c) \in \Pi^*(C)$ and where $\pi^*(c)$ is a set of payoffs which result from the maximization of players according to a particular rule, a given sharing rule of the gains from cooperation and a given coalition structure. The first argument in $\pi_i^*(c_j, c)$ refers to the particular coalition to which country i belongs, the second to a particular coalition structure.

First, note that the definition assumes away one of the original problems of endogenous coalition formation, namely how the gains from cooperation are divided among coalition members. That is, the equilibrium valuation assumes a *fixed sharing rule*. Though for symmetric countries the assumption of an equal sharing rule seems quite plausible (see the Introduction to Chapter 13), such an assumption (like any other assumption) is exogenous to the model.[3] However, since almost all general results which have been obtained so far by applying 'new' concepts to the problem of coalition formation rely on the assumptions of '*ex ante* symmetric players'[4] and an equal sharing rule, we shall make these assumptions for simplicity in this chapter, too.[5]

Second, note that a similar critique can be raised with respect to the assumption of the choice of the strategies within the coalition. Again, for symmetric countries it is plausible that the coalition members will maximize joint welfare (see the discussion in Chapter 13). It is also plausible that a coalition will play a Nash strategy toward outsiders since this implies that a Nash equilibrium is played in the final stage of the finite game which is a necessary prerequisite for a strategy to be subgame-perfect. However, as pointed out in Chapter 13, this implies either that players are assumed to abide by the rules of the coalition or that the instant reactions of players

do not allow players to reap a transitory free-rider gain. Therefore, this weakness should be kept in mind when interpreting the following results in Section 15.4.[6]

Third, note that due to the 'external' specification of the rules of the game, players' equilibrium valuations can solely be identified by the coalition structure.[7] This eases computation tremendously.

The key difference between the coalition models discussed in this chapter may be structured with respect to four aspects:

1. The first aspect is the *time dimension* of the coalition formation process. There are two basic assumptions: either the formation process is viewed as a one-shot game or it is seen as a dynamic process. That is, one can distinguish between games with (a) *simultaneous* choice of membership and (b) *sequential* choice of membership. In the latter case, the first stage of the game is divided into further sub-stages.
2. The second aspect concerns the membership. In *open-membership* games all players can freely accede to a coalition just by announcing that they will do so. The conjectural variation models are typical representatives already encountered in previous chapters. In *exclusive-membership* games the coalition members decide whether to allow other players to accede to their coalition or whether to merge with other coalitions. The formation of sub-coalitions in the supergame model in Chapter 14 is a representative of an exclusive membership game.
3. The third aspect concerns the *stability concept*. Of course, the exact definition depends on the time dimension of the coalition formation process and we shall be more specific on this issue below. For the simultaneous move membership game we consider the *perfectly coalition-proof Nash equilibrium* (CPNE), the *strongly subgame-perfect equilibrium* (SNE), some forms of *core stability* and the *largest consistent set.* These concepts are closely related to those discussed in previous chapters. For sequential move membership games stability concepts are tailored to the special assumptions of the formation process. We shall discuss the *equilibrium binding agreement* and the *sequential equilibrium*.
4. The fourth aspect concerns the *direction of spillovers* (Yi 1997).[8] One can distinguish positive and negative spillover (externality) games. Most generally, in positive externality games actions taken by a coalition increase *ceteris paribus* the payoff of outsiders. Typical games which belong to this category are coalition formation in Cournot-oligopoly to reduce output and in Bertrand-oligopoly to avoid excessive price underbidding so as to stabilize profits. Another instance is the

provision of a public good such as a 'clean environment'. Countries cooperating within an IEA reduce emissions compared to some status quo. In both instances, outsiders of the coalitions benefit from such kind of cooperation.[9]

In negative externality games actions taken by coalition members have a negative effect on outsiders. Typical examples include the provision of club goods. For instance, the gains from cooperation on R&D (which may materialize in a cut in production costs) accrue only to coalition members. Since this reduces the competitiveness of outsiders, such actions exhibit a negative externality on them. A similar instance is the formation of custom unions whereby tariffs within the union are abolished but tariffs against outsiders protect the union's market. By such actions goods produced by outsiders are partially replaced by goods produced by union members and the terms of trade are worsened from the outsiders' perspective.

Thus according to this categorization the *issue linkage games* mentioned in Chapter 13, Sub-section 13.2.7, combine both types of externalities. In the following we restrict our attention exclusively to positive externality games.

In what follows we characterize the per-member-partition function for positive externality games in Section 15.3. Results derived in this section will form the basis for the computations of the coalition structure in subsequent sections in which the various concepts are illustrated with the help of a simple example. In Section 15.3 we introduce the stability concepts of coalition-proof Nash equilibrium and strong Nash equilibrium. For simplicity this is done in a static context. The extension to the dynamic model framework will be indicated subsequently. In Section 15.4 simultaneous move coalition formation concepts will be discussed, and in Section 15.5 this is done assuming a sequential formation process.

15.2 CHARACTERIZATION OF THE PER-MEMBER-PARTITION FUNCTION FOR POSITIVE EXTERNALITY GAMES

In this section we characterize the conditions of the equilibrium valuation for *positive externality games* (see Bloch 1997; Yi 1997). These conditions are useful when analyzing a *global emission game* below in the context of various coalition formation concepts.

The following example assumes for simplicity payoff function (13.1), implying concave benefits and linear damages from emissions. Countries

are assumed *ex ante* symmetric. The *component game* is a two-stage game where in the first stage countries choose coalitions and in the second stage coalitions choose their emission levels. In the second stage coalitions maximize the payoffs of their members and play Nash–Cournot against other coalitions (see Section 13.1). Since actions taken by the players in the second stage are fixed and already known in stage one by backward induction, and since payoffs are received at the end of the second stage, the whole game can be analyzed in a reduced form of a single stage. The equilibrium valuation function $\pi^*(c) = \{\pi_1^*(c_j, c), ..., \pi_N^*(c_k, c)\}$ captures the relevant payoffs.

Four conditions characterize such a positive externality game (Bloch 1997; Yi 1997):[10]

$$C_1: \pi_i^*(c_i, c) < \pi_i^*(c_i, c') \text{ where } c_i \subset c, c' \text{ and } c' \backslash c_i \text{ can be derived from } c \backslash c_i \text{ by merging coalitions in } c \backslash c_i.$$

If coalitions merge to form larger coalitions, coalitions which are not involved in the merge are better off:

$$C_2: \pi_j^*(c_j, c) < \pi_i^*(c_i, c) \text{ iff } |c_i| < |c_j|.$$

Members of small coalitions enjoy a higher payoff than members of large coalitions for any given coalition structure.

The next two conditions deal with the effect on members of an old and a new coalition if a member leaves coalition i to join coalition j:

$$C_3: \pi_i^*(c_i, c) < \pi_i^*(c_i \backslash \{k\}, c') \text{ where } c' = c \backslash \{c_i, c_j\} \cup \{c_j \cup \{k\}, c_i \backslash \{k\}\} \text{ and } |c_j| \geq |c_i| \geq 2.$$

If a member of the coalition i leaves its coalition to join a larger or equal-sized coalition j, the members of the old coalition are better off.

$$C_4: \pi_k^*(c_i, c) > \pi_k^*(c_j \cup \{k\}, c') \text{ where } c' = c \backslash \{c_i, c_j\} \cup \{c_j \cup \{k\}, c_i \backslash \{k\}\} \text{ and } |c_j| \geq |c_i| \geq 2.$$

If a member k of coalition i leaves its coalition to join a larger or equal-sized coalition j, then the deviator becomes worse off.

Now we can show the following:

Proposition 15.1

The symmetric global emission game with payoff functions (13.1) in which the coalition maximizes the aggregate payoff of its members and

plays a non-cooperative strategy toward outsiders (Nash–Cournot assumption) satisfies conditions C_1–C_4.

Proof: Appendix XII.1. QED[11]

Moreover, it turns out that in the subsequent analysis in Sections 15.4 and 15.5 the following definition is useful:

Definition 15.2: Stand-alone stable
$c = \{c_1, ..., c_M\}$ is stand-alone stable iff $\pi_i(c_i, c) \geq \pi_i(\{i\}, c')$ where $c' = c \backslash c_i \cup \{c_i \backslash \{i\}, \{i\}\}$ $\forall$ $i \in$ I.

A coalition structure c is *stand-alone stable* if and only if no player finds it profitable to leave its coalition to be a singleton, holding the rest of the coalition structure constant. From this it follows immediately that, by definition, the degenerate coalition structure consisting only of singletons is stand-alone stable. Moreover, it is straightforward to establish the following result:

Proposition 15.2
In the symmetric global emission game with payoff functions (13.1) in which the coalition maximizes the aggregate payoff of its members and plays a non-cooperative strategy toward outsiders (Nash–Cournot assumption) the largest coalition in a coalition structure which is stand-alone stable comprises no more than three coalition members.

Proof: Is straightforward and therefore omitted. QED

Conditions C_1–C_4, Proposition 15.2 and two additional conditions which are derived in Appendix XII.1 are used in Appendix XII.2 to derive the preference profile for the example based on payoff function (13.1). The preference profile is derived for $N \in [3, 4, 5]$. It forms the basis for deriving the coalition structure for all coalition formation concepts discussed below.

15.3 STATIC GAMES

In this section we want to introduce the stability concept of a strong Nash equilibrium (SNE) and a coalition-proof Nash equilibrium (CPNE) to which we refer in Section 15.4. This is most easily done in a static context. At the end of this section we shall indicate how the definition must be extended to the component game.

Both stability concepts require that a coalition structure is only called

stable if no *single player* has an incentive to change his/her strategy. This corresponds to the notion of a Nash equilibrium (NE). Thus, as the name indicates, any SNE and CPNE must be an NE, too. In particular, consider the following definition of an SNE (see also Ordeshook 1986, pp. 304ff.).

Definition 15.3: Strong Nash Equilibrium (SNE) (Aumann 1959)
A strong Nash equilibrium is a strategy profile s^* for which no coalition $c_i \subseteq \mathrm{I}$ can increase its payoff by changing its strategy. That is, $\pi_i(s^*) \geq \pi_i(s_j, s_{-j})\ \forall\ c_j \subseteq \mathrm{I}$ and $i \in \mathrm{I}$ where s_j is coalition j's strategy vector and s_{-j} the strategy vector of players in $\mathrm{C} \backslash c_j$.

From the definition it follows that the payoff vector of an SNE must be Pareto-efficient with respect to the *entire* payoff space since otherwise a sub-group of countries or the coalition as a whole would jointly alter their strategies. From this it is immediately evident that no SNE exists in a standard type of global emission game. This is so since the Nash equilibrium (for example, payoff function of types 1 and 3) or Nash equilibria for example, payoff function of type 2 in the case of symmetric countries) are inefficient in positive externality games (as defined in this chapter).

As has been argued already in the two-player context (see Chapters 6 and 7), requiring equilibrium strategies to be *Pareto-efficient* with respect to the entire strategy space may be an unnecessarily restrictive condition. In the N-player context, the SNE concept is fraught with an additional weakness. Deviation by a sub-group of countries is deemed to be feasible even though the deviation itself may be subject to further deviations which may lead to an outcome which is Pareto-dominated by the original agreement. It is this very restrictive definition of a stable equilibrium that is responsible for no SNE existing in many games.

The CPNE concept takes up this concern and rules out such non-credible deviations. It allows an equilibrium to be challenged only by self-enforcing deviations. That is, efficiency is only required with respect to the set of *self-enforcing* agreements.[12]

The basic idea of a CPNE may be illustrated with the following example (Bernheim *et al.* 1987). Suppose there are four players in a room who agree on some strategy vector. Then one player, say 4, leaves the room and hence cannot change his/her strategy any more. Given player 4's strategy, the three remaining players may now think about altering their strategies. Again, one player leaves the room after the agreement and cannot change his/her strategy any more. The process continues until one player is left in the room. The original agreement among the four players is a CPNE provided that, at any stage of the process, no sub-group of players left in the room wishes to

switch to another strategy vector, regardless of the order in which players leave the room.

Such an agreement can be found by *backward induction*. The last player left in the room has no incentive to deviate if his/her last agreement had the property of best response. The last two players left in the game must play a Pareto-undominated NE in the game induced on them by the strategies of the players who left the room. The induction process continues through the total number of players.

Definition 15.4: Coalition-proof Nash equilibrium (CPNE) (Bernheim *et al.* 1987)
Let Γ/s_{-i} denote the game induced on sub-group c_i by strategy vector s_{-i}. (a) Then if $N=1$, $s^* \in S$ is a coalition-proof Nash equilibrium iff s^* maximizes $\pi_i(s)$. (b) Let $N>1$ and assume that coalition-proof Nash equilibria have been defined for all games with $n<N$ players. Then (i) for any game Γ with N players s^* is self-enforcing if, for all $c_i \subset I$, s_i^* is the coalition-proof equilibrium in the game Γ/s^*_{-i}, and (ii) s^* is a coalition-proof Nash equilibrium if it is self-enforcing such that there does not exist another self-enforcing strategy s' for which $\pi_i(s') > \pi_i(s^*)\ \forall\ c_i \subset I$.

Thus every sub-group of players plays a CPNE against the other players. From the definition it is evident that every SNE is a CPNE, too; however, the opposite does not hold. Hence, $S^{SNE} \subseteq S^{CPNE} \subseteq S^N$.[13,14]

It is important to note that the CPNE imposes not only a restriction on the deviation strategies (which must be self-enforcing) but also on the set of players who deviate. Only among the players who originally deviated are further deviations considered. That is, if players 1 and 2 deviate, only these players may further deviate, but not players 1 and 3 for instance. Though this a conceptual weakness of a CPNE, it reduces the number of deviations to consider and therefore eases computations.

In a (static) global emission game the NE is also a CPNE. By the definition of an NE, no single country wants to deviate. Moreover, any sub-group of players which jointly deviated would again be subject to a further deviation. This follows simply from the fact that any such deviation implies an emission tuple which does not simultaneously lie on the reaction function of all deviators, implying an incentive to further deviations by at least one player.

To highlight the difference between the CPNE and the SNE concept consider the following simple matrix game. Each of three countries has two actions a_i and na_i where country 1 chooses rows, country 2 columns and country 3 matrices. In this game countries 1 and 2 may be small countries whose interests more or less coincide, whereas country 3 is a big country.

Matrix 15.1 could represent a modified chicken-assurance game as laid out in Section 3.5. Action a_i could represent 'invention of catalytic converters (cats)' for cars, action na_i could represent the status quo (see Section 3.4). If the big country does not make cats mandatory (na_3), it is a best strategy for countries 1 and 2 to make cats mandatory (a_1, a_2).[15] (a_1, a_2, na_3) is an NE since country 3 has *no* interest in changing its strategy (given the strategies of countries 1 and 2) to a_3. Similarly, if big country 3 introduces cats, it is best for countries 1 and 2 to produce cars without cats. (na_1, na_2, a_3) is the second NE in this game.

Matrix 15.1 Three-player abatement game with two Nash equilibria, one coalition-proof Nash equilibrium and no strong Nash equilibrium

	a_3	
	a_2	na_2
a_1	3, 3, 9	0, 1, 4
na_1	1, 0, 4	*4*, *4*, *6*

	na_3	
	a_2	na_2
a_1	**5**, **5**, **10**	0, 1, 2
na_1	1, 0, 2	6, 6, 0

From the payoff structure it is evident that $\pi_i(a_1, a_2, na_3) > \pi_i(na_1, na_2, a_3)$ $\forall\, i \in I$ and therefore one should expect at first thought that countries would play (a_1, a_2, na_3). However, this NE is upset if countries 1 and 2 coordinate their strategies and jointly deviate to (na_1, na_2). It is evident from Matrix 15.1 that no further deviations will occur from (na_1, na_2, na_3) and hence this deviation is a credible challenge to the NE (a_1, a_2, na_3).[16] Consequently, (a_1, a_2, na_3) is not a coalition-proof equilibrium. In contrast, the (Pareto-inferior) NE (na_1, na_2, a_3) is coalition-proof since no group of two players has an incentive to deviate jointly. Also the group of all countries will not deviate, though it has an immediate incentive to move to the Pareto-superior NE, since this deviation would be subject to a further deviation to (na_1, na_2, na_3). Since the payoff vector resulting from the play of (na_1, na_2, na_3) does not Pareto-dominate the NE (na_1, na_2, a_3), the deviation of the group as a whole is not credible in the sense of a CPNE.

Note that in this game no SNE exists. Only the Pareto-efficient NE (a_1, a_2, na_3) qualifies as a potential candidate; this, however, is subject to a joint deviation by players 1 and 2, as argued above.

We saw above that the two-stage coalition formation game may be analyzed in a reduced (one-stage) form (since payoffs are received at the end of the second stage and equilibrium strategies of the second stage are fixed and known in the first stage). Thus, for our purposes, the definition of an

SNE and a CPNE in a dynamic setting is not necessary.[17] Only in Definitions 15.3 and 15.4 do strategies s_i, s_{-i} and s have to be replaced by c_i, $c\backslash c_i$ and c and $\pi_i(...)$ by the equilibrium valuation $\pi_i^*(...)$.

15.4 COALITION FORMATION MODELS: SIMULTANEOUS MOVES

In this section we look at coalition formation models where countries choose their coalition membership simultaneously. The incentive profile to derive the subsequent results is provided in Appendix XII.2.

15.4.1 Open-membership Game

In the open-membership game of Yi and Shin (1995) players can form coalitions freely as long as no outsider is excluded from joining a coalition (see also Yi 1996, 1997). Each player's strategy space is a message space M where $|M| \geq N$. All players simultaneously announce a message m_i (or in the words of Yi and Shin they 'announce an address'). Players which have announced the same message form a coalition. For instance, if $N=4$ and $m_1=m_2=m_3=1$ and $m_4=2$, $\{\{1, 2, 3\},\{4\}\}$ forms. However, this coalition structure is no NE since country 3 has (given the announcements of the other countries) an incentive to announce $m_3=2$ instead. Though country 4 is better off in the coalition structure $\{\{1, 2, 3\},\{4\}\}$ than in the coalition structure $\{\{1, 2\},\{3, 4\}\}$, it cannot deny country 3 accession to its singleton coalition according to the open-membership rule.

Table 15.1 summarizes the equilibrium coalition structures for $N=3$, $N=4$ and $N=5$ if the CPNE and the SNE concepts are applied. Appendix XII.3 provides the details of the derivation of these results.

With respect to the CPNE concept it is interesting to note that the equilibrium coalition structure implies the coexistence of several coalitions. Though Table 15.1 covers only $N \in [2, 5]$, for large N one finds that more than two coalitions coexist. For instance, for $N=9$ three coalitions of three countries form. Except for $N=3$ the grand coalition is never a CPNE. The reason is simple. Any CPNE must be an NE too. Any NE under the open-membership rule must be immune to deviations by a single country. In other words, an equilibrium coalition structure must be stand-alone stable according to this rule. Since by Proposition 15.2 the largest coalition in a stand-alone coalition structure is 3, one can conclude that *not only* for $N=4$ and $N=5$ will no grand coalition form in the emission game but also not for $N>5$. In fact, one can conjecture that coalition structures are rather symmetric for large N. Consider – with a slight abuse of notation –

Table 15.1 Equilibrium coalition structures of simultaneous move coalition formation games[a]

Game	Simultaneous move coalition formation games: Coalition-proof equilibrium	Simultaneous move coalition formation games: Strong Nash equilibrium
Open-membership game	$N = 3$: {{1, 2, 3}}, {{1, 2}, {3}} $N = 4$: {{1, 2}, {3, 4}} $N = 5$: {{1, 2, 3}, {4, 5}}	$N = 3$: {{1, 2, 3}}, {{1, 2}, {3}} $N = 4$: no equilibrium $N = 5$: no equilibrium
Exclusive membership Γ game	$N = 3$: {{1, 2, 3}}, {{1, 2}, {3}} $N = 4$: {{1, 2, 3, 4}}, {{1, 2, 3},{4}} $N = 5$: {{1, 2, 3, 4, 5}}, {{1, 2, 3, 4}, {5}}	$N = 3$: {{1, 2, 3}}, {{1, 2}, {3}} $N = 4$: {{1, 2, 3, 4}}, {{1, 2, 3},{4}} $N = 5$: {{1, 2, 3, 4, 5}}, {{1, 2, 3, 4}, {5}}
Exclusive membership Δ game	$N = 3$: {{1, 2, 3}}, {{1, 2}, {3}} $N = 4$: {{1, 2, 3},{4}}, {{1, 2}, {3, 4}} $N = 5$: {{1, 2, 3}, {4, 5}}, {{1, 2}, {3, 4}, {5}}	$N = 3$: {{1, 2, 3}}, {{1, 2}, {3}} $N = 4$: {{1, 2, 3},{4}} $N = 5$: no equilibrium
Core stability	$N = 3$: {{1, 2, 3}}, {{1, 2}, {3}} $N = 4$: no equilibrium $N = 5$: no equilibrium	
α stability	$N = 3$: {{1, 2, 3}}, {{1, 2}, {3}} $N = 4$: {{1, 2, 3, 4}}, {{1, 2, 3} {4}} $N = 5$: {{1, 2, 3, 4, 5}}, {{1, 2, 3, 4}, {5}}	
β stability	$N = 3$: {{1, 2, 3}}, {{1, 2}, {3}} $N = 4$: {{1, 2, 3, 4}}, {{1, 2, 3} {4}} $N = 5$: {{1, 2, 3, 4, 5}}, {{1, 2, 3, 4}, {5}}	
Farsighted equilibrium	$N = 3$: {{1, 2, 3}}, {{1, 2}, {3}} $N = 4$: {{1, 2, 3}, {4}}, {{1, 2}, {3, 4}} $N = 5$: {{1, 2, 3, 4, 5}}	

Note: [a] Assumption: payoff function (13.1) and symmetric countries.

the following coalition structure $c = \{|c_1|, |c_2|, ..., |c_M|\}$, $|c_1| \geq |c_2| \geq ... \geq |c_M|$ with $|c_1| \geq |c_M| + 2$. According to condition C_4 above, a member of c_1 has an incentive to leave its coalition to join a smaller coalition. Thus, $|c_1| \leq |c_M| + 1$ can only be an equilibrium coalition structure.[18] Consequently, ignoring integer constraints, $\{\{N/2\},\{N/2\}\},\{\{N/3\},\{N/3\},\{N/3\}\}$ etc. are potential equilibrium NE coalition structures. Of course, coalition structures involving smaller coalitions are also stand-alone stable. For instance, for $N = 4$ $\{\{1, 2\},\{3\},\{4\}\}$ or $\{\{1\},\{2\},\{3\},\{4\}\}$ are stand-alone stable. However, these coalition structures are Pareto-dominated by larger coalitions (though $|c_i| \leq 3 \; \forall \; i \in M$), (for example, $\{\{1, 2\},\{3\},\{4\}\}$ is dominated by $\{\{1, 2\},\{3, 4\}\}$), so that it can be expected that for $N \geq 4$ the CPNE coalition structure(s) involve(s) the coarsest coalition structure with two or three countries in a coalition.

From the discussion it comes as no surprise that for $N \geq 4$ no SNE exists since $|c_i| \leq 3 \; \forall \; i \in M$ implies a Pareto-inferior coalition structure. The result stresses once more that the SNE concept may be too strong in that no outcome can be predicted:

Proposition 15.2
In the symmetric global emission game with payoff functions (13.1) and where the coalition maximizes the aggregate payoff of its members and plays a non-cooperative strategy toward outsiders (Nash–Cournot assumption) a CPNE coalition structure implies for any $N \geq 5$ (a) $[N/3]$ coalitions of size 3 if $R = 0$; (b) $[N/3]$ coalitions of size 3 and one coalition of size 2 if $R = 2$; or (c) $[N/3] - 1$ coalitions of size 3 and two coalitions of size 2 if $R = 1$ where $R = N - k \cdot 3$, $k = [N/3]$. This equilibrium is unique. There is no SNE equilibrium coalition structure for $N \geq 4$.

Proof: Since the proof is straightforward using results of the discussion above, conditions C_1–C_4, Proposition 15.2 and results provided in Appendix XII.2, it is omitted here. QED

An obvious disadvantage of all simultaneous move membership games, and therefore also of the open-membership game, is that there is no story of how a particular coalition structure evolves. That is, there is no explanation as to how countries coordinate on a particular equilibrium.

A particular disadvantage of open-membership games is the assumption that each country can freely join any coalition. That is, in the language of the conjectural variation models encountered in Chapter 13, an equilibrium coalition structure must be externally stable. As already mentioned in Sub-section 13.2.8, in a global emission context there is no obvious reason why a coalition cannot restrict membership if the accession of outsiders

implies a welfare loss. In other games the assumption of open membership can be more easily justified. For instance, in international trade the accession to GATT/WTO is generally open to all countries if they abide by the rules. Nevertheless, in reality there are many bilateral or multilateral trade agreements which favor only members and where the members put up high barriers to entry for outsiders. An obvious instance is the current discussion about the enlargement of the European Union to include Eastern European countries.

In particular, in issue linkage games it may be more effective to restrict membership (see Sub-section 13.2.8). If too many members accede to a linked agreement, the gains accruing from the club good agreement may become too small, so that the public good part of the agreement cannot be stabilized any more.

One advantage of the open-membership game is that if the CPNE stability concept is applied there is a unique equilibrium for $N \geq 4$. Of course, this result rests on the assumption of symmetric countries and payoff function (13.1). In fact, it seems very likely that this result does not hold in more general environments.

Another advantage of the open-membership game is that the results for the CPNE concept correspond to intuition. Due to the free-rider incentive, there is no grand coalition for $N \geq 4$. Moreover, the coexistence of several coalitions also seems plausible considering the fact that it may be easier to reach an agreement among a smaller group of countries than among all countries. However, due to the restriction to symmetric countries, final conclusions must be drawn with caution.[19]

15.4.2 Exclusive Membership Games

Γ games

In Γ games all players announce which coalition they want to belong to, including the list of all participants of this coalition (Hart and Kurz 1983). In order for a coalition to form, the *unanimous agreement* of all players is required. More precisely, in the Γ game the message space of country i is a set of coalitions to which it wants to belong $\mathrm{M}_i = \{\mathrm{C}_i \subset N, i \in \mathrm{C}_i, c_i \subseteq \mathrm{C}_i\}$. A coalition c_i forms if and only if all members i of c_i have announced $m_i = c_i$. This definition implies that, whenever a member deviates and leaves the coalition, the whole coalition breaks apart. Due to this assumption, the set of NE is very large (for example, for $N = 5$ it comprises all permutations, see Appendix XII.4). For instance, if each country announces a singleton coalition, this announcement is immune to a deviation. This is so since an NE considers only single deviations, and forming any other coalitions requires a change of at least two messages. Thus, in Γ games the equilibrium

refinements SNE and CPNE are particularly useful. Nevertheless, Table 15.1 reveals that a single coalition structure can hardly be expected in the example. From the table it appears that an equilibrium SNE or CPNE coalition structure involves typically either the grand coalition or a coalition structure with one coalition of $N-1$ countries and a singleton coalition. The reason for such large coalitions is the assumption that a coalition breaks apart into singletons once a player or group of players deviates. This assumption has already been criticized in the context of the γ core as a not particularly credible threat.

Δ games

In Δ games this last-mentioned concern is taken up (Hart and Kurz 1983). As in Γ games, all members announce a coalition to which they want to belong and a list of coalition members. In contrast to Γ games, however, in Δ games the participation of all listed members is not needed for a coalition to form. Thus, as in the open-membership game, the message space is used as a coordination device (Bloch 1997, pp. 322ff.). However, membership is exclusive because only coalition structures which have been announced can eventually form. The definition implies that, whenever a member leaves the coalition, it assumes that all other coalition members remain together (as in the conjectural variation models in Section 13.2).

For instance, consider $N=4$. Now, in contrast to the open-membership game, $\{\{1, 2, 3\},\{4\}\}$ is an NE-coalition structure. Though any of the coalition members in the larger coalition would like to join the singleton coalition, this is not in 4's interest. Since unanimous agreement is required by the exclusive membership rule, and 4 will never propose $\{\{1, 2\},\{3, 4\}\}$, $\{\{1, 2, 3\}, \{4\}\}$ is stable.

From Table 15.1 it is evident that according to the CPNE concept more than one equilibrium coalition structure exists. (See Appendix XII.5 for a derivation of the equilibria.) For $N\geq 4$ the grand coalition is no CPNE and the coalition structure comprises several smaller coalitions. As in the open-membership game, the reason is that any CPNE in a Δ game must be an NE and an NE must be stand-alone stable. Thus, no coalition structure will be an equilibrium involving coalitions of size greater than 3.

Obviously, the coalition structure in the Δ game allows for finer equilibrium coalition structures than the open-membership game. The reason is the following. Consider – with a slight abuse of notation – the coalition structure $c=\{2, 2, ..., 1\}$ where the numbers indicate the size of each coalition. Then any of the coalitions of size 2 would like to merge with the singleton coalition. Since, however, the singleton member is indifferent to such a move, the coalition structure above is a CPNE in the Δ emission game. In contrast, in the open-membership game such a merger could be enforced.

Again, from $|c_i| \leq 3$ it follows that for $N \geq 5$ no SNE exists, confirming our reservation against the SNE concept.

Thus, taken together, the coalition formation process in Δ games seems quite plausible. Coalitions are not resolved once a country deviates and exclusive membership seems a plausible assumption in the global emission context.[20]

15.4.3 Core-stable Coalition Structures

The concepts discussed in this sub-section are interesting in that they combine the stability concept of the core, as laid out in Section 13.3, and the conjectural variation setting of Section 13.2. This setting is constituted by the component game which has been introduced in Section 15.1. There are two extreme assumptions. Shenoy (1979) proposes a concept which he calls *core stability*. The concept assumes that, following a deviation, other players react as if they were maximizing the payoff of the deviating coalition. That is, a deviating group of players supposes that the remaining players support their deviation by forming an optimal coalition structure. This is a very optimistic assumption. In contrast, Hart and Kurz (1983) propose α and β stability where a deviating coalition expects the worst from external players. This is a very pessimistic assumption. As in the α and β core in Section 13.3, an α-stable coalition structure is a partition for which there is no group of players who could obtain a higher payoff irrespective of the behavior of the external players. In a β-stable coalition structure, there is no group of players who can expect to obtain a higher payoff whatever the reaction of the external players may be (see Bloch 1997, pp. 330ff.):

Definition 15.5: Core-stable coalition structure
A coalition structure c is core-stable if there is no group of n players and a coalition structure $c' \supset n$ such that for all $i \in n$ $\pi_i^*(c') > \pi_i^*(c)$ holds.

Definition 15.6: α-stable coalition structure
A coalition structure c is α-stable if there is no group of n players and a partition c' such that for all partitions $c_{N/n}$ formed by external players $\pi_i^*(c' \cup c_{N/n}) > \pi_i^*(c)$ holds for all $i \in n$.

Definition 15.7: β-stable coalition structure
A coalition structure c is β-stable if there is no group of n players such that, for all partitions $c_{N/n}$ of external players, there exists a partition c_n of n such that for all $i \in n$ $\pi_i^*(c_n \cup c_{N/n}) \geq \pi_i^*(c)$ holds.

From the definitions it is evident that, in general, deviations occur more frequently under core stability, than under β stability and under β stability more often than under α stability. Consequently, if C^C, C^α and C^β denote the set of core-, α- and β-stable coalition structures, $C^C \subseteq C^\beta \subseteq C^\alpha$ holds (Kurz 1988). In fact, it is easily checked for the global emission game example that $C^\beta = C^\alpha$ holds.

As it is evident from Table 15.1, a core-stable coalition structure for $N \geq 4$ does not exist. (See Appendix XII.6 for a derivation of the equilibria.) This is not surprising since, given the optimistic conjectures about external players, there is always a coalition structure for which it is beneficial for a group of players to deviate. As under the SNE concept no restriction is imposed on such a deviation to be immune to further deviations.

Under α and β stability either the grand coalition or one large coalition and a singleton coalition is stable. The reason is that any deviation triggers further deviations, leading to small partitions which are Pareto-dominated. Again, no restriction is imposed on such a deviation to be immune to further deviations, which allows us to stabilize large coalitions.

Thus, an obvious weakness of all three stability concepts is the *ad hoc* assumption regarding the behavior of external players. Neither the optimistic assumption under core stability nor the pessimistic assumption under α and β stability is usually a best reply for external players. Thus, these behavioral assumptions are rather unrealistic and confirm our reservations against the core concept, as has been raised already in Section 13.3.

15.4.4 Farsighted Coalitional Stability

As the term 'farsighted coalitional stability' (FCS) indicates, this concept is designed to take account fully of all reactions following an action by a group of players. According to Chwe (1994), the advantage of this concept is that it remedies some of the deficiencies of other concepts. In contrast to core stability, an equilibrium exists in most environments. Moreover, reactions by external players to a deviation by a group of players is defined more consistently. In contrast to the SNE concept, a deviation is only accepted if it improves upon the payoffs of the deviating players and if and only if all possible further deviations are taken into account. In contrast to the CPNE concept, where only deviations which are self-enforcing are valid, the FCS also considers the possibility that some players intentionally deviate to a coalition structure which is subject to further deviations since they are interested in the final outcome. A set of coalition structures which is immune to such deviations is called the *largest consistent set* and is denoted by C^{FCS}. It is defined as follows:[21]

Definition 15.8: Largest consistent set
A set $C^{LCS} \subseteq C$ is consistent if no coalition c_i or a member of a coalition c_i has an incentive to move from the coalition structure c to another coalition structure $c' \in C$ since c' is not strictly preferred by all deviators or c' is indirectly dominated by another coalition structure c''.

It is important to note that at each stage of the deviation process, all players who currently deviate must prefer the new coalition structure. It does not suffice if the original members would like a further deviation, but not all the members of the present coalition of which the original deviators are members.

As laid out in Appendix XII.7, to determine the largest consistent set, first, a list of all possible permutations is needed. Second, a preference relation between different coalition structures must be established. Third, a coalition structure which is not obviously Pareto-dominated by another coalition structure is picked and any possible deviation and 'chain deviations' are tested for.[22]

Obviously, the farsighted coalitional stability concept introduces some interesting strategic aspects into the coalition formation process which are absent from the concepts discussed so far. For instance, consider $N = 5$, for which the grand coalition is stable (see Table 15.1). Since the grand coalition is not stand-alone stable, a country, say 5, has an incentive to leave the grand coalition, that is, $\{\{1, 2, 3, 4, 5\}\} \Rightarrow \{\{1, 2, 3, 4\}, \{5\}\}$, which is subject to a further deviation, say by country 4, that is, $\{\{1, 2, 3, 4\}, \{5\}\} \Rightarrow \{\{1, 2, 3\}, \{4\}, \{5\}\}$. Since $\{\{1, 2, 3\}, \{4, 5\}\}$ is strictly preferred by all players to $\{\{1, 2, 3\}, \{4\}, \{5\}\}$, this would be, potentially, the final coalition structure. However, since $\{\{1, 2, 3, 4, 5\}\}$ strictly Pareto-dominates $\{\{1, 2, 3\}, \{4, 5\}\}$, this chain deviation is not valid and therefore the grand coalition is consistent.

In contrast, for $N = 4$, the grand coalition is not consistent since $\{\{1, 2, 3, 4\}\} \Rightarrow \{\{1, 2, 3\}, \{4\}\}$ and $\{\{1, 2, 3\}, \{4\}\}$ is not dominated by any other coalition structure. $\{\{1, 2\}, \{3, 4\}\}$ is also stable. Though $\{\{1, 2\}, \{3, 4\}\}$ is directly dominated by $\{\{1, 2, 3, 4\}\}$, $\{\{1, 2, 3, 4\}\}$ is dominated by $\{\{1, 2, 3\}, \{4\}\}$ and hence for $\{\{1, 2\}, \{3, 4\}\}$ to be *not* consistent it must be indirectly dominated by $\{\{1, 2, 3\}, \{4\}\}$. This is, however, not the case since not all countries (here it is country 3) forming a grand coalition from $\{\{1, 2\}, \{3, 4\}\}$ benefit from coalition structure $\{\{1, 2, 3\}, \{4\}\}$.

Note that $\{\{1, 2, 3\}, \{4\}\}$ is only consistent by the definition of a *strict* Pareto-dominance. Since country 3 would prefer to join coalition $\{4\}$ one could think of the following scenario: $\{\{1, 2, 3\}, \{4\}\} \Rightarrow \{\{1, 2\}, \{3\}, \{4\}\} \Rightarrow \{\{1,2\}, \{3,4\}\}$, where the final coalition structure is stable according to the discussion above. That is, country 3 leaves the coalition to become a

singleton and thereby creates an incentive for country 4 to accept its membership. Since the first move leaves the payoff to country 3 unchanged (though the second move implies an increase in its payoff), this chain deviation is ruled out by Chwe's definition.

On the one hand, requiring only weak dominance would produce a unique equilibrium in the example, and therefore the above definition seems unnecessarily restrictive. On the other hand, this definition avoids cyclical deviations in other examples and may therefore be justified.

Taken together, the farsighted coalitional stability concept appears the most advanced among the simultaneous move concepts. It takes the strategic considerations of players fully into account. It therefore satisfies the condition of a rational conjectural variation equilibrium at *every* state (see Chapter 10, Definition 10.1), and may therefore be regarded as a convincing static representation of dynamic games. It is a further development of the farsighted concept mentioned in Sub-section 13.2.8 in that the coexistence of several coalitions is in addition considered. The strict Pareto-dominance at each stage of a deviation process implies *de facto* exclusive membership, for which we have indicated some sympathy already. In contrast to the Γ and Δ games, the FCS neither assumes in an *ad hoc* fashion that the coalition breaks apart nor that the remaining coalition structure is preserved following a deviation but that the external players play a best-reply strategy. Thus, reactions to a deviation are consistently modeled. Therefore, it seems promising to apply the FCS concept to a more general model of coalition formation (such as payoff function (13.3) and heterogeneous countries and so on) in future research.[23]

15.5 SEQUENTIAL MOVE COALITION MODELS

15.5.1 Equilibrium Binding Agreements

Ray and Vohra (1997) motivate their *equilibrium binding agreement* (EBA) with the help of the following story behind a coalition formation process. Initially the grand coalition gathers. Then some leading perpetrators may propose a different coalition structure if this is in their interest. The perpetrators split up to form a coalition, say c_i, by themselves. In a next step, either the coalition c_i or the coalition $C \backslash c_i$ may be subject to further deviations. Those countries which initiate further deviations are called secondary perpetrators. The process of disintegration continues until a coalition structure has been reached where no country wishes to split up into finer partitions. From this brief introduction it is already apparent that the term equilibrium binding agreement is a misnomer: it implies that the EBA

concept belongs to the realm of cooperative game theory. However, except for the weakness which is pertinent to all models discussed in this chapter, namely that stability within the component game is either assumed *ad hoc* or via the assumption of instant reactions by players, the EBA concept clearly belongs to non-cooperative game theory since an EBA must be a self-enforcing coalition structure.

Important for understanding the concept is the assumption that coalitions can only become finer but not coarser and that only members of a coalition can form smaller coalitions but this is not possible across coalitions.

Similar to the FCS concept, countries will not deviate from a given coalition structure if the final outcome following subsequent deviations implies a payoff loss to them. Leading perpetrators and external players will further deviate if it is in their interest to do so. Thus the reaction of players is consistently defined.

The formal definition of equilibrium binding agreements (like that of a CPNE) is recursive:[24]

Definition 15.9: Equilibrium binding agreement (EBA)

1. The finest coalition structure consisting of singletons is an equilibrium binding agreement.
2. Suppose that equilibrium binding agreements have been defined for all partitions finer than a coalition structure c and let the set of these finer coalition structures be denoted by $C^{EBA(c)}$. Then c is an EBA if no player or group of players wishes to deviate to any $c' \in C^{EBA(c)}$. That is, $\pi_i^*(c_j, c) \geq \pi_i^*(c'_j, c')\ \forall\ i \in c_i,\ c_i \subset c$ and $c' \in C^{EBA(c)}$ where $c'_j \subset c_j$ and $c' = \{c'_j\} \cup \{c'_{-j}\}$.

Due to the recursive definition and no further selection criteria, the set of EBA is very large. For instance, for $N = 4$ $\{\{1\},\{2\},\{3\},\{4\}\}$ is stable by definition. $\{\{1, 2\},\{3\},\{4\}\}$ is stable since it Pareto-dominates the finer partition $\{\{1\},\{2\},\{3\},\{4\}\}$ which is stable. $\{\{1\},\{2, 3, 4\}\}$ is stable since it Pareto-dominates $\{\{1\},\{2\},\{3\},\{4\}\}$ and so on. By continuing this recursive process, it is evident that the set of EBA is quite large. Thus, it seems sensible to introduce a selection device to reduce the number of equilibria. An obvious criterion follows by recalling the story at the beginning of this sub-section to illustrate the EBA concept. By assuming that the coalition formation process evolves from the grand coalition, fine partitions are Pareto-dominated by coarser ones.[25] Therefore, in Table 15.2 only $C^{EBA*} = \text{Eff}(C^{EBA})$ are displayed. Though the original concept of Ray and Vohra does not have this selection device, we take the illustrative story as

Table 15.2 Equilibrium coalition structures of sequential move coalition formation games

Sequential move coalition formation games	
Equilibrium binding agreements[a]	N = 3: {{1, 2, 3}} N = 4: {{1}, {2, 3, 4}}, {{1, 2}, {3, 4}} N = 5: {{1, 2, 3, 4, 5}}
Sequential coalition formation (Bloch)	N = 3: {{1, 2, 3}}, {{1, 2}, {3}} N = 4: {{1, 2, 3, 4}}, {{1}, {2, 3, 4}} N = 5: {{1, 2, 3, 4, 5}}, {{1}, {2, 3, 4, 5}}
Sequential coalition formation (Ray and Vohra)	N = 3: {{1, 2, 3}} N = 4: {{1}, {2, 3, 4}} N = 5: {{1, 2, 3, 4, 5}}
Sequential coalition formation (Finus)	N = 3: {{1, 2}, {3}} N = 4: {{1, 2, 3, 4}} N = 5: {{1}, {2, 3, 4, 5}}

Note: [a] Only the coarsest coalition structures have been selected.

an important and distinct feature of the formation process and have therefore placed this concept in the context of sequential move coalition models.

From Table 15.2 it appears that rather large coalitions can be supported: either the grand coalition or a coalition comprising $N-1$ countries and a singleton coalition. (See Appendix XII.8 for a derivation of the equilibria.) In particular for large N large coalitions are supported by the unfavorable prospect of further deviations following a deviation. Thus, the set of efficient EBA contains large coalitions if coalition structures of intermediate coalition sizes are not stable.

To summarize, the EBA concept is convincing regarding the behavior of players following a deviation. The assumption that only players belonging to the same coalition can jointly deviate seems rather restrictive as in the CPNE concept. However, in one respect the EBA concept is even more restrictive than the CPNE concept in that only finer but not coarser partitions can be formed by perpetrators. This is an obvious disadvantage.

15.5.2 Sequential Move Unanimity Game

The original idea of the *sequential move unanimity game* (SMUG) goes back to Bloch (1995, 1996). The game is in the spirit of Rubinstein's (1982)

two-player alternating-offers bargaining game and is a generalization of Chatterjee *et al.*'s (1993) extension to an N-country bargaining game. The game proceeds as follows. First, countries are ordered according to some (external) rule, for example, countries are indexed. The country with the lowest index (initiator) starts by proposing a coalition to which it wants to belong. Each prospective member is asked whether it accepts the proposal. According to the external rule the country with the lowest index in the prospective coalition is asked first, then that with the second lowest index and so forth. If all prospective members agree, the coalition, say c_i, is formed and the remaining players $N \backslash c_i$ may form coalitions among themselves. The country with the lowest index among $N \backslash c_i$ becomes the new initiator. If a country rejects a proposal, it can make a new proposal. That is, for a coalition to form, unanimous agreement is required which corresponds to the assumption of exclusive membership in the Γ game.

Bloch's game is set up in extensive form.[26] He assumes an infinite time horizon which makes it impossible to solve the game by backward induction. Therefore, though he assumes no discounting, he supposes that if players cannot agree on a coalition they will receive a payoff of zero which is Pareto-dominated by any other payoff in a coalition (Bloch 1996, p. 97). Through this trick, it is ensured that any sequential equilibrium must be played in finite time (and can therefore be determined by backward induction), though a plausible explanation is missing as to why the status quo payoff (resulting from the singleton coalition structure) does not function as a threat point.

In order to reduce the set of sequential equilibria emerging from the coalition formation process, Bloch considers only stationary perfect equilibrium strategies, also called Markov strategies. That is, a strategy depends only on the 'current state' in the negotiation process. There are basically three states which can occur in the game:

1. There is an ongoing proposal which the player who has the move may accept or reject.
2. A player has rejected a proposal and has him/herself to make a proposal.
3. A coalition has formed and a player becomes the new initiator. Therefore, the payoff relevant part of the history at stage V is the set of players who have left the game already, the partition they have formed and the current offer.[27]

If an initiator makes a proposal s/he will think about two things. First, is the proposal s/he makes acceptable to the proposed members? Obviously, a proposal which is unacceptable makes no sense since the right to make a proposal is passed on to the next player. Second, if those players who are asked accept the proposal, the question arises which coalition the

remaining players will form? The answer to the latter question will of course affect the proposal at the initial stage. Thus, an initiator must solve the entire game backward for all players to find his/her best strategy.

We define now more formally the SMUG and the equilibrium concept. For this we need the following definitions. Some of these definitions have already been encountered above but are listed here for convenience.

Let the set of players be denoted by I, $i \in \mathrm{I}$. A coalition c_i is a non-empty sub-set of players. A coalition c is a partition on the set I and the set of all coalitions is denoted C. For any sub-set of players K of I, the set of partitions on K is denoted by $\mathrm{C_K}$ with typical element c_K. $\pi_i^*(c_j, c)$ is the payoff to player i in coalition j and coalition structure c.

Definition 15.10: History of the game
A history h^V at stage V is a list of all actions taken from stage 0 to $\mathrm{V}-1$. Possible actions are coalition offers, acceptances and rejections up to stage $\mathrm{V}-1$. At any point in the game a history h^V determines:

1. a set $\hat{\mathrm{K}}(h^\mathrm{V})$ of players who have already formed coalitions;
2. a coalition structure $c_{\hat{\mathrm{K}}(h^\mathrm{V})}$ formed by the players in $\hat{\mathrm{K}}(h^\mathrm{V})$;
3. an ongoing proposal (if any) $\hat{c}_j(h^\mathrm{V})$;
4. a set of players $\tilde{c}_j(h^\mathrm{V})$ who have already accepted the proposal (including the initiator); and
5. a player i who moves at stage V.

Player i is called active at stage V if it is his/her turn to move after history h^V. The set of histories at which player i is active is denoted by H_i.

Definition 15.11: Strategy of a player
A continuation strategy σ_i of player i is a mapping from H_i to his/her set of actions, namely (a) $\sigma_i(h^\mathrm{V}) \in \{\text{yes, no}\}$ if $\hat{c}_i(h^\mathrm{V}) \neq 0$; and (b) $\sigma_i(h^\mathrm{V}) \in \{c_i \subseteq \mathrm{I} \backslash \hat{\mathrm{K}}(h^\mathrm{V}),\ i \in c_i\}$ if $\hat{c}_i(h^\mathrm{V}) = 0$.

As pointed out above, Bloch considers only Markov strategies, so the history at stage V is described by $h^\mathrm{V}(\hat{\mathrm{K}}, c_{\hat{\mathrm{K}}}, \hat{c}_i)$, that is, the set of players who have left the game already, the partition they have formed and the current offer. Then a stationary subgame-perfect equilibrium in the SMUG can be defined by using Definition 15.11:[28]

Definition 15.12: Subgame-perfect equilibrium in the sequential move unanimity game
A subgame-perfect equilibrium in the sequential move unanimity game is a continuation strategy combination $\sigma_i^*(h^\mathrm{V})$ for which $\pi_i^*(\sigma_i^*(h^\mathrm{V}), \sigma_{-i}^*(h^\mathrm{V})) \geq \pi_i^*(\sigma_i'(h^\mathrm{V}), \sigma_{-i}^*(h^\mathrm{V}))\ \forall i \in \mathrm{I}$ and $h^\mathrm{V} \in \mathrm{H}_i$.

Bloch (1996) has shown that for symmetric valuations a simple finite procedure can be used to determine the equilibrium/equilibria in the game. The algorithm works as follows. The first player proposes an integer $|c_1| \in [1, N]$ which indicates the coalition size. The player $|c_1|+1$ proposes a coalition of size $|c_2| \in [1, N-|c_1|]$. This process continues until $|c_1|+|c_2|+...+|c_M|=N$.

From Table 15.2 it appears that the SMUG according to Bloch does not produce a unique equilibrium in the emission game. The reason is that in this symmetric game players face an indifference between two strategies (see Appendix XII.9) at some stages. Therefore, two equilibria can be supported.

It is also apparent from the example that rather large coalitions can be supported. However, for a larger number of countries this may be different. In the extension considered below a more exact statement is possible.

Ray and Vohra (1999) have generalized the sequential formation game of Bloch. In particular, in contrast to the previous concepts which assume a fixed sharing rule, they have endogenized the sharing rule. Since their extension is quite involved, we shall not pursue this issue further and consider only a slight modification of Ray and Vohra. They assume that in the case of indifference between two strategies a country selects the largest coalition size. In the emission game this assumption seems plausible since all other players (except the player who has the move) would prefer the larger coalition to form. (For details, see Appendix XII.9.) This selection device produces a unique equilibrium coalition structure which can be characterized according to Bloch (1997, pp. 338ff.) by a *Fibonnaci decomposition*. A Fibonnaci decomposition is derived from a sequence of Fibonnaci numbers where $f_0=1, f_1=2$ and $f_n=f_{n-1}+f_{n-2}$. One starts by choosing the largest Fibonnaci number equal to or smaller than N. Denote this Fibonnaci number f_n^k. Then one looks for the largest Fibonnaci number equal to or smaller than $N-f_n^k$. This process continues until $\Sigma f_i^k = N$.

For instance, consider the Fibonnaci numbers 1, 2, 3, 5, 8, 13, 21 and so on. Thus if $N=4$, the largest Fibonnaci number is 3 and $N-3=1$. Hence, with some abuse of notation, the coalition structure is {3, 1} where the numbers indicate the coalition size. If $N=20$ the coalition structure would be {13, 5, 3} and for $N=8$ the grand coalition forms. Thus, whenever N is not a Fibonnaci number rather asymmetric coalition structures form.

In contrast to Ray and Vohra, however, it may well be argued that a player, though s/he may not bother about exhibiting a positive externality on other players by choosing the larger coalition size if s/he is indifferent between two strategies, s/he may be concerned about another issue. Suppose there is a slight probability that a player, say k, who follows a player say, i, in the sequence of moves, unintentionally makes a small error by choosing a wrong strategy. In the terminology of Selten (1975) a player

may make a small tremble.[29] Given this possibility, it might be in the interest of player i to choose a strategy which is immune to such an error.[30]

For instance, suppose $N=3$ and country 1 proposes the grand coalition. Alternatively, it could propose a singleton coalition, knowing that it is in the interest of country 2 to propose a coalition comprising countries 2 and 3 which country 3 will accept, that is, $\pi^*_{2/3}(\{\{1\}, \{2, 3\}\}) > \pi^*_{2/3}(\{\{1\}, \{2\}, \{3\}\})$. According to Bloch, countries 2 and 3 will accept the proposal of a grand coalition since $\pi^*_2(\{\{1, 2, 3\}\}) = \pi^*_2(\{\{1,3\}, \{2\}\})$ and $\pi^*_3(\{\{1, 2, 3\}\}) = \pi^*_3(\{\{1,2\}, \{3\}\})$. However, suppose that, say, country 2 rejects the proposal by mistake and proposes itself as a singleton coalition, after which countries 1 and 3 will form a coalition. Since $\pi^*_1(\{\{1, 2, 3\}\}) = \pi^*_1(\{\{1\}, \{2,3\}\}) > \pi^*_1(\{\{1, 3\}, \{2\}\})$ country 1 may be on the safe side by proposing itself as a singleton coalition instead of the grand coalition, and the resulting equilibrium coalition structure is $\{\{1\}, \{2, 3\}\}$.

As is evident from Table 15.2, this procedure also delivers a unique equilibrium in the example. In fact, denoting Bloch's equilibrium set by $C^{SE(B)}$, that of Ray and Vohra $C^{SE(V/R)}$ and the trembling hand equilibrium by $C^{SE(F)}$, it is easy to see that $C^{SE(B)} = C^{SE(V/R)} \cup C^{SE(F)}$.

An obvious advantage of the SMUG is that the formation process is explicitly modeled. In contrast to the simultaneous move games, there is an explicit story of how a coalition builds and how players coordinate on an equilibrium. It seems plausible for many games that there is an initiator at the beginning of the game who proposes a coalition which is in his/her interest. This is particularly true in asymmetric games. For instance, in an asymmetric emission game it should be expected that governments with a higher environmental preference will approach other governments with a similar high preference. Governments with a lower preference may form coalitions among themselves. Membership is only offered to outsiders if all coalition members unanimously agree. Thus, Bloch's framework is the most sophisticated among the sequential move coalition formation models. It seems promising for future research to derive the coalition structures in an asymmetric emission game by applying his concept.[31]

A first step in this direction has been undertaken by Finus and Rundshagen (1999). The authors investigate an issue linkage game which is a mix of a positive and negative externality game. The positive externality part of the game concerns the reduction of global emissions emanating from the production of a homogeneous good which is sold in the domestic market and exported abroad. The negative externality part of the game concerns the formation of a customs union where the members abolish tariffs within the union but protect their market against outsiders. Since the authors assume that firms may migrate to other countries if the environmental policy of governments is too strict, the endogenous plant location

game is quite complex. Therefore, the authors restrict the number of players to three countries only but they consider heterogeneous countries. Although Finus and Rundshagen derive some quite interesting results worth reporting, due to lack of space we restrict the following discussion to some conceptual extensions of their model.

The first extension is the construction of a finite game. This is simply done by assuming that there is a record of all past proposals which is known to all players. Since no player is allowed to make the same proposal twice, the game ends in finite time and can be solved by backward induction. Thus, neither the assumption of symmetric countries nor the assumption that players receive a payoff of zero if they do not form coalitions is needed to solve the game.

Second, the authors consider non-stationary strategies. Thus the history of the game is not (arbitrarily) restricted to the immediate state in the game: an assumption which seems plausible for rational players.

The third extension concerns the order in which players move. Instead of assuming a fixed rule according to which players move, they introduce two new features. The first feature is that if there is an ongoing proposal the player who has the move decides who will be the next player to be asked whether to accept or reject the proposal. The second feature is that at the beginning of the game and after the last player belonging to a proposed coalition has accepted a proposal, nature, which is modeled as an additional player, selects randomly the player who will move next. This has the advantage that in asymmetric environments the exogenous rule according to which players move does not affect the equilibrium outcome. Consequently, results are more general. Of course, such extensions imply a more complex definition of a strategy than in Definition 15.11. The interested reader is referred to Finus and Rundshagen (1999) for a full account of the extended definition.

NOTES

1. Bloch (1997) and Yi (1997) provide an excellent survey of most concepts discussed in this chapter. They have applied these concepts to the provision of a public good. Their example could be interpreted as a global emission game defined in reduction space.
2. The equilibrium valuation is, apart from the normal and extensive form, a third possibility for representing a non-cooperative game. It may also be called the coalitional representation of games. It corresponds to the characteristic function in cooperative games (see Section 13.3).
3. So far not much progress has been made in endogenizing the sharing rule in the coalition formation process. An exception is Ray and Vohra (1999). They provide some vindication for the equal sharing rule if countries are symmetric.
4. This term refers to the fact that, although all players have the same strategies and payoff

spaces at the beginning of the play, asymmetric coalition structures may *ex post* emerge during the formation process.

5. Obviously, for symmetric countries any kind of transfers are therefore not relevant to the coalition formation process. This is at least true as long as one assumes that transfers can only be paid among coalition members but not to outsiders. We shall make this assumption in what follows.
6. It is interesting to note that in neither of the papers quoted below is this inconsistency recognized.
7. Note that this definition does not imply that the coalition structure itself is an equilibrium. It only implies that a coalition maximizes the payoffs of its members for a given coalition structure.
8. Coalition games without spillovers are discussed in Konishi *et al.* (1997).
9. To avoid confusion: so far we have called transboundary emissions a negative externality. Now, in this chapter it is more convenient to keep with the tradition of Yi (1997). Hence, we call an emission game a positive externality game since we think in terms of emission reductions. A deviation from Nash equilibrium emissions by a coalition implies an emission reduction.
10. Conditions C_1–C_3 apply in general to positive externality games. Condition C_4 may not hold in general but is true for payoff function (13.1). A general proof may be found in Yi (1997).
11. We state the proposition with respect to payoff function (13.1) since the subsequent discussion is illustrated by this example. This procedure saves space, though a more general proof for a larger class of payoff functions is certainly possible but more involved. See Yi (1997) for a proof in a game on the provision of a public good.
12. Thus the CPNE is identical to a renegotiation-proof equilibrium in a two-player game. See Chapter 6.
13. Unfortunately, there are no conditions which guarantee the existence of an SNE or a CPNE.
14. For an extension to cover correlated strategies also, see Moreno and Wooders (1993).
15. One reason may be that some consumers prefer to buy an environmentally friendly car and some consumers prefer a cheaper car without a cat. Thus the market is segmented and makes heterogeneous production attractive.
16. Of course, country 3 would like to deviate to a_3 to gain a payoff of 6 instead of 0. However, since player 3 is not among the original deviators, this deviation is not considered by the CPNE concept. See the discussion on p. 290.
17. Of course, strictly speaking a dynamic definition would be required. However, for convenience we follow the somewhat sloppy notation and definition of Bloch (1997) and Yi (1997). For an extended definition in a supergame framework, see Stähler (1996). He applies his definition to a global emission game comprising three countries.
18. Hence, one can immediately conclude that for $N = 9$ $\{\{1, 2, 3, 4\}, \{5, 6, 7\}, \{8, 9\}\}$ cannot be an NE coalition structure under the open-membership rule.
19. An application of the open-membership game to the formation of custom unions in international trade (negative externality game) may be found in Yi (1996).
20. An application to the formation of custom unions in international trade (negative externality game) may be found in Burbidge *et al*. (1997).
21. For a more formal definition, see Chwe (1994, pp. 302ff.).
22. Though the FCS concept considers sequential deviations, the coordination of an equilibrium coalition structure occurs by assumption simultaneously. This is why the concept is classified as a simultaneous move game.
23. A concept similar to the FS concept has been applied by Ecchia and Mariotti (1997) in the context of a global emission game.
24. For a full account of the concept the reader is referred to Ray and Vohra (1997).
25. Alternatively, one could assume that the formation process starts from the singleton coalition and that coalitions gradually become larger until no further enlargement is stable.
26. See Chapter 4 for the definition of the extensive-form representation of a game.

27. Note that, since payoffs are received at the end of the game, the game comprises (according to the terminology of this book) only one period but of several stages V. See Chapter 4.
28. Recall that a continuation strategy has been defined in Chapter 4 and applied in Definition 4.9 to characterize a subgame-perfect equilibrium in a repeated game.
29. This does not imply that this player is irrational but only that s/he may choose a wrong strategy with probability ϵ where $\lim \epsilon = 0$. For details of the motivation of this concept, see Selten (1975).
30. An issue for future research is to formalize the idea of the trembling hand sequential equilibrium in the SMUG.
31. Bloch (1995) applied his concept to the coalition formation in Cournot and Bertrand oligopoly where firms may form associations in order to reduce production costs or costs of R&D (negative externality game). Yi (1996) applies the sequential formation game to the formation of custom unions (negative externality game).

16. Summary and conclusions

After important terms, the notation of this book and the structure of game theory had been laid out in Chapter 2, we started out in Chapter 3 by analyzing four simple matrix games: the prisoners' dilemma, the chicken, the assurance and the no-conflict games. The analysis first focused on two countries only (Sections 3.2–3.5) and was then extended to N countries. Depending on the cost–benefit structure of an abatement policy, either *no* cooperation, cooperation among *some* countries or cooperation among *all* countries could be explained. Though these matrix games are rather simple by their nature, basic features of the coalition models in Chapters 13–15 could already be depicted.

In Chapter 3 it was also shown that though there is no external coordinator endowed with the power to enforce an IEA, and hence countries must play a correlated Nash equilibrium, coordination may improve upon the non-cooperative outcome. This result was demonstrated with the help of the chicken game. It was pointed out that future research should derive conditions under which coordination is possible and that it should scrutinize whether it is possible to transform a game such that coordination can be applied more effectively. Moreover, retrospectively of the coalition formation models, it seems promising to look at the effect of coordination in an N-country world. Due to the complexity of this issue, it seems promising to look first at some of the reduced coalition formation games of Chapter 15. Due to their simple structure, the simultaneous move games seem particularly suited for such an undertaking. For instance, in the open-membership game and the exclusive membership game, one could determine the set of correlated coalition-proof Nash equilibria and find out whether large coalitions could be established as a self-enforcing outcome. Of course, since we established that the farsighted coalitional stability of Chwe is a particularly convincing concept, it would also be interesting to investigate whether this concept could be extended to correlated strategies too.

In Chapters 4–7 the framework was extended to cover dynamic games. For finitely repeated games Chapters 4 and 6, and for infinitely repeated games Chapters 5 and 7 traced the historical development of equilibrium concept refinements. It became evident that the steps in this development were closely related to more sophisticated definitions of credible threat

strategies. Moreover, it appeared that those refinements allow us to make more precise predictions about the outcome of a game. The various equilibrium concepts were applied to simple matrix games with discrete strategy space. In finitely repeated games the concept of a renegotiation-proof equilibrium and in infinitely repeated games the concepts of weakly and strongly renegotiation-proof equilibria were identified as being a particularly convincing tool for analyzing the stability of IEAs. Chapters 3–7 elaborated some basic results of practical importance. For instance, it could be shown that the following conditions are conducive to the chances of cooperation: a long-term relationship between governments; regular meetings between governments to monitor compliance of a treaty; immediate reactions to the violation of a treaty; the usefulness of harsh but also credible threat strategies; and low discounting of the gains accruing from cooperation between countries. In contrast, if environmental projects involve a high amount of sunk costs or if the structure of international environmental problems is such that a change in environmental policies takes a considerable time, this is unfavorable for cooperation. For practical purposes these results suggest that governments should pursue an environmental policy of 'small steps' and that regular meetings should be institutionalized to build up mutual confidence among signatories. Moreover, treaties should establish credible punishments which are transparent to all participants and which are so simple that possible violations of treaties can immediately be punished. It remains for future research to specify exactly what this means for the design of IEAs.

In Chapter 8 it was shown that the linkage of two or more issues may be conducive to cooperation. First, it was demonstrated that in an infinitely repeated framework issue linkage may help to avoid asymmetric payoffs so that it is easier to stabilize an IEA. Second, in a finitely repeated game framework it was shown that if one game possesses the necessary properties to be stabilized as a subgame-perfect or renegotiation-proof equilibrium, a second game which does not possess these properties can be stabilized by linking both games to each other. Third, departing from the classical framework in which issue linkage games are commonly analyzed, it was shown that issue linkage may not always be conducive to cooperation. For a prisoners' dilemma game it was established that if issues are complements in governments' objective functions, then issues are better negotiated separately. As one possibility to achieve this objective delegating decision-making power to independent agencies with concave utility functions was suggested.

From the analysis three topics suggest themselves for treatment in future research. First, for various environmental problems and possible issues to which they could be linked it should be investigated empirically whether

they constitute substitutes or complements in governments' objective function. Based on these results, recommendations on the design of future IEAs and the redesign of current ones could be given. Second, if it turns out that there is no suitable substitutional relationship between issues, the design of national or international agencies should be given high priority in future research. Third, though this has not been analyzed explicitly, issue linkage may involve considerable transaction costs. Therefore, empirical estimates of these costs would be useful for an empirically founded forecast of whether the possible gains from issue linkage will actually be realized in certain situations. In the light of the results of Chapter 15 one could also analyze whether the coexistence of several IEAs may be associated with lower transaction costs and therefore issue linkage may be more effective among small groups of countries. Particularly in cases where a large number of countries suffer from an externality, it may be easier to find suitable issues to be linked among small coalitions than among large ones.

In Chapter 9 a simple global emission game was introduced and fundamental benchmarks were derived which formed the basis for the subsequent analysis. Particular emphasis was given to the conditions which guarantee the existence of a unique Nash equilibrium and the location and curvature of reaction functions.

In Chapter 10 three models of the literature on the provision of public goods were discussed which helped to explain those IEAs in which signatories depart substantially from non-cooperative abatement targets. It became evident that the only charm of models based on 'non-Nash behavior' and the 'theory of reciprocity' lies in their simplicity. A thorough game theoretical investigation revealed that those models are based on inconsistent behavior by agents. This was also found to be true for the 'strategic matching' approach, though it appeared that this approach is far from being simple. Chapter 10 also derived the bargaining equilibrium of an auction of emission reductions, the stability of which was investigated in a supergame framework in Chapter 12.

Similar to the auctioning equilibrium, two more bargaining equilibria were analyzed in Chapter 11. One equilibrium was derived by assuming that countries negotiate on the level of a uniform emission reduction quota, and the other equilibrium was derived by assuming that countries negotiate on the level of a uniform effluent charge. Abstracting from stability considerations, a comparison revealed that under a quota agreement global emissions may be lower and global welfare may be higher than under a tax agreement. Particularly in those cases where an agreement puts a particular strain on one bargaining partner, a quota agreement provides more favorable conditions to the 'bottleneck country' in the negotiations than a

tax agreement, so higher abatement targets can be achieved in a treaty. This cost–benefit effect may compensate for the inherent inefficiency of the quota as a typical command and control instrument. Additionally, it was shown that the tax agreement may be subject to strategic proposals which lead to an inefficient bargaining outcome whereas the quota agreement is immune to biased proposals.

In Chapter 12 the stability properties of the auctioning equilibrium and also of the tax and quota bargaining equilibria were tested in a supergame framework. One important result was that if punishment options are restricted for some reason the tax agreement may not be stable. This was shown to be always true in those cases where biased proposals occur under a tax agreement. In contrast, the stability of the quota agreement is rather robust.

However, the main purpose of Chapter 12 was to characterize the weakly and strongly renegotiation-proof as well as the strongly perfect equilibrium payoff space in the global emission game in a two-player environment. Particular emphasis was given to restricted punishment profiles. This analysis was motivated by several restrictions which may occur in reality. First, technical conditions may not allow a country either to reduce emissions rapidly from some agreed level during the repentance phase of a punishment or to increase emissions quickly above a certain level in order to punish non-compliance. Second, international law requires punishments to be proportional to the severity of the violation. It was shown that those restrictions make it more difficult to stabilize an agreement. In particular, agreements involving asymmetric payoffs and agreements which depart substantially from the non-cooperative status quo have scarcely a chance of being realized. Moreover, it turned out that in a global emission game it is a very complex undertaking to operationalize 'relative proportional punishments', also called 'reciprocal punishments'. It requires non-simple punishment profiles – different from those of Abreu.

For future research it would be promising to consider such reciprocal punishments in case studies of actual IEAs as conducted by Finus and Tjøtta (1998) and Murdoch and Sandler (1997) in order to capture an important restriction under which IEAs are operating in reality.

In Chapters 13–15 the analysis of a global emission game was extended to an N-country framework. Whereas Chapters 13 and 14 presented models which basically constitute the state of the art in the literature on coalition formation, Chapter 15 reported on new developments in the field of non-cooperative game theory. A distinguishing feature of the models in Chapters 13 and 14 was that they assume only one group of signatories, whereas all other countries behave as singletons. In contrast, the concepts of Chapter 15 allow for the coexistence of several coalitions.

The models presented in Chapter 13 were classified as static representations of dynamic games. Among the conjectural variation models, Barrett's model was identified as possessing the highest explanatory power. It can explain IEAs ranging from two to N countries and it can relate the outcome to the cost–benefit structure from emissions. The two central results are seemingly paradoxical: (a) IEAs are only signed by many countries if they achieve little; (b) whenever the cost–benefit structure of an environmental problem would allow the netting of high gains from cooperation, IEAs achieve only little. These apparently paradoxical results were confirmed by the supergame coalition model in Chapter 14.

Apart from other weaknesses, the inconsistency of the stability concept was identified as a major shortcoming of conjectural variation models. A modification in the spirit of the farsighted coalitional concept of Chwe was suggested to remedy this deficiency.

Shortcomings were also identified for the core models of Section 13.3. Apart from the inconsistent behavior of external players, it appeared that the explanatory power of these models is rather low; they predict that the grand coalition will form as long as an appropriate transfer scheme among countries is established.

In Chapter 14 a coalition formation process was analyzed in a supergame framework. It was shown that if the concept of weakly renegotiation-proofness is applied, it is very unlikely that a grand coalition will form. This was shown to apply for a socially optimal agreement but also for agreements which require less demanding abatement targets to be realized within the coalition of all countries. Therefore, the formation of a smaller coalition among the most environmentally concerned countries was analyzed. It turned out that for global environmental problems, where the number of countries affected by an externality is large, signatories will agree on an emission quota but not on an effluent charge. The result served as an additional explanation of the frequent occurrence of emission quotas in many IEAs. Moreover, it turned out that it might be rational for only a small coalition of rather homogeneous countries to sign an IEA (instead of the grand coalition); this is particularly true in those cases in which cooperation is difficult to achieve (critical parameter constellations).

In future research it would certainly be interesting to confirm those results if the supergame model were modified such that it allowed for the coexistence of several coalitions. This seems suggestive since almost all the concepts of Chapter 15 predict the coexistence of several coalitions as a stable coalition structure. For this purpose the definition of a coalition-proof equilibrium has to be extended to infinitely repeated games. First attempts in this direction have been undertaken by Stähler (1996).[1]

From the 'new' game theoretical concepts on the coalition formation

process presented in Chapter 15 two concepts appeared to be particularly fruitful for an application in future research. Among the simultaneous move concepts Chwe's farsighted coalitional concept was identified as a consistent static representation of a dynamic coalition formation process. It takes into account all forms of reactions and counter-reactions of coalitions and countries. Among the sequential move concepts it turned out that Bloch's sequential move unanimity game, and its extensions by Ray and Vohra and by Finus and Rundshagen, models a coalition formation process in a plausible and consistent manner. It was argued that the sequential move concepts have the advantage that they explicitly model the coordination on an equilibrium coalition structure via the activities of an initiator who seeks equally minded coalition partners.

Almost all the concepts of Chapter 15 suggest that the stability of IEAs could be increased if IEAs were tailored to individual groups of countries which form a coalition among themselves instead of trying 'to get all countries into one boat'. This result should play an important role in the design of future IEAs.

Despite the fact that the two 'new' coalition concepts mentioned above already exhibit a high degree of sophistication, a long list of issues which may be treated in future research come to mind. First, both concepts await to be applied to a global emission game with heterogeneous countries and strictly concave benefit and cost functions from emissions (or from abatement). Second, and particularly important in a world of heterogeneous countries, the choice of the sharing rule for the gains from cooperation among the coalition members has to be endogenized in these models. For Bloch's model, Ray and Vohra (1999) have made a first attempt in this direction; in the context of Chwe's model such an attempt is still missing. Third, the choice of equilibrium strategies among coalition members must be justified more convincingly. So far a consistent and entirely endogenous derivation from the models is missing. Fourth, the coalition models should be extended to cover cases of incomplete information. Of course, in reality many forms of incomplete information may be associated with the problem of coalition formation in international pollution control; however, a particularly fruitful extension might be an integration of incomplete monitoring. Though scientists may identify the total amount of pollutants released to the atmosphere, it may be difficult for them to assign these emissions to single countries. There may remain some considerable uncertainty about the exact amount of emissions each country releases. This is particularly true since in most IEAs monitoring relies on self-reporting by countries, and this is normally rather patchy.

A possible extension could borrow from work on stable coalitions in oligopoly (for example, Green and Porter 1984; Abreu *et al.* 1986; Vives 1984).

In this literature a Cournot oligopoly is assumed where the market price can be publicly observed; not so, however, the output by each firm. The price is a function of the output and of some erratic fluctuations due to (not observed) changes in demand. Thus, though the price is a proxy of the compliance record of firms which are obliged to reduce output to cooperative levels, it is not an indicator which allows for unequivocal conclusions. The similarity of this example to the problem of international pollution is obvious. In the international environmental pollution context, the price is the observable aggregate emissions and the output by each firm is the non-observable emissions of each country. For this suggested extension one may also want to consult the work of Avenhaus (1992); Güth and Pethig (1992); Russell (1990, 1992); and Russell *et al.* (1986) on the monitoring of emissions. Moreover, equilibrium concepts in the spirit of a renegotiation-proof or coalition-proof equilibrium must be extended to games of incomplete information. A good source for a possible definition of equilibrium strategies in such an incomplete information framework is undoubtedly Fudenberg and Levine (1992) and Fudenberg *et al.* (1994) and their public equilibrium concept.

Finally, we should like to finish with a last general remark. As pointed out in the Introduction, there are two strands in the literature. One strand is empirically oriented, estimating the costs and benefits of the control of particular pollutants. The second strand game theoretically investigates the stability of IEAs and the incentive structure of countries. So far, both strands have coexisted more or less in isolation. Therefore, it seems promising for future research to combine both approaches in the spirit of Finus and Tjøtta (1998) and Murdoch and Sandler (1997). Though this book has tried hard to make the point that game theory is not only a toy for academics but also a useful device for analyzing problems of cooperation in international pollution control, critics of game theory could be even better convinced if the theoretical concepts were more often applied to concrete environmental problems. Such applications would also constitute a major step toward sound policy recommendations with respect to the design of more efficient and effective IEAs in the future: a concern which most readers will share.

NOTE

1. Bernheim *et al.* (1987) define their concept of a coalition-proof equilibrium only for static and finite dynamic games. Farrell and Maskin's (1989a) concept of a weakly and strongly renegotiation-proof equilibrium is defined for two countries only or for N countries but where only single deviations are considered. Stähler (1996) combines both concepts and adapts them to a supergame framework. However, his analysis is restricted to three countries only.

Appendices

I CHAPTER 3: APPENDIX

Proposition 3.1 claims that in a chicken game there is a correlated strategy equilibrium with aggregate payoffs higher than in any of the pure or uncorrelated strategy equilibria.

Proof: For the proof of Proposition 3.1 it is helpful to note the following relations. If two events, say A and B, occur independently with probability $p(\mathrm{A})$ and $p(\mathrm{B})$, then the probability that they occur jointly is given by $p(\mathrm{A} \cap \mathrm{B}) = p(\mathrm{A}) \cdot p(\mathrm{B})$. If events A and B do *not* occur independently, then the following relation applies:

$$p(\mathrm{A/B}) = \frac{p(\mathrm{A} \cap \mathrm{B})}{p(\mathrm{B})} \tag{I.1}$$

where $p(\mathrm{A/B})$ reads as 'probability that event A occurs given event B occurred' (see, for example, Rasmusen 1989, pp. 54ff). For instance, in Matrix 3.9 the probability that strategy a_{11} is played is $z_1 + z_2$. Therefore, the probability that strategy combination a_{21} is played, given player 1 plays a_{11}, is therefore $p(a_{21}/a_{11}) = z_1/(z_1 + z_2)$. Moreover, we find $p(a_{22}/a_{11}) = z_2/(z_1 + z_2)$ and so on.

Now, let payoffs be given by the General Payoff Matrix 3.2 and the correlated distribution z by Matrix 3.9. Assume for notational convenience a symmetric game so that $a_1 = a_2 = a$ and so on.[1] Then, if the coordinator recommends country 1 to play a_1, its expected payoff is given by the two LHS terms in inequality (I.2). Alternatively, if country 1 deviates, it receives an expected payoff represented by the two terms on the RHS of this inequality. For stability, it is thus required that the expected payoff when complying is at least as high as when defecting:

$$a \cdot \frac{z_1}{z_1 + z_2} + b \cdot \frac{z_2}{z_1 + z_2} \geq c \cdot \frac{z_1}{z_1 + z_2} + d \cdot \frac{z_2}{z_1 + z_2} \Leftrightarrow z_1 \cdot \frac{c - a}{b - d} \leq z_2. \tag{I.2}$$

The second possibility to be considered is that the coordinator recommends playing the second strategy to country 1. For this to be an equilibrium recommendation the following inequality must hold:

$$c\cdot\frac{z_3}{z_3+z_4}+d\cdot\frac{z_4}{z_3+z_4}\geq a\cdot\frac{z_3}{z_3+z_4}+c\cdot\frac{z_4}{z_3+z_4}\Leftrightarrow z_4\leq z_3\cdot\frac{c-a}{b-d}. \tag{I.3}$$

By the same token, it is straightforward to find the following equilibrium conditions for country 2:

$$z_1\cdot\frac{c-a}{b-d}\leq z_3 \tag{I.4}$$

$$z_4\leq z_2\cdot\frac{c-a}{b-d} \tag{I.5}$$

where we may recall that $c>a>b>d$ and $2a>b+c>2d$ holds in a chicken game. Moreover:

$$z_1+z_2+z_3+z_4=1 \tag{I.6}$$

must hold by definition. Since we look for an equilibrium with a high aggregate payoff, $z_4=0$ is assumed. Thus, equations (I.3) and (I.5) become nonbinding and we are left with two inequalities. Moreover, since strategy combination (a_1, a_2) delivers the highest aggregate payoff in this game, z_1 should be chosen as large as possible subject, however, to the two constraints (I.2) and (I.4). Hence z_2 is chosen equal to the LHS term in (I.2) and z_3 equal to the LHS term in (I.4). Substituting this information into (I.6) and solving for z_1, z_2 and z_3 respectively, gives:

$$z_1=\frac{b-d}{b-d+2c-2a},\ z_2=\frac{c-a}{b-d+2c-2a},\ z_3=\frac{c-a}{b-d+2c-2a} \tag{I.7}$$

from which the aggregate equilibrium payoff $(\Sigma\pi_i(z^{N^*})=2a\cdot z_1+(b+c)\cdot z_2+(b+c)\cdot z_3)$

$$\Sigma\pi_i(z^*)=\frac{2(bc+c^2-ca-da)}{b-d+2c-2a} \tag{I.8}$$

follows. In any of the two NE in pure strategies, aggregate payoffs are given by $\Sigma\pi_i(s^*)=b+c$ and in the mixed strategy equilibria aggregate payoffs $(\Sigma\pi_i(p^{N^*})=2\cdot a\cdot p^2+2\cdot(b+c)(1-p)\cdot p+2\cdot d\cdot(1-p)^2)$ are computed to be:

$$\Sigma\pi_i(p^{N^*})=\frac{2(bc-da)}{b-d+c-a} \tag{I.9}$$

which follows from $p_1^*=p_2^*=p=(b-d)/(b+c-a-d)$. Then:

$$\Sigma\pi_i(z^{N^*})-\Sigma\pi_i(s^{N^*})=\frac{2(2a-b-c)(b-d)}{2(c-a)+b-d}>0 \tag{I.10}$$

$$\Sigma\pi_i(z^{N^*}) - \Sigma\pi_i(p^{N^*}) = \frac{2(c-a)^2(c-d)}{(2c-2a+b-d)(c-a+b-d)} > 0 \quad \text{(I.11)}$$

where we make use of the relations $c > a > b > d$ and $2a > b + c > 2d$ in a chicken game. QED

For the example in Matrix 3.3 where $a = 4.6$, $b = 2.2$, $c = 5.2$, $d = 2$ we therefore have $z_1^* = 1/7$ and $z_2^* = z_3^* = 3/7$ and $\Sigma\pi_i(z^{N^*}) = 7.656$.

II CHAPTER 4: APPENDIX

In Theorem 4.4 we claim that in a simultaneous (sequential) move game any payoff tuple which gives each player more than his/her worst stage game NE (SPE) can be sustained as an average payoff vector in a finitely repeated game for large T and discount factors close to 1.

Proof: We start by considering simultaneous moves and assume, as Friedman (1985), that there is a 'good' equilibrium $s^{N(1)}$ and a 'bad' equilibrium $s^{N(2)}$, that is, $\pi_i^{N(1)}(s^{N(1)}) > \pi_i^{N(2)}(s^{N(2)})\ \forall\ i \in \text{I}$. This assumption simplifies notation. We extend the proof subsequently to cover the other cases of Theorem 4.4 as well.

Consider the following trigger strategy:

$$\sigma_i = \begin{cases} s_i^j & \text{if in any } h^t(0) \ldots h^t(t-1) = (s_j', s_{-j}),\ h^t(0) \ldots h^t(t-1) = (s_j'', s_{-j}^k), \\ & h^t(0) \ldots h^t(t-1) = (s_j^j, s_{-j}^j)\ \forall t = 0 \ldots \text{T, or} \\ & h^t(t^*+1) \ldots h^t(t-1) = (s_j''', s_{-j}^{N(1)})\ \forall t = t^*+1 \ldots \text{T} \\ s_i & \text{in } t = 0 \\ s_i & \text{if in any } h^t(0) \ldots h^t(t-1) = (s_i, s_{-i})\ \forall t = 0 \ldots t^* \\ s_i^{N(1)} & \text{if in any } h^t(0) \ldots h^t(t-1) = (s_i, s_{-i})\ \forall t = 0 \ldots t^* \\ & \wedge\ h^t(t^*+1) \ldots h^t(t-1) = (s_i^{N(1)}), s_{-i}^{N(1)})\ \forall t = t^*+1 \ldots \text{T} \end{cases} \quad \text{(II.1)}$$

where s_j', s_j'' and s_j''' are arbitrary strategies played by country j, $s_j' \neq s_j$, $s_j'' \neq s_j^k$, $s_j''' \neq s_j^{N(1)}$ and $j \neq i \neq k$. s_i^j is a strategy of player i to punish player j. As pointed out above, we start by assuming $s_i^j = s_i^{N(2)}$.

Strategy σ_i may be summarized as follows: along the equilibrium path a stage game strategy combination $s = (s_i, s_{-i})$ is played in every round for $t^* < \text{T}$ times and payoffs $\pi_i^*(s) \geq \pi_i^{N(2)}(s^{N(2)})$ are received. From $t = t^* + 1$ until the end of the game, the good Nash strategy combination is played. During the playing of s, stage game payoffs of $\pi_i^*(s) \geq \pi_i^{N(2)}(s^{N(2)})$ are received. If a

player deviates at any time $t = t^0$ prior to t^*, that is $t^0 \leq t^*$, the threat involves playing the bad NE strategy combination until the end of the game. Thus, deviation triggers a punishment until the end of the game.

This punishment strategy is played against player j on four occasions:

1. All players play the agreed equilibrium strategy up to time $t \leq t^*$ except player j.
2. Player k deviates in a previous round and all players punish player k except player j.
3. Punishment has started already in a previous round and therefore will be continued.
4. All players play the good equilibrium strategy except player j.

Of course, cases 2 and 4 include an irrational move by player j (because they imply a deviation from a stage game NE), but are listed for completeness to emphasize that an equilibrium strategy should specify a best reply for all possible, though unlikely, events. But also in the cases 1 and 3 it obviously does not pay *not* to conduct the punishment.[2] Hence, it follows immediately that strategy σ_i is an equilibrium strategy during the punishment phase and by Theroem 4.1 it is subgame-perfect. It remains to be shown that neither does it pay to deviate in the cooperative phase. Basically, a player can deviate during the first t^* stages or the last $T - t^*$ stages. However, since in the last $T - t^*$ stages a Nash stage equilibrium is played along the equilibrium path, deviation does not pay by definition. Consequently, we only have to pay attention to stability in the first t^* stages.

A player i does not deviate provided the payoff stream when complying (for t^* periods a player receives π_i^* and for $T - t^*$ periods $\pi_i^{N(1)}$) is higher than when taking a free-ride in any period t^0 and then being punished afterwards (for $t^0 - 1$ periods a player receives π_i^*, in period t^0 π_i^D, and afterwards $\pi_i^{N(2)}$).[3]

$$\sum_{t=0}^{t^*} \delta_i^t \pi_i^* + \sum_{t=t^*+1}^{T} \delta_i^t \pi_i^{N(1)} \geq \sum_{t=0}^{t^0-1} \delta_i^t \pi_i^* + \delta_i^{t^0} \pi_i^D + \sum_{t=t^0+1}^{T} \delta_i^t \pi_i^{N(2)}. \qquad \text{(II.2)}$$

Rearranging gives:

$$\delta_i^{t^0} \pi_i^* + \sum_{t^0+1}^{t^*} \delta_i^t \pi_i^* \geq \delta_i^{t^0} \pi_i^D + \sum_{t^0+1}^{t^*} \delta_i^t \pi_i^{N(2)} - \sum_{t^*+1}^{T} \delta_i^t (\pi_i^{N(1)} - \pi_i^{N(2)}) \qquad \text{(II.3)}$$

$$\sum_{t^0+1}^{t^*} \delta_i^t (\pi_i^* - \pi_i^{N(1)} + \pi_i^{N(1)} - \pi_i^{N(2)}) + \sum_{t^*+1}^{T} \delta_i^t (\pi_i^{N(1)} - \pi_i^{N(2)}) \geq \delta_i^{t^0} (\pi_i^D - \pi_i^*) \qquad \text{(II.4)}$$

or

$$\sum_{t^0+1}^{t^*}\delta_i^t(\pi_i^*-\pi_i^{N(1)})+\sum_{t^0+1}^{T}\delta_i^t(\pi_i^{N(1)}-\pi_i^{N(2)})\geq\delta_i^{t^0}(\pi_i^D-\pi_i^*) \tag{II.5}$$

or making use of the formulae given in (4.2) and (4.3) in the text we get:

$$\delta_i^{t^0+1}\frac{1-\delta_i^{t^*-t^0}}{1-\delta_i}(\pi_i^*-\pi_i^{N(1)})+\delta_i^{t^0+1}\frac{1-\delta_i^{T-t^0}}{1-\delta_i}(\pi_i^{N(1)}-\pi_i^{N(2)})\geq\delta_i^{t^0}(\pi_i^{\mathrm{D}}-\pi_i^*). \tag{II.6}$$

Dividing through by $\delta_i^{t^0}$ and rearranging leads to:

$$\delta_i[(1-\delta_i^{t^*-t^0})(\pi_i^*-\pi_i^{N(1)})+(1-\delta_i^{\mathrm{T}-t^0})(\pi_i^{N(1)}-\pi_i^{N(2)})]\geq(1-\delta_i)(\pi_i^{\mathrm{D}}-\pi_i^*) \tag{II.7}$$

or

$$\delta_i[\pi_i^*-\pi_i^{N(1)}-\delta_i^{t^*-t^0}(\pi_i^*-\pi_i^{N(1)})+\pi_i^{N(2)}-\pi_i^{N(1)}-\delta_i^{\mathrm{T}-t^0}(\pi_i^{N(1)}-\pi_i^{N(2)})+\pi_i^{\mathrm{D}}-\pi_i^*]\geq\pi_i^{\mathrm{D}}-\pi_i^*. \tag{II.8}$$

Dividing through by the expression in the square brackets gives:

$$\delta_i\geq\frac{\pi_i^{\mathrm{D}}-\pi_i^*}{\pi_i^{\mathrm{D}}-\pi_i^{N(1)}-\delta_i^{t^*-t^0}(\pi_i^*-\pi_i^{N(2)})-\delta_i^{T-t^0}(\pi_i^{N(1)}-\pi_i^{N(2)})}\ \text{if}\ t^0<t^* \tag{II.9}$$

$$\delta_i\geq\frac{\pi_i^{\mathrm{D}}-\pi_i^*}{\pi_i^{\mathrm{D}}-\pi_i^*+\pi_i^{N(1)}-\pi_i^{N(2)}-\delta_i^{T-t^0}(\pi_i^{N(1)}-\pi_i^{N(2)})}\ \text{if}\ t^0=t^*. \tag{II.10}$$

For large T (II.9) and (II.10) become (see the Annex below):

$$\delta_i\geq\frac{\pi_i^{\mathrm{D}}-\pi_i^*}{\pi_i^{\mathrm{D}}-\pi_i^*+\pi_i^{N(1)}-\pi_i^{N(2)}}=\delta_i^{\min}\ \forall\ \pi_i^*>\pi_i^{N(1)}\wedge\forall\ i\in\mathrm{I} \tag{II.11}$$

$$\delta_i\geq\frac{\pi_i^{\mathrm{D}}-\pi_i^*}{\pi_i^{\mathrm{D}}-\pi_i^{N(2)}}=\delta_i^{\min}\forall\ \pi_i^{N(1)}\geq\pi_i^*\geq\pi_i^{N(2)}\wedge\forall\ i\in\mathrm{I}. \tag{II.12}$$

That is, the RHS terms in (II.11) and (II.12) constitute upper bounds for the RHS expressions (II.9) and (II.10). Notice that (II.11) and (II.12) can generally be satisfied for any $\pi_i^*(s)\geq\pi_i^{N(2)}\ \forall\ i\in\mathrm{I}$ and δ_i close to 1 and therefore σ_i can be a subgame-perfect strategy. It remains to be shown that $\pi_i^*(s)$ is the average payoff of such a strategy provided $\mathrm{T}\to\infty$ (and $\delta_i\to1$). For this suppose t^0, t^* and δ_i are given. Further let t_i^{R} be the smallest integer value of $\mathrm{T}-t^*$ that satisfies (II.9) or (II.10), whichever is appropriate, and define $t^{\mathrm{R}}:=\max t_i^{\mathrm{R}}$. That is, t^{R} is the minimum time span required at the end of the game, such that the trigger strategy σ_i is an SPE, that is, $\mathrm{T}\geq\mathrm{T}^{\min}=t^*+t^{\mathrm{R}}$. Then, it is possible to choose $t^*=\mathrm{T}-t^{\mathrm{R}}$ and the fraction of periods in which the cooperative strategy is played in relation to T is given by:

$$\mu = t^*/\mathrm{T} \Leftrightarrow \frac{\mathrm{T} - t^{\mathrm{R}}}{\mathrm{T}} \Leftrightarrow 1 - \frac{t^{\mathrm{R}}}{T} \Rightarrow 1 \text{ for } \mathrm{T} \to \infty. \tag{II.13}$$

Hence, for large T the fraction of periods in which the cooperative outcome will be obtained approaches 1. Hence, the formula of the average payoff as given in (4.5) in the text applies (assuming $\mathrm{T} \to \infty$ and $\pi_1 = \ldots = \pi_{\mathrm{T}}$) and $\pi_i^*(s)$ is the average payoff to player i from such a strategy.

To adapt the proof for sequential move games the stage game strategies $s_i^{N(1)}$ and $s_i^{N(2)}$ have simply to be replaced by subgame-perfect strategies $s_i^{\mathrm{SPE}(1)}$ and $s_i^{\mathrm{SPE}(2)}$ respectively.

Finally, we have to discuss games where $\pi_i^{N(1)} > \pi_i^{N(2)}$ ($\pi_i^{\mathrm{SPE}(1)} > \pi_i^{\mathrm{SPE}(2)}$) does not hold for all players but where one or some stage game equilibria are good equilibria for some players but at the same time bad equilibria for some other players. Then in the last periods $\mathrm{T} - t^*$ each good equilibrium must be played at least once, otherwise punishment is not a deterrent. A typical sequence at the end of play would be $\pi^{(\mathrm{B})}_{it^*+1}$, $\pi^{(\mathrm{B})}_{it^*+2}$, …, $\pi^{(\mathrm{G})}_{it^*+m}$, where the superscript B stands for bad equilibrium, G for good equilibrium and m for the length of the sequence. This implies that if a player deviates, say, at time $t^0 = t^*$, punishment might not hurt this player immediately but m rounds later, because at times $t^* + 1$ to $t^* + (m-1)$ a bad equilibrium from his/her point of view would have been played anyway. Nevertheless, Theorem 4.4 holds as long as T is large enough.

To see this, note first that for large discount factors a player deviates, if at all, at time $t^0 = t^*$. Then, s/he nets a gain of $\pi_i^{\mathrm{D}} - \pi_i^*$. Punishment involves a loss of $\pi_i^{\mathrm{G}} - \pi_i^{\mathrm{B}}$ at time $t^* + m$, $t^* + 2m$, $t^* + 3m$, …, $t^* + h{\cdot}m$, where h is the biggest integer value of $(T - t^*)/m$. Discounting all payoffs to time t^* (or setting $t^* = 0$ above), we therefore must have:

$$\sum_{l=1}^{h} \delta_i^{h \cdot l} (\pi_i^{\mathrm{G}} - \pi_i^{\mathrm{B}}) \geq (\pi_i^{\mathrm{D}} - \pi_i^*) \Leftrightarrow (\pi_i^{\mathrm{G}} - \pi_i^{\mathrm{B}}) \left[\frac{\delta_i^m}{(1 - \delta_i^m)} + \frac{(\delta_i^m)^{h+1}}{(1 - \delta_i^m)} \right] \geq (\pi_i^{\mathrm{D}} - \pi_i^*) \tag{II.14}$$

to deter free-riding where $l \in \{1, \ldots, h\}$ is an index. According to the definition of h, $\mathrm{T} \to \infty$ implies $h \to \infty$. Consequently, (II.14) becomes:

$$(\pi_i^{\mathrm{G}} - \pi_i^{\mathrm{B}}) \frac{\delta_i^m}{(1 - \delta_i^m)} \geq (\pi_i^{\mathrm{D}} - \pi_i^*) \tag{II.15}$$

for large T. Since $\delta_i^m/(1 - \delta_i^m)$ approaches infinity for $\delta_i \to 1$, (II.15) can always be satisfied regardless of how large is the free-rider gain $\pi_i^{\mathrm{D}} - \pi_i^*$ and how long the sequence m is. QED

Annex

To show that for large T ($T\to\infty$) the RHS terms in (II.9) and (II.10) are bounded from above by the RHS expression in (II.11) if $\pi_i^* > \pi_i^{N(1)}$ and by (II.12) if $\pi_i^{N(2)} \le \pi_i^* \le \pi_i^{N(1)}$. Suppose first that $\pi_i^{N(2)} \le \pi_i^* \le \pi_i^{N(1)}$ and $t^0 < t^*$ is true. Then, we have to show that:

$$\frac{\pi_i^D - \pi_i^*}{\pi_i^D - \pi_i^{N(2)}} \ge \frac{\pi_i^D - \pi_i^*}{\pi_i^D - \pi_i^{N(2)} - \delta_i^{t^*-t^0}(\pi_i^* - \pi_i^{N(1)}) - \delta_i^{T-t^0}(\pi_i^{N(1)} - \pi_i^{N(2)})} \quad \text{(II.16]}$$

for $T\to\infty$. That is:

$$\pi_i^D - \pi_i^{N(2)} \le \pi_i^D - \pi_i^{N(2)} - \delta_i^{t^*-t^0}(\pi_i^* - \pi_i^{N(1)}) - \delta_i^{T-t^0}(\pi_i^{N(1)} - \pi_i^{N(2)}) \quad \text{(II.17)}$$

for $T\to\infty$ or:

$$-\delta_i^{t^*-t^0}(\pi_i^* - \pi_i^{N(1)}) - \delta_i^{T-t^0}(\pi_i^{N(1)} - \pi_i^{N(2)}) \ge 0 \quad \text{(II.18)}$$

for $T\to\infty$. Since $-\delta_i^{t^*-t^0}(\pi_i^* - \pi_i^{N(1)}) \ge 0$ due to $\pi_i^* \le \pi_i^{N(1)}$ by assumption and $\delta_i^{T-t^0}(\pi_i^{N(1)} - \pi_i^{N(2)}) \to 0$ for large T ($\delta_i^{T-t^0} \to 0$ for $T\to\infty$ because $\delta_i < 1$), (II.18) is satisfied. If $t = t^0$, $-\delta_i^{t^*-t^0}(\pi_i^* - \pi_i^{N(1)})$ becomes $-(\pi_i^* - \pi_i^{N(1)}) \ge 0$, and, again (II.18) holds.

Next consider $\pi_i^* > \pi_i^{N(1)}$ and $t^0 < t^*$. Then, we have to show that:

$$\frac{\pi_i^D - \pi_i^*}{\pi_i^D - \pi_i^{N(2)} - \pi_i^* + \pi_i^{N(1)}} \ge$$

$$\frac{\pi_i^D - \pi_i^*}{\pi_i^D - \pi_i^{N(2)} - \delta_i^{t^*-t^0}(\pi_i^* - \pi_i^{N(1)}) - \delta_i^{T-t^0}(\pi_i^{N(1)} - \pi_i^{N(2)})} \quad \text{(II.19)}$$

if $T\to\infty$ which is equivalent to:

$$\pi_i^D - \pi_i^{N(2)} - \pi_i^* + \pi_i^{N(1)} \le \pi_i^D - \pi_i^{N(2)} - \delta_i^{t^*-t^0}(\pi_i^* - \pi_i^{N(1)}) - \delta_i^{T-t^0}(\pi_i^{N(1)} - \pi_i^{N(2)}) \quad \text{(II.20)}$$

if $T\to\infty$ or:

$$\pi_i^* + \pi_i^{N(1)} - \delta_i^{t^*-t^0}(\pi_i^* - \pi_i^{N(1)}) - \delta_i^{T-t^0}(\pi_i^{N(1)} - \pi_i^{N(2)}) \ge 0 \quad \text{(II.21)}$$

if $T\to\infty$. Note that $\pi_i^* - \pi_i^{N(1)} > 0$ by assumption and $\pi_i^* - \pi_i^{N(1)} - \delta_i^{t^*-t^0}(\pi_i^* - \pi_i^{N(1)}) > 0$ because $\delta_i < 1$. Again, $-\delta_i^{T-t^0}(\pi_i^{N(1)} - \pi_i^{N(2)})$ approaches zero from below. Taken together, this shows that (II.21) holds if $t^0 < t^*$. Moreover, if we let $t^0 \to t^*$ in (II.21) we can always find a T large enough so

that $\pi_i^* + \pi_i^{N(1)} - \delta_i^{*-t^0}(\pi_i^* - \pi_i^{N(1)}) \to 0 > \delta_i^{T-t^0}(\pi_i^{N(1)} - \pi_i^{N(2)}) \to 0$. This completes the proof. QED

III CHAPTER 5: APPENDIX

The objective of this appendix is to provide a detailed and intuitive proof of Folk Theorem V. The three-phase strategy, as laid out in Section 5.2, works as follows:

Phase 1 Play the cooperative phase strategy s_i as long as nobody deviates. Players receive the stage game payoff $\pi_i^*(s)$.

Phase 2 If a player i deviates from s_i start punishment by minimaxing him/her for t_i^P periods. If s/he or any other player deviates in phase 2, restart phase 2. In the following, the minimax payoff to player i, $\pi_i^{M(i)}(m_i^i, m_{-i}^i)$, will be normalized to zero for convenience.

Phase 3 If there was no deviation in phase 2 for t_i^P (if player i was punished) or t_j^P (if any player j among the punishers was punished) periods, then play strategy s_i' for the rest of the game which gives a payoff of $\pi_i^{*\prime} + \varepsilon$ to each player i, except to the last deviator j in phase 2 who receives only $\pi_j^{*\prime}, \pi_j^{*\prime} < \pi_j^*$. Moreover, $\pi^{*\prime} = (\pi_1^{*\prime}, \ldots, \pi_N^{*\prime}) \in \Pi^{IR}$ and therefore $(\pi_1^{*\prime} + \varepsilon, \pi_2^{*\prime} + \varepsilon, \ldots, \pi_j^{*\prime}, \pi_N^{*\prime} + \varepsilon) \in \Pi^{IR}$, too. If any player deviates in phase 3, restart phase 2.

In order to check whether this strategy profile is subgame-perfect, we have to demonstrate that deviation in each phase does not pay. We consider phases 1, 2 and 3 sequentially.

Phase 1

In this phase each player receives a payoff of π_i^* in each stage. If s/he deviates s/he can net a payoff of at most π_i^U in the first period. Then, s/he is minimaxed for t_i^P periods and receives a payoff of zero. From period $t_i^P + 1$ until perpetuity s/he receives a payoff of $\pi_i^{*\prime}$ if s/he complies in phase 2. Since $\pi_i^{*\prime} \geq \pi_i^M = 0$ by assumption, compliance is more attractive than deviation in phase 2. Thus, the maximal gain from a deviation in phase 1 is given by:

$$\pi_i^U + 0 \cdot \sum_{t=1}^{t_i^P} \delta_i^t + \pi_i^{*\prime} \cdot \sum_{t=t_i^P+1}^{\infty} \delta_i^t - \pi_i^* \cdot \sum_{t=0}^{\infty} \delta_i^t \Leftrightarrow \pi_i^U + \frac{\delta_i^{t_i^P+1}}{1-\delta_i} \cdot \pi_i^{*\prime} - \frac{\pi_i^*}{1-\delta_i}. \quad \text{(III.1)}$$

Since $\pi_i^* > \pi_i^{*\prime}$, this *maximum gain* is bounded from above by:

$$\pi_i^{\mathrm{U}} + \frac{\delta_i^{t_i^P+1}}{1-\delta_i} \cdot \pi_i^{*\prime} - \frac{\pi_i^{*\prime}}{1-\delta_i} \Leftrightarrow \pi_i^{\mathrm{U}} - \frac{1-\delta_i^{t_i^P+1}}{1-\delta_i} \cdot \pi_i^{*\prime}. \tag{III.2}$$

Equation (III.2) can generally be satisfied for sufficiently big discount factors and sufficiently long punishment durations. To see this, note that

$$\lim_{\delta_i \to 1} \frac{1-\delta_i^{t_i^P+1}}{1-\delta_i} = t_i^P + 1.$$

Then (III.2) reads $\pi_i^{\mathrm{U}} - (t_i^{\mathrm{P}} + 1)\pi_i^{*\prime}$ which is negative provided:

$$\frac{\pi_i^{\mathrm{U}}}{\pi_i^{*\prime}} \leq [t_i^{\mathrm{P}}] + 1 \tag{III.3}$$

holds, where the bracket indicates that t_i^{P} must be an integer value. That is, t_i^{P} is sufficiently long that deviation in phase 1 is not attractive. Thus, condition (III.3) is a *sufficient* condition to ensure negative gains from deviation, which is assumed to hold in the remainder.

Phase 2

As pointed out above, it is not attractive for a deviator in phase 1 to continue with deviation in phase 2 because the maximum payoff s/he nets is zero and phase 2 is only prolonged. In contrast, if s/he complies in phase 2 s/he will eventually receive a payoff of $\pi_i^{*\prime} \geq 0$ in phase 3.

A player j who is a punisher in phase 2 receives a payoff of $\pi_j^{\mathrm{M}(i)}(m_j^i, m_{-j}^i)$ for t_i^{P} periods and in phase 3 a payoff of $\pi_j^{*\prime} + \varepsilon$ in each period for the rest of the game if s/he complies. Alternatively, if s/he does not fulfill her punishment obligations in phase 2, s/he receives at most a payoff of π_j^{U} in the first period and then s/he is minimaxed him- or herself for t_j^{P} periods (phase 2 starts anew) which, as argued above, s/he will accept in order to reach phase 3 finally. Therefore, the discounted payoff stream of a deviation in phase 2 is given by:

$$\pi_j^{\mathrm{U}} + 0 \cdot \sum_{t=1}^{t_j^{\mathrm{P}}} \delta_j^t + \pi_j^{*\prime} \cdot \sum_{t=t_j^{\mathrm{P}}+1}^{\infty} \delta_j^t \Leftrightarrow \pi_j^{\mathrm{U}} + \frac{\delta_j^{t_j^{\mathrm{P}}+1}}{1-\delta_j} \cdot \pi_j^{*\prime}. \tag{III.4}$$

Since $\delta_j \leq 1$ and hence $\delta_j^{t_j^{\mathrm{P}}+1} \leq 1$ holds, an upper bound for this deviation payoff is given by:

$$\pi_j^{\mathrm{U}} + \frac{\pi_j^{*\prime}}{1-\delta_j}. \tag{III.5}$$

Thus, the maximum gain from a deviation in phase 2 does not exceed:

$$\pi_j^{\mathrm{U}} + \frac{\pi_j^{*\prime}}{1-\delta_j} - \pi_j^{\mathrm{M}(i)} \cdot \sum_{t=0}^{t_j^{\mathrm{P}}-1} \delta_j^t - (\pi_j^{*\prime} + \varepsilon) \sum_{t=t_j^{\mathrm{P}}}^{\infty} \delta_j^t \Leftrightarrow$$

$$\pi_j^{\mathrm{U}} + (\pi_j^{*\prime} - \pi_j^{\mathrm{M}(i)}) \frac{1-\delta_j^{t_j^{\mathrm{P}}}}{1-\delta_j} - \varepsilon \frac{\delta_j^{t_j^{\mathrm{P}}}}{1-\delta_j}. \tag{III.6}$$

Now let $\delta_j \to 1$, then $(1-\delta_j^{t_j^{\mathrm{P}}})/(1-\delta_j)$ approaches t_j^{P} from below. That is, the first two terms are finite. However, the third term goes to minus infinity for $\delta_j \to 1$, so that (III.6) becomes negative. Thus if the discount factor is sufficiently close to 1, deviation in phase 2 does not pay a punisher.

Phase 3

Last but not least, we have to check for the incentive to deviate in phase 3. In this phase a player who was punished previously in phase 2 will not deviate in phase 3 because otherwise phase 2 is started again. Whereas compliance in phase 3 gives him/her a payoff of $\pi_i^{*\prime}$, deviation gives him/her at most a payoff of π_i^{U} in the first period and subsequently the payoffs of phases 2 and 3. Thus, the gain from a deviation is given by:

$$\pi_i^{\mathrm{U}} + \frac{\delta_i^{t_i^P+1}}{1-\delta_i} \cdot \pi_i^{*\prime} - \frac{\pi_i^{*\prime}}{1-\delta_i} \Leftrightarrow \pi_i^{\mathrm{U}} - \frac{1-\delta_i^{t_i^P+1}}{1-\delta_i} \cdot \pi_i^{*\prime} \tag{III.7}$$

which is equivalent to (III.2) and negative by (III.3). For a punisher *j* who complied with his/her punishment obligations in phase 2, the incentive to deviate in phase 3 is given by:

$$\pi_j^{\mathrm{U}} + \frac{\delta_j^{t_j^{\mathrm{P}}+1}}{1-\delta_j} \cdot \pi_j^{*\prime} - \frac{\pi_j^{*\prime} + \varepsilon}{1-\delta_j} \tag{III.8}$$

which is obviously smaller than (III.7) (of course, *i* has to be replaced by *j* in (III.7)) and hence deviation does not pay either.

Taken together, deviation does not pay in phases 1, 2 or 3, provided discount factors are close to 1. Choosing $\pi_i^{*\prime}$ sufficiently close to π_i^{M} and letting $\varepsilon \to 0$, any payoff vector $\pi^* \in \Pi^{\mathrm{IR}}$ can be obtained with $\pi_i^{*\prime} \to \pi_i^{\mathrm{M}}$ at the limit of $\delta_i \to 1$. QED

IV CHAPTER 7: APPENDICES

IV.1 Appendix 1

In Section 7.1.1 we claim that for δ_i close to 1 (implying $\delta_i^{t_i^{\mathrm{P}}} \to 1$) (7.3) and (7.4) in the text reduce to (7.7).

Proof:

(a) Assume $s_i \neq s_i(s_j)$. Then $\pi_i^D > \pi_i^*$ and from (7.4) $\pi_i^P < \pi_i^*$ follows. Accordingly, in (7.6) $\pi_i^R < \pi_i^*$ must be true. Consequently for $\delta_i \to 1$, π_i^P approaches π_i^* from below in (7.6), so that (7.3) becomes $\pi_i^* > \pi_i^C$, which is the first possibility in (7.7).

(b) If $s_i = s_i(s_j)$, then $\pi_i^D = \pi_i^*$. In words: s_j is a stage game Nash strategy and therefore the best deviation strategy is $s_i^N = s_i(s_j^N)$. Consequently, $\pi_i^D = \pi_i^N$. Since there is no incentive for player i to deviate in the co-operative phase, the strategy tuple $s = (s_i, s_j)$ may also be chosen as punishment for player i. Hence, $\pi_i^* = \pi_i^C$ is possible. QED

IV.2 Appendix 2

Theorem 7.2 in Sub-section 7.1.1 claims: in an infinitely repeated (ordinary) PD game with stage game payoffs as given in the General Payoff Matrix 3.2, assuming $c_i > a_i > d_i > b_i$, $a_1 + a_2 > b_1 + c_2$, $a_1 + a_2 > b_2 + c_1$ and $a_1 + a_2 > d_1 + d_2$, $\overline{\Pi}^{SPE} = \overline{\Pi}^{WRPE}$ for $\delta_i \to 1 \; \forall \; i \in I$.

Proof: The proof is illustrated with the help of Figure IV.1. The WRPE conditions assuming player 1 to be the potential deviator are given by:

$$\pi_1^* \geq \pi_1^C(p_1, q_1^1) \tag{IV.1}$$

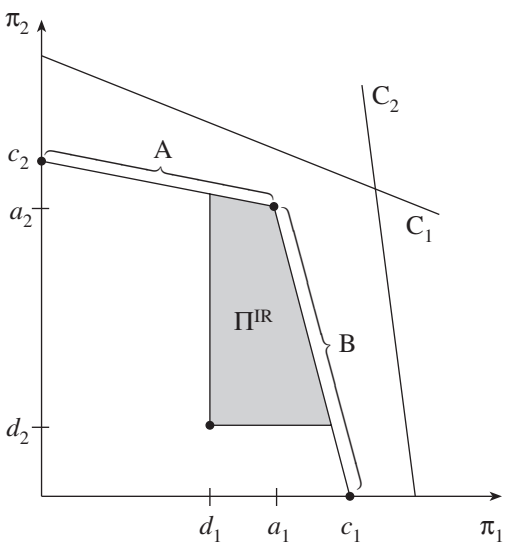


Figure IV.1 Weakly renegotiation-proof payoff space in the PD game

$$\pi_2^* \leq \pi_2^R(p_1^1, q_1^1). \tag{IV.2}$$

A dominant deviation strategy is $p_1 = 0$ and the RHS term in (IV.2) is maximized for $p_1^1 = 1$. Then using the notation of the General Payoff Matrix 3.2, we have:

$$\pi_1^* \geq c_1 q_1^1 + d_1(1 - q_1^1) \tag{IV.3}$$

$$\pi_2^* \leq a_2 q_1^1 + c_2(1 - q_1^1). \tag{IV.4}$$

Assuming (IV.3) to be binding, we get:

$$q_1^1 + \frac{\pi_1^* - d_1}{c_1 - d_1} \tag{IV.5}$$

and upon substitution in (IV.4):

$$C_1 : = \pi_2^* \leq c_2 + d_1 \frac{(c_2 - a_2)}{(c_1 - d_1)} - \frac{(c_2 - a_2)}{(c_1 - d_1)} \cdot \pi_1^*. \tag{IV.6}$$

Without loss of generality, we may normalize payoffs such that $b_1 = b_2 = 0$. Then the line segment A of the Pareto frontier in Figure IV.1 may be expressed as:

$$\pi_2^* = c_2 - \frac{(c_2 - a_2)}{a_1} \cdot \pi_1^*. \tag{IV.7}$$

Since segment A of the Pareto frontier and condition C_1 are both straight lines, it suffices to show that line C_1 lies above line segment A at two points, for example, $\pi_1^* = d_1$ and $\pi_1^* = a_1$. For this we compute the difference between line C_1 and line A to be:

$$C_1 - A = \frac{(c_2 - a_2)(a_1 d_1 + \pi_1^*(c_1 - a_1 - d_1))}{a_1(c_1 - d_1)}. \tag{IV.8}$$

Substituting $\pi_1^* = d_1$ in (IV.8) gives $C_1 - A = d_1(c_2 - a_2)/a_1 > 0$ and for $\pi_1^* = a_1$ we find $C_1 - A = ((c_2 - a_2)(c_1 - a_1))/(c_1 - d_1) > 0$.

A similar procedure establishes $C_2 - B$ and the claim above is proved. QED

IV.3 Appendix 3

In Sub-section 7.1.3 we claim that with the help of a two-phase punishment strategy $\delta_i^{\min}$ (WRPE) $= \delta_i^{\min}$ (SPE) in an ordinary PD game. The intention of this appendix is to demonstrate this for the example discussed in that section (Matrix 3.1). Since the game is symmetric, $\delta_i^{\min}$ (WRPE) is the same for both players and we only have to consider the case if, say, player 1 is the potential defector.

In the particular example $\pi^* = (3.2, 3.2)$ and in the first phase $\pi^{R(1)} = (p_1^1 = p = 1, q_1^1 = 0) = (1.4, 4.4)$. Moreover, let $\pi^I(p_1^1 = p, q_1^1 = q)$ denote the 'intermediate payoff tuple' of the second phase. Then, the average continuation punishment payoffs to players 1 and 2 are given by:

$$\begin{aligned}\pi_1^P &= (1-\delta_1^{t_1^P})\pi_1^{R(1)} + \delta_1^{t_1^P}(1-\delta_1)\pi_1^I + \delta_1^{t_1^P+1}\pi_1^* \\ \pi_2^P &= (1-\delta_2^{t_1^P})\pi_2^{R(1)} + \delta_2^{t_1^P}(1-\delta_2)\pi_2^I + \delta_2^{t_1^P+1}\pi_2^*. \end{aligned} \tag{IV.9}$$

Now there are *two* possibilities for a deviation during the punishment, either in phase 1 or phase 2. Hence, compliance with the punishment (repentance phase) must be ensured in both phases:

$$C_1 := \pi_1^P \geq \pi_1^C(p_1 = 1, q_1^1 = 0) \tag{IV.10}$$

$$C_2 := (1-\delta_1)\pi_1^I(p_1^1 = p, q_1^1 = q) + \delta_1\pi_1^* \geq (1-\delta_1)\pi_1^C(p_1 = 1, q_1^1 = q) + \delta_1\pi_1^P \tag{IV.11}$$

must hold. (IV.10) is the familiar condition (7.3) in the text which ensures that deviation in phase 1 does not occur. (IV.11) is a similar condition for phase 2. Condition C_2 states that the average payoff to player 1 in phase 2 and subsequently resuming cooperation must be greater than deviating in phase 2 (recall $p_1 = 1$ is a dominant defection strategy in this game) and then starting punishment anew.

Of course, deviation during the cooperative phase should not pay either (see inequality (7.4) in the text):

$$C_3 := \pi_1^* \geq (1-\delta_1)\pi_i^D + \delta_1\pi_1^P. \tag{IV.12}$$

Finally, we have to take care of condition (7.5) in the text. Using (IV.9) we have:

$$C_4 := \pi_2^* \leq (1-\delta_2^{t_1^P})\pi_2^{R(1)} + \delta_2^{t_1^P}(1-\delta_2)\pi_2^I + \delta_2^{t_1^P+1}\pi_2^*. \tag{IV.13}$$

We start solving for δ_i^{min} by assuming $t_1^P = 1$, $p_1^1 = p = 1$ and $q_1^1 = q = 1/3$ in phase 2. Then $\pi_1^I = 2$, $\pi_2^I = 4$, $\pi_1^C(p = 1, q_1^1 = 0) = 2$ and $\pi_1^C(p = 1, q_1^1 = 1/3) = 2.8$. Consequently, condition C_1 implies $\delta_1 \geq 0.5$, C_2 $\delta_1 \geq 0.359$ and C_3 $\delta_1 \geq 0.5$. C_4 is always satisfied for any $0 \leq \delta_2 \leq 1$. Hence, δ_1^{min} (WRPE) $= 0.5$ and by symmetry δ_2^{min} (WRPE) $= 0.5$. It is easily checked that if we substitute $\delta_1 = 0.5$ into (IV.9) $\pi_1^P = \pi_1^M = 2$ from which δ_i^{min} (WRPE) $= \delta_i^{min}$ (SPE) follows.

V CHAPTER 8: APPENDIX

In Proposition 8.3 we claim that issue linkage improves upon the chances for cooperation if issues are substitutes in governments' objective function and that cooperation becomes more difficult if issues are complements.

Proof: Since Proposition 8.3 focuses on the Pareto-efficient cooperative stage game strategy tuple (a_1, a_2), $\delta_i^{\min}(\text{SPE}) = \delta_i^{\min}(\text{SRPE})$ by Theorem 7.4.

Linking a New Issue to an Existing Issue

As laid out in the text, four sub-cases have to be distinguished.

Sub-case 1
Governments are cooperating on issue 1 if:

$$u_i(a_{i1}) \geq (1-\delta_i)u_i(c_{i1}) \Leftrightarrow \frac{\delta_i}{(1-\delta_i)}u_i(a_{i1}) \geq u_i(c_{i1}) - u_i(a_{i1}) \;\forall\, i \in \text{I} \quad \text{(V.1)}$$

holds. If the new issue 2 is viewed as a separate game, independent of game I, cooperation can be sustained in game II provided:

$$u_i(a_{i1},a_{i2}) - (1-\delta_i)u_i(a_{i1},c_{i2}) - \delta_i u_i(a_{i1}) \geq 0 \quad \text{(V.2)}$$

holds where in (V.2) and in all following formulae we skip $\forall\, i \in \text{I}$ for convenience. That is, (V.2) expresses that a country deviates only with respect to the second issue (second term) and is also only punished with respect to this issue (third term). In contrast, if both issues are linked, a sufficient condition for cooperation is:[4]

$$u_i(a_{i1},a_{i2}) - (1-\delta_i)u_i(c_{i1},c_{i2}) \geq 0. \quad \text{(V.3)}$$

This implies that deviation as well as punishment will be conducted with respect to both issues. In order to find out whether issue linkage eases cooperation on issue 2, we subtract the LHS term in (V.2) from the LHS term in (V.3). Upon division by $(1-\delta_i)$ we obtain:

$$\frac{\delta_i}{(1-\delta_i)}u_i(a_{i1}) - [u_i(c_{i1},c_{i2}) - u_i(a_{i1},c_{i2})]. \quad \text{(V.4)}$$

From (V.1) it is known that a lower bound for $(\delta_i/(1-\delta_i))u_i(a_{i1})$ is $u_i(c_{i1}) - u_i(a_{i1})$ which may be substituted into (V.4):

$$[u_i(c_{i1}) - u_i(a_{i1})] - [u_i(c_{i1},c_{i2}) - u_i(a_{i1},c_{i2})]. \quad \text{(V.5)}$$

In other words, the effect of issue linkage is evaluated assuming (V.1) to be binding (knife-edge case; there is no slackening of enforcement power). Now if both issues are substitutes, that is, $\partial^2 u_i/\partial x_{i1}\partial x_{i2} < 0$ holds, the first term in brackets is greater than the second term, and (V.5) is positive. This implies that cooperation will be easier to sustain in game II if both issues are linked. By the same token, if both issues are complements (V.5) will be negative and cooperation in the linked game II is more difficult to sustain.

In order to analyze what changes for game I after the new issue 2 emerges and both games are linked, we compare the condition in the initial situation of game I, (V.1), with the issue linkage condition (V.3). For this we rewrite (V.1) to have:

$$u_i(a_{i1}) - (1-\delta_i)u_i(c_{i1}) \geq 0. \tag{V.1′}$$

and subtract (V.1′) from (V.3). Rearranging terms, we get:

$$\frac{\delta_i}{(1-\delta_i)}[u_i(a_{i1},a_{i2}) - u_i(a_{i1})] + [u_i(c_{i1}) - u_i(a_{i1})] - [u_i(c_{i1},c_{i2}) - u_i(a_{i1},a_{i2})]. \tag{V.6}$$

To sign this expression, we rewrite (V.2) as:

$$\frac{\delta_i}{(1-\delta_i)}[u_i(a_{i1},a_{i2}) - u_i(a_{i1})] \geq u_i(a_{i1},c_{i2}) - u_i(a_{i1},a_{i2}) \tag{V.2′}$$

from which it is evident that a lower bound for the first term in brackets in (V.6) is the RHS expression in (V.2′). Upon substitution we get:

$$u_i(a_{i1},c_{i2}) - u_i(a_{i1},a_{i2}) + [u_i(c_{i1}) - u_i(a_{i1})] - [u_i(c_{i1},c_{i2}) - u_i(a_{i1},a_{i2})]. \tag{V.7}$$

Collecting terms show that (V.7) is equivalent to (V.5) which we have signed already. Hence, again, if issues are substitutes (complements) linking makes cooperation easier (more difficult) on issue 1.

Sub-case 2

Next we have to clarify how issue linkage affects issues 1 and 2 if on issue 1 cooperation cannot be sustained on its own but on issue 2 it can. In this case we may write condition (V.1) as:

$$\delta_i + \varepsilon_i = \frac{u_i(c_{i1}) - u_i(a_{i1})}{u_i(c_{i1})} \tag{V.8}$$

where ε_i is some positive number which might be interpreted as a slack variable. Accordingly, the issue linkage condition (V.3) becomes a strict inequality:

$$\delta_i + \varepsilon_i > \frac{u_i(c_{i1}, c_{i2}) - u_i(a_{i1}, a_{i2})}{u_i(c_{i1}, c_{i2})}. \tag{V.9}$$

To show that (V.8) implies a higher minimum discount factor requirement if issues are substitutes, we insert (V.8) into (V.9) to have:

$$\frac{u_i(c_{i1}) - u_i(a_{i1})}{u_i(c_{i1})} > \frac{u_i(c_{i1}, c_{i2}) - u_i(a_{i1}, a_{i2})}{u_i(c_{i1}, c_{i2})} \tag{V.10}$$

which turns out to be equivalent to:

$$\frac{u_i(a_{i1}, a_{i2})}{u_i(a_{i1})} > \frac{u_i(c_{i1}, c_{i2})}{u_i(c_{i1})}. \tag{V.11}$$

Unfortunately, it is not possible to verify (V.11) at a general level. In the RHS term two variables have changed compared to the LHS term so that the levels of utility are not directly comparable. Thus to make further progress we have to reduce the number of variables to one. In particular we may focus on games where $c_{i1}/a_{i1} = c_{i2}/a_{i2} (\Leftrightarrow a_{i1}/a_{i2} = c_{i1}/c_{i2})$ holds. That is, deviations are proportional in both games. Using $a_i = (a_{i1} + a_{i2})$ and $\gamma_i = a_{i1}/(a_{i1} + a_{i2})$ $(\Rightarrow (1 - \gamma_i) = a_{i2}/(a_{i1} + a_{i2}))$ we can write the LHS term in (V.11) as:

$$\mathrm{A} := \frac{u_i(a_i)}{u_i(\gamma_i a_i)}. \tag{V.12}$$

Since $c_{i1} > a_{i1}$ and $c_{i2} > a_{i2}$ we have to show that A is decreasing in the payoff a_i. For this we look at some typical concave utility functions for which $\partial^2 u_i/\partial x_{ik}^2 < 0$ and $\partial^2 u_i/\partial x_{i1}\partial x_{i2} < 0$ is true. For instance, consider the utility function $u_i = \theta a_i - \frac{1}{2}\omega a_i^2$. Then we find for (V.12):

$$\mathrm{A} := \frac{\theta a_i - \frac{1}{2}\omega a_i^2}{\theta\gamma a_i - \frac{1}{2}\omega(\gamma a_i)^2}; \quad \frac{\partial A}{\partial a_i} = -\frac{2\theta\omega(1-\gamma)}{\gamma(2\theta - \omega a_i\gamma)^2} < 0. \tag{V.13}$$

For other typically frequently used utility functions, such as the exponential function $u_i = (1/\alpha) - (1/\alpha)\mathrm{e}^{(-\alpha a_i)}$ or the logarithmic function $u_i = \ln(1 + a_i)$, $\partial \mathrm{A}/\partial a_i < 0$ holds as well and therefore (V.11) holds for a great class of utility functions with a substitutional relation between issues and homogeneous payoffs in the two games. This implies that linking issues also improves upon the condition for the existing issue 1 even though cooperation was not possible if this issue was viewed as a single game. A similar procedure would establish that the opposite holds if issues are complements.

Now we have to check the impact on issue 2 through issue linkage. If cooperation is not sustainable on issue 1 but on issue 2, that is, $x_{i1}=0$, the condition for cooperation on issue 2 reads:

$$\delta_i \geq \frac{u_i(c_{i2}) - u_i(a_{i2})}{u_i(c_{i2})}. \tag{V.14}$$

This condition has to be compared to issue linkage condition (V.3), which may be formulated as:

$$\delta_i \geq \frac{u_i(c_{i1}, c_{i2}) - u_i(a_{i1}, a_{i2})}{u_i(c_{i2}, c_{i2})}. \tag{V.3′}$$

To show that issue linkage reduces the discount factor requirement if issues are substitutes, we have to show that the RHS term in (V.3′) is smaller than the RHS term in (V.14). Thus:

$$\frac{u_i(c_{i2}) - u_i(a_{i2})}{u_i(c_{i2})} > \frac{u_i(c_{i1}, c_{i2}) - u_i(a_{i1}, a_{i2})}{u_i(c_{i1}, c_{i2})} \tag{V.15}$$

or

$$\frac{u_i(a_{i1}, a_{i2})}{u_i(a_{i2})} > \frac{u_i(c_{i1}, c_{i2})}{u_i(c_{i2})} \tag{V.16}$$

must be true, which is a condition similar to (V.11) and hence the same arguments apply.

Sub-case 3
This sub-case is symmetric to sub-case 2 and therefore a proof is omitted.

Sub-case 4
Here we assume that cooperation cannot be sustained on both issues if they are viewed as single issues. Thus:

$$u_i(a_{i1}) - (1-\delta_i)u_i(c_{i1}) < 0 \tag{V.17}$$

$$u_i(a_{i2}) - (1-\delta_i)u_i(c_{i2}) < 0 \tag{V.18}$$

are true, which of course also applies to the sum of (V.17) and (V.18). Thus, introducing a slack variable ε_i, we have alternatively:

$$\delta_i + \varepsilon_i = \frac{u_i(c_{i1}) + u_i(c_{i2}) - u_i(a_{i1}) - u_i(a_{i2})}{u_i(c_{i1}) + u_i(c_{i2})}. \tag{V.19}$$

This condition has to be compared with the issue linkage condition in the form (V.3′). Upon substitution we derive:

$$\frac{u_i(c_{i1}) + u_i(c_{i2}) - u_i(a_{i1}) - u_i(a_{i2})}{u_i(c_{i1}) + u_i(c_{i2})} > \frac{u_i(c_{i1}, c_{i2}) - u_i(a_{i1}, a_{i2})}{u_i(c_{i1}, c_{i2})} \quad \text{(V.20)}$$

or, rearranging terms, we get:

$$\frac{u_i(a_{i1}, a_{i2})}{u_i(a_{i1}) + u_i(a_{i1})} > \frac{u_i(c_{i1}, c_{i2})}{u_i(c_{i1}) + u_i(c_{i2})}. \quad \text{(V.21)}$$

Note that (V.21) is similar to condition (V.11). Using the same assumptions, we can write the LHS term in (V.21) as:

$$\mathrm{B} := \frac{u_i(a_i)}{u_i(\gamma_i a_i) + u_i((1-\gamma_i)a_i)}. \quad \text{(V.22)}$$

Again, using typical utility functions with a substitutional relation between issues would show that $\partial \mathrm{B}/\partial a_i < 0$, proving that (V.21) holds. Once more, the opposite would hold if issues are complements.

Linking Two Existing Issues

Sub-case 1

Cooperation on both single issues can be sustained. That is, if both games are not linked, condition (V.2) holds in game II and by analogy in game I we require:

$$u_i(a_{i1},a_{i2}) - (1-\delta_i)u_i(c_{i1},a_{i2}) - \delta_i u_i(a_{i2}) \geq 0. \quad \text{(V.23)}$$

If (V.2) and (V.23) hold, this implies that (V.2) + (V.23) must hold too. Thus, to analyze the effect issue linkage has on the stability requirements we have to find out whether (V.2) + (V.23) or (V.3) is more stringent. For this we compute (V.3) − ((V.2) + (V.23)):[5]

$$\frac{\delta_i}{(1-\delta_i)}[u_i(a_{i1}) + u_i(a_{i2}) - u_i(a_{i1},a_{i2})] + [u_i(c_{i1},a_{i2}) - u_i(a_{i1},a_{i2})] - [u_i(c_{i1},c_{i2}) - u_i(a_{i1},c_{i2})] \quad \text{(V.24)}$$

which is easily signed. If both issues are substitutes, then the first term is positive and the same applies to the difference between the second and the third terms. By the same token, the opposite holds if both issues are complements, and hence issue linkage has a positive (negative) effect on the stability of contracts if issues are viewed as substitutes (complements) in the objective function of governments.

Sub-cases 2, 3 and 4

Next consider the sub-case where cooperation on one single issue, say 2, can be sustained but not on issue 1. This is identical with sub-case 2 above, for

which the impact of issue linkage has been shown. Of course, by symmetry if issue 1 can be sustained as a single issue but not issue 2, sub-case 3 above applies. Finally, the sub-case where cooperation on single issues cannot be sustained is equivalent to sub-case 4 above, and we are done. QED

VI CHAPTER 9: APPENDICES

VI.1 Appendix 1

We claim in Section 9.3, (9.12), that:

$$\frac{1}{r_1''} = \frac{\partial^2 e_2}{\partial e_1^2} = -\frac{\beta_1''' \phi_1''^2 - \beta_1''^2 \phi_1'''}{(-\phi_1'')^3}. \tag{VI.1}$$

Proof: The inverse of the slope of country 1's reaction function may be written as:

$$\frac{1}{r_1'} = \frac{\partial e_2}{\partial e_1} = \frac{\beta_1'' - \phi_1''}{\phi_1''} = -\frac{\pi_{1e_1e_1}}{\pi_{1e_1e_2}} \tag{VI.2}$$

which follows from the first-order condition in the Nash equilibrium according to the implicit function rule. $\pi_{1e_1e_1}$ denotes the second derivative of the net benefit function with respect to e_1 and $\pi_{1e_1e_2}$ the cross-derivative and so on. Thus:

$$\frac{1}{r_1''} = \frac{\partial^2 e_2}{\partial e_1^2} = \frac{\partial\left(-\frac{\pi_{1e_1e_1}}{\pi_{1e_1e_2}}\right)}{\partial e_1} + \frac{\partial\left(-\frac{\pi_{1e_1e_1}}{\pi_{1e_1e_2}}\right)}{\partial e_2} \cdot \frac{\partial e_2}{\partial e_1} \tag{VI.3}$$

where the last term on the RHS may be replaced by (VI.2). Then after some basic manipulations we derive:

$$\frac{\partial^2 e_2}{\partial e_1^2} = -\frac{\pi_{1e_1e_1e_1} \cdot \pi_{1e_1e_2}{}^2 - 2\pi_{1e_1e_2e_2} \cdot \pi_{1e_1e_1} \cdot \pi_{1e_1e_2} + \pi_{1e_1e_2e_2} \cdot \pi_{1e_1e_1}{}^2}{(\pi_{1e_1e_2})^3} \tag{VI.4}$$

where:

$$\pi_{1e_1e_1} = \beta_1'' - \phi_1'',\ \pi_{1e_1e_1e_1} = \beta_1''' - \phi_1''',\ \pi_{1e_1e_2} = -\phi_1'',\ \pi_{1e_1e_1e_2} = -\phi_1''',$$
$$\pi_{1e_1e_2e_2} = -\phi_1''' \tag{VI.5}$$

so that, after collecting terms, we get (VI.1). QED

VI.2 Appendix 2

We claim in Section 9.7 that country 1's (country 2's) indifference curve (iso-welfare curve) is concave (convex) with respect to the origin. We prove this exemplarily for country 1's indifference curve. A proof for country 2 would follow exactly along the same lines.

Proof: Country 1's indifference curve is described by $I_1 := \pi_1(e_1, e_2) = \bar{\pi}_1$ which may be written in implicit form as $\pi_1(e_1, e_2) - \bar{\pi}_1 = 0$. In Section 9.7 (9.27), the slope of this function has already been derived in the $e_1 - e_2$ emission space, namely:

$$I_1' = \frac{\partial e_2}{\partial e_1} = \frac{\beta_1' - \phi_1'}{\phi_1'} = -\frac{\pi_{1e_1}}{\pi_{1e_2}} \tag{VI.6}$$

where the same notation as in Appendix VI.1 has been used. In order to show concavity of this function we compute the second-order derivative:

$$I_1'' = \frac{\partial^2 e_2}{\partial e_1^2} = \frac{\partial\left(-\frac{\pi_{1e_1}}{\pi_{1e_2}}\right)}{\partial e_1} + \frac{\partial\left(-\frac{\pi_{1e_1}}{\pi_{1e_2}}\right)}{\partial e_2} \cdot \frac{\partial e_2}{\partial e_1} \tag{VI.7}$$

where the last term on the RHS may be replaced by (VI.6). Proceeding in a similar way as in Appendix VI.1 we arrive at:

$$\frac{\partial^2 e_2}{\partial e_1^2} = -\frac{\pi_{1e_1e_1} \cdot \pi_{1e_2}^2 - 2\pi_{1e_1e_2} \cdot \pi_{1e_1} \cdot \pi_{1e_2} + \pi_{1e_2e_2} \cdot \pi_{1e_1}^2}{(\pi_{1e_2})^3} \tag{VI.8}$$

which, after substitution of:

$$\pi_{1e_1} = \beta_1' - \phi_1'; \ \pi_{1e_1e_1} = \beta_1'' - \phi_1''; \ \pi_{1e_2} = -\phi_1'; \ \pi_{1e_1e_2} = -\phi_1''; \ \pi_{1e_2e_2} = -\phi_1'' \tag{VI.9}$$

and collecting terms gives:

$$\frac{\partial^2 e_2}{\partial e_1^2} = -\frac{\phi_1'^2 \beta_1'' - \phi_1'' \beta_1'^2}{(-\phi_1')^3} < 0$$

due to $\phi_1' > 0$, $\phi_1'' > 0$, $\beta_1'^2 > 0$ and $\beta_1'' < 0$ which establishes strict concavity. In the case $\beta_1'' > 0$, β_1'' is required to be sufficiently small so as to ensure concavity. QED

VII CHAPTER 10: APPENDICES

VII.1 Appendix 1

In Section 10.2, note 3, we indicate that the SOC may not hold in general and therefore they must be assumed to hold. Moreover, it is claimed that if the SOC hold, then a unique equilibrium exists.

Proof: Taking the second derivative of the objective function of the leader as stated in (10.1) leads to:

$$\frac{\partial^2 \pi_i}{\partial e_i^2} = \underset{(-)}{\frac{\partial^2 \beta_i}{\partial e_i^2}} - \underset{(+)}{\frac{\partial^2 \phi_i}{\partial e_i^2}} - \underset{(+)\cdot(+)}{\frac{\partial^2 \phi_i}{\partial e_j^2}\left(\frac{\partial e_j}{\partial e_i}\right)^2} - \underset{(+)\cdot(?)}{\frac{\partial \phi_i}{\partial e_j}\frac{\partial^2 e_j}{\partial e_i^2}} \tag{VII.1}$$

from which it appears that all terms are negative except the last term which might be positive if $\partial^2 e_j/\partial e_i^2 < 0$. That is, if the second-order derivative of the reaction function is negative the whole term in (VII.1) could be positive. From Section 9.3 it is known that strictly concave reaction functions cannot be ruled out on theoretical grounds. Hence, to guarantee a unique maximum, either convex reaction functions have to be required or the effect of the last term in (VII.1) has to be of minor importance. (For quadratic benefit and damage cost functions the last term vanishes since $\partial^2 e_j/\partial e_i^2 = 0$.)

If $\partial^2 e_j/\partial e_i^2 < 0$ the damage cost function becomes concave instead of convex.[6] Then if $|\beta_i''| > |\phi_i''|$, where the last three terms in (VII.1) have been summarized to ϕ_i'', the payoff function will still be concave and we get an interior solution. Only if $|\beta_i''| < |\phi_i''|$ would the SOC be violated (that is, $\pi_i'' > 0$) which therefore has to be ruled out by assumption.

If the SOC hold, the leader's objective function is strictly concave and his/her equilibrium strategy is unique. Since the follower's profit function is also strictly concave as established in note 3 of Chapter 9, the uniqueness of the Stackelberg equilibrium follows. QED

VII.2 Appendix 2

Proposition 10.1 claims $e_i^{ST} \geq e_i^N$, $e_j^{ST} \leq e_j^N$, $\Sigma e_k^{ST} \geq \Sigma e_k^N$, $\pi_i^{ST} \geq \pi_i^N$ and $\pi_j^{ST} \leq \pi_j^N$ where country i is the leader and j the follower. The strict inequality sign cannot be used since a boundary Nash equilibrium has not been ruled out. If $e_j^N = 0$, then, obviously, there is no incentive for the Stackelberg leader i to expand emission beyond e_i^N and $e_i^N = e_i^{ST}$, $e_j^N = e_j^{ST}$ and $\Sigma e_k^{ST} = \Sigma e_k^N$ follows. For all other cases, however, the strict inequality sign holds.

To prove the above-stated relation, assume $e_i^{ST} < e_i^N$ (which requires $e_i^N > 0$)

instead, $e_i^{ST} > e_i^N$ would be true. Then from $-1 < \partial e_j/\partial e_i < 0$, $\Sigma e_k^{ST} < \Sigma e_k^N$ would follow. Consequently, $\beta_i'(e_i^N) < \beta_i'(e_i^{ST})$ and $\phi_i'(\Sigma e_k^N) > \phi_i'(\Sigma e_k^{ST})(1 + k_i)$. Hence in the Stackelberg equilibrium $\pi_i' > 0$ must hold which violates the FOC in (10.2). $e_i^{ST} > e_i^N$ implies $e_j^{ST} < e_j^N$ due to negatively sloped reaction functions. Since $-1 < \partial e_j/\partial e_i < 0$ country j reduces its emissions less than country i expands its emissions and hence $\Sigma e_k^{ST} \geq \Sigma e_k^N$ follows.

It should be obvious that the leader must gain (provided $e_j^N > 0$) otherwise he/she could choose $e_i^{ST} = e_i^N$. The follower's loss is obvious since due to $\Sigma e_k^{ST} \geq \Sigma e_k^N$ and $e_j^{ST} < e_j^N$ damage will be higher and benefits lower than in the Nash equilibrium $_{QED}$[7]

VII.3 Appendix 3

Proposition 10.2 claims that in a transferable externality game $\Sigma e_k^N < \Sigma e_k^S$ and $\Sigma e_k^N < \Sigma e_k^{ST}$.

Proof: Since transportation coefficients are not important for the subsequent analysis, we use the general formulation $\phi_i = \phi_i(e_i, e_{-i})$ to describe damages instead of $\phi_i = \phi_i(e_i + e_{-i})$ as previously. The payoff function thus reads $\pi_i = \beta_i(e_i) - \phi_i(e_i, e_{-i})$.

For the FOC in the Nash equilibrium itself the assumption of transferable externalities has no implication since $\partial\phi_j/\partial e_i$ does not play a role in country i's optimization task (see (VII.3) below). However, reaction functions are upward-sloping in emission space since:

$$r_i' = \frac{\partial\phi_i/\partial e_i \partial e_j}{\beta_i'' - \partial\phi_i^2/\partial e_i^2} \tag{VII.2}$$

holds and the nominator – different from previous sections – is now negative, whereas the denominator is still negative, so that $r_i' > 0$. From (VII.2) it is evident that a sufficient condition for $r_i' < 1$ is $|\partial\phi_i^2/\partial e_i^2| > |\partial\phi_i/\partial e_i \partial e_j|$ (as assumed in Proposition 10.2) and hence by Theorem 9.2 there is a unique Nash equilibrium.

For the social optimum the assumption of transferable externalities implies a major change. This is evident from (VII.3) where the FOC in the Nash equilibrium and in the social optimum are contrasted with each other (assuming an interior solution) and the signs of the derivatives have been reproduced for convenience:

$$\text{Nash eq.: } \underset{+}{\frac{\partial\beta_i}{\partial e_i}}\bigg|_{e_i^N} - \underset{+}{\frac{\partial\phi_i}{\partial e_i}}\bigg|_{\Sigma e_k^N} = 0; \quad \text{Social opt. } \underset{+}{\frac{\partial\beta_i}{\partial e_i}}\bigg|_{e_i^N} - \underset{+}{\frac{\partial\phi_i}{\partial e_i}}\bigg|_{\Sigma e_k^N} - \underset{+}{\frac{\partial\phi_i}{\partial e_i}}\bigg|_{\Sigma e_k^N} > 0. \tag{VII.3}$$

From (VII.3), together with the assumptions with respect to the second-order derivatives, it is evident that emissions in the Nash equilibrium are *too low* from a global point of view.[8,9] Hence, the relation derived in the previous sections for the case of filterable externalities is reversed and $\Sigma e_k^N < \Sigma e_k^S$ is proved.

In the Stackelberg equilibrium the leader's FOC are given by:[10]

$$\underset{(+)}{\frac{\partial \beta_i}{\partial e_i}} - \underset{(+)}{\frac{\partial \phi_i}{\partial e_i}} - \underset{(-)\cdot(+)}{\frac{\partial \phi_i}{\partial e_j}\frac{\partial e_j}{\partial e_i}} = 0. \qquad \text{(VII.4)}$$

Since the last term in (VII.4) is negative, marginal damages are less valued by the Stackelberg leader and s/he will choose higher emissions than in the Nash equilibrium. The follower, due to his/her *positively* sloped reaction function, will respond by increasing emissions. Hence, aggregate emissions in the Stackelberg equilibrium exceed those in the Nash equilibrium, that is, $\Sigma e_k^N < \Sigma e_k^{ST}$. QED

VII.4 Appendix 4

At the end of Section 10.4, in note 20 we mention that the FOC and the SOC may not be satisfied in a negative conjectural variation equilibrium.

Proof: The FOC for an interior conjectured equilibrium are:

$$\beta_i' - \phi_i'(1 + k_i) = 0. \qquad \text{(VII.5)}$$

If $k_i < -1$ and $\beta_i' \geq 0 \; \forall \; e_i \geq 0$, which is implied by benefit curves of the type depicted in Figure 9.1(a) and (b), (VII.5) cannot hold. Then $\pi_i' > 0 \; \forall \; e_i \geq 0$. $k_i < -1$ implies that benefits increase and damages decrease in *own* emissions and therefore in the conjectured equilibrium countries choose their maximum emission level e_i^{max} and $\pi_i'' > 0$ holds. Only if benefit curves exhibit a pattern as shown in Figure 9.1(c) may the FOC for an interior solution hold, because above e_i^0 $\beta_i' < 0$ is true. In the case where the damage cost function is convex as drawn in Figure VII.1(a) the interior equilibrium lies between e_i^0 and e_i^Z. If the damage cost function is convex, as suggested by Figure VII.1(b), an interior equilibrium can be expected if $|\beta_i''| > |\phi_i''|$. However, in the opposite case we have a boundary equilibrium at e_i^Z. Then $\pi_i' > 0$ and $\pi_i'' > 0$ in equilibrium. QED

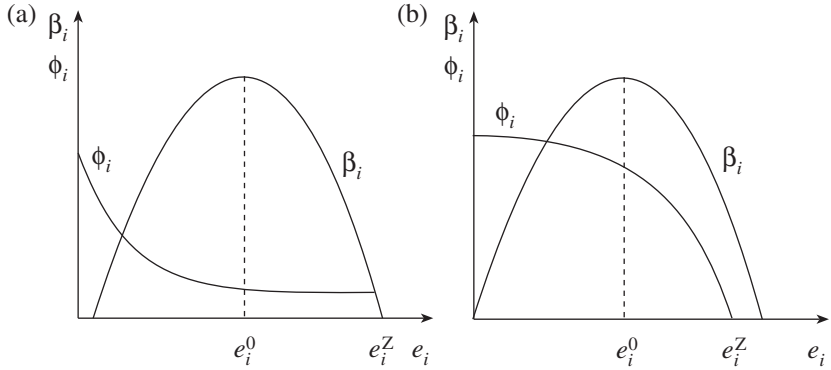


Figure VII.1 Boundary negative conjectural variation equilibria

VII.5 Appendix 5

Proposition 10.4 claims that $e_i^A < e_i^I$ and $\pi_i(e^A) > \pi_i(e^I)$, π_i^M $i \in I$ and hence $\Sigma e_i^A < \Sigma e_i^I$ and $\Sigma\pi_i(e^A) > \Sigma\pi_i(e^I)$ hold provided $e_i^I \geq e_i^N$ $\forall$ $i \in I$.

Proof: First, it is shown that provided $e_i^I \geq e_i^N$ each country makes a positive offer for any $\mu > 0$, that is, $r^i > 0$. Second, it is demonstrated that net benefit curves are strictly concave in r^i for any μ. Hence, in the equilibrium μ^*, where both countries make the same offer, that is, $r^i = \mu^* \cdot r^j$, both countries must have gained compared to the initial situation. Third, $\pi_i(e^A) > \pi_i^M$ is shown.

1. The FOC in the initial situation ($r^i = 0$) for any $e_i^I \geq e_i^N$ $\forall$ $i \in I$ are (recalling $e_i = (1 - r^i) \cdot e_i^I$):

$$\frac{\partial \beta_i}{\partial e_i}\frac{\partial e_i}{\partial r^i} - \frac{\partial \phi_i}{\partial e_i}\frac{\partial e_i}{\partial r^i} \leq 0 \Leftrightarrow -\frac{\partial \beta_i}{\partial e_i} \cdot e_i^I + \frac{\partial \phi_i}{\partial e_i} \cdot e_i^I \leq 0 \qquad \text{(VII.6)}$$

   where the equality sign holds in the (interior) Nash equilibrium and $<$ for any greater emission levels. The FOC of an offer (assuming an interior solution) may be written as (see also (10.10) in the text):[11]

$$\frac{\partial \beta_i}{\partial e_i}\frac{\partial e_i}{\partial r^i} - \frac{\partial \phi_i}{\partial \Sigma e_k}\frac{\partial \Sigma e_k}{\partial r^i} = 0 \Leftrightarrow -\frac{\partial \beta_i}{\partial e_i} \cdot e_i^I + \frac{\partial \phi_i}{\partial e_i} \cdot (e_i^I + \mu \cdot e_j^I) = 0. \qquad \text{(VII.7)}$$

   Since $|\partial \phi_i / \partial e_i(e_i^I)| < |\partial \phi_i(e_i^I + \mu e_j^I)/\partial e_i|$ $\forall$ r^i and $\mu > 0$ and recalling $\beta_i'' < 0$, the offer will always be positive, $r^i > 0$. Clearly, if initial emissions are below those in the Nash equilibrium the $\leq$ sign in (VII.6) becomes $\geq$ and $r^i > 0$ cannot be deduced any more by comparing (VII.6) with (VII.7). The optimal offer may involve $r^i \leq 0$.

2. A sufficient condition for strict concavity of the net benefit function with respect to the offer r^i is if the second-order derivative is negative over the entire domain of r^i. Hence:

$$\frac{\partial^2 \pi_i}{\partial r^{i^2}} = \frac{\partial^2 \beta_i}{\partial e_i^2}\left(\frac{\partial e_i}{\partial r^{i^2}}\right)^2 + \frac{\partial \beta_i}{\partial e_i}\frac{\partial^2 e_i}{\partial r^{i^2}} - \frac{\partial^2 \phi_i}{\partial (\Sigma e_k)^2}\left(\frac{\partial \Sigma e_k}{\partial r^i}\right)^2 - \frac{\partial \phi_i}{\partial \Sigma e_k}\frac{\partial^2 \Sigma e_k}{\partial r^{i^2}} < 0 \tag{VII.8}$$

must be true; which it is since:

$$\frac{\partial^2 \beta_i}{\partial e_i^2} < 0,\ \left(\frac{\partial e_i}{\partial r^i}\right)^2 > 0,\ \frac{\partial^2 \phi_i}{\partial (\Sigma e_k)^2} > 0,\ \left(\frac{\partial \Sigma e_k}{\partial r^i}\right)^2 > 0,\ \frac{\partial^2 e_i}{\partial r^{i^2}} = 0 \text{ and } \frac{\partial^2 \Sigma e_k}{\partial r^{i^2}} = 0. \tag{VII.9}$$

The last two equalities follow from $e_i = (1 - r^i) \cdot e_i^{\mathrm{I}}$ and $\Sigma e_k = (1 - r^i) \cdot e_i^{\mathrm{I}} + (1 - \mu \cdot r^i) \cdot e_i^{\mathrm{I}}$.

3. Suppose that $e_i^{\mathrm{I}} = e_i^{\max}$ and $e_j^{\mathrm{I}} = e_j^{\max}$. If $\mu^* = 0$, implying that country j would not contribute to abatement, country i would propose r_i such that $e_i^{\mathrm{A}} = e_i(e_j^{\max})$. For any $\mu^* > 0$, country j contributes something to abatement and hence in equilibrium where $r_i = \mu^* r_j$ country i's payoff must exceed its minimax payoff. By symmetry, the same holds for country j as long as $\mu^* < \infty$. QED

VIII CHAPTER 11: APPENDICES

VIII.1 Appendix 1

In Section 11.4 we claim that to ensure strict concavity of the payoff functions with respect to a uniform tax t_i, third-order effects must be sufficiently small (see assumption A_2 in (11.7)).

Proof: The first derivative of payoffs with respect to t_i is given by:

$$\frac{\partial \pi_i}{\partial t_i} = \frac{\partial \beta_i}{\partial e_i}\frac{\partial e_i}{\partial t_i} - \frac{\partial \phi_i}{\partial \Sigma e_k}\frac{\partial \Sigma e_k}{\partial t_i} \tag{VIII.1}$$

where $\partial \pi_i / \partial t_i = 0$ in the optimum. The sufficient condition for strict concavity is $\partial^2 \pi_i / \partial t_i^2 < 0\ \forall\ t_i \in [0, t_i^{\max}]$. Therefore, we require:

$$\frac{\partial^2 \pi_i}{\partial t_i^2} = \frac{\partial^2 \beta_i}{\partial e_i^2}\left(\frac{\partial e_i}{\partial t_i}\right)^2 + \frac{\partial \beta_i}{\partial e_i}\frac{\partial^2 e_i}{\partial t_i^2} - \frac{\partial^2 \phi \mathrm{D}_i}{\partial \Sigma e_k^2}\left(\frac{\partial \Sigma e_k}{\partial t_i}\right)^2 - \frac{\partial \mathrm{D}_i}{\partial \Sigma e_k}\frac{\partial^2 \Sigma e_k}{\partial t_i^2} < 0. \tag{VIII.2}$$

From the equilibrium condition (11.3) in the text, $t=\beta_i'(e_i)$ if $t\leq\beta_i'(0)$ we have:

$$\frac{\partial e_i}{\partial t_i}=\frac{1}{\beta_i''} \text{ and } \frac{\partial^2 e_i}{\partial t_i^2}=-\frac{1}{\beta_i''^3}\cdot\beta_i'''$$

by applying the implicit function rule. Using $\Sigma e_k=e_i+e_j$ and therefore $\partial\Sigma e_k/\partial e_i=1$, (VIII.2) can be written as:

$$(\phi_i''-\beta_i'')\left(\frac{1}{\beta_i''}\right)^2+(\phi_i'-\beta_i')\left(-\frac{1}{\beta_i''^3}\cdot\beta_i'''\right)>0. \qquad \text{(VIII.3)}$$

The first term is positive because $\phi_i''>0$ and $\beta_i''<0$ by assumption A_1. Since $\phi_i'-\beta_i'<0$ is possible if $\beta_i'(0)\geq t>t_i$, we have to require β_i''' to be sufficiently small in this case to ensure that the whole term is positive. (However, for $t>\beta_i'(0)$, we set $e_i=0$ and then $\partial e_i/\partial t_i=0$ and $\partial^2 e_i/\partial t_i^2=0$ in (VIII.3) and because $\phi_i'>0$ and $\phi_i''>0$ hold, concavity follows immediately.) QED

VIII.2 Appendix 2

In this appendix we provide equilibrium emissions under the quota and tax regimes, the Nash equilibrium, the social optimum and the minimax emission tuples for the payoff functions in (11.8). All relevant information is provided in Table VIII.1.

Equilibrium emissions in the Nash equilibrium, social optimum and when a country is minimaxed are derived as described in Chapter 11; however, they are now based on payoff functions (11.8).

From the table it is evident that the following non-negativity conditions (abbreviated NNC_i) have to be imposed:

$$\begin{aligned} A_3\text{: } NNC_1\text{: } \gamma>|\omega-1|(\Theta+1)/\omega \text{ and } NNC_2\text{: } \gamma>|\omega-\Theta|/\omega \\ NNC_3\text{: } \gamma>\max\{\Theta/\omega, 1\},\ e_i\in[0, d]. \end{aligned} \qquad \text{(VIII.4)}$$

That is, NNC_1 follows from requiring $e_i^S\geq 0$, NNC_2 from $e_i^N\geq 0$ and NNC_3 from $e_i^{M(i)}\geq 0\ \forall\ i\in I$ where the upper bound of the strategy space is, again, assumed to be e_i^0 which follows from $\beta_i'(e_i^0)=0$. Using NNC_1 and NNC_2 we can show that emissions under the tax and quota regimes are positive.

Moreover, note that the proposals under the quota regime are given by:

$$\begin{aligned} r_1&=\frac{(\gamma\omega+\Theta-\omega)(\gamma\omega+\Theta+\omega)}{\omega^2\gamma^2+2\omega\Theta\gamma+2\omega^2\gamma+\Theta^2-2\Theta\omega+\omega^2}, \\ r_2&=\frac{(\gamma\omega-\Theta+\omega)(\gamma\omega+\Theta+\omega)}{\omega^2\gamma^2+2\omega\Theta\gamma+2\omega^2\gamma+\Theta^2-2\Theta\omega+\omega^2} \end{aligned} \qquad \text{(VIII.5)}$$

from which it follows that $r_1 >< = r_2$ if $\Theta >< = \omega$. That is, $r_1 = r^A$ if $\Theta > \omega$, $r_2 = r^A$ if $\Theta < \omega$ and $r_1 = r_2 = r^A$ if $\Theta = \omega$. The equilibrium emissions follow from $e_i^Q = (1 - r^A) \cdot e_i^N$.

VIII.3 Appendix 3

Proposition 11.3 claims $\Sigma e_k^Q < \Sigma e_k^N$, $\Sigma e_k^T < \Sigma e_k^N$, $\Sigma e_k^Q \geq < \Sigma e_k^S$ and $\Sigma e_k^T \geq \Sigma e_k^S$.

Proof: To prove Proposition 11.3 we need the following lemma as preliminary information:

Lemma VIII.1
Let $k^* \in \{r^*, t^*\}$ denote a global welfare maximizing uniform policy level which solves $\max_{(k)}[\Sigma \pi_k(e_i(k), e_j(k))]$, $\pi_i(e_i(k), e_j(k))$ a strictly concave function in k, and let k_i and k_j be the bargaining proposals for countries i and j. Then either $k^* \in]k_i, k_j[$ or $k^* = k_i = k_j$ and therefore $k^A \leq k^*$.

Proof: The idea of the proof is to show that the maximum of the sum of two strictly concave functions is located in the interval of the maxima of the two single functions. Concavity of the payoff functions under both regimes has been established in Section 11.4. Hence, $\partial \pi_i / \partial k > 0 \ \forall \ k < k_i \wedge \partial \pi_i / \partial k \leq 0 \ \forall \ k \geq k_i$. We show $k_i > k^* > k_j$ or $k_i = k_j = k^*$ by contradiction, assuming, first, $k_i, k_j > k^*$ which implies:

$$\left.\frac{\partial \Sigma \pi_k}{\partial k}\right|_{k^*} = \underset{>0}{\left.\frac{\partial \pi_i}{\partial k}\right|_{k^*}} + \underset{>0}{\left.\frac{\partial \pi_j}{\partial k}\right|_{k^*}} > 0. \tag{VIII.6}$$

Second, we assume $k_i, k_j < k^*$ which implies:

$$\left.\frac{\partial \Sigma \pi_k}{\partial k}\right|_{k^*} = \underset{<0}{\left.\frac{\partial \pi_i}{\partial k}\right|_{k^*}} + \underset{<0}{\left.\frac{\partial \pi_j}{\partial k}\right|_{k^*}} < 0. \tag{VIII.7}$$

However, (VIII.6) and (VIII.7) should both be zero evaluated at k^* due to the FOC. Of course, $k^A \leq k^*$ is then an immediate implication of the LCD decision rule. QED

$\Sigma e_k^Q < \Sigma e_k^N$
Since $r^A > 0$ has been established, $e_i^Q = (1 - r^A) \cdot e_i^N < e_i^N \ \forall \ i \in I$ holds and $\Sigma e_k^Q < \Sigma e_k^N$ follows.

Table VIII.1 Emission levels under the quota and tax regimes for the payoff functions in (11.8)

	e_1	Quota regime	e_2
RE_1 $\omega \leq \Theta$	$\dfrac{d(\omega\gamma + \Theta - \omega)^2}{(\omega^2\gamma^2 + 2\omega\Theta\gamma + 2\omega^2\gamma + \Theta^2 - 2\omega\Theta + \omega^2)}$		$\dfrac{d(\omega\gamma + \Theta - \omega)(\omega\gamma - \Theta + \omega)}{(\omega^2\gamma^2 + 2\omega\Theta\gamma + 2\omega^2\gamma + \Theta^2 - 2\omega\Theta + \omega^2)}$
RE_2 $0 \leq \Theta \leq \omega$	$\dfrac{d(\omega\gamma + \Theta - \omega)(\omega\gamma - \Theta + \omega)}{(\omega^2\gamma^2 + 2\omega\Theta\gamma + 2\omega^2\gamma + \Theta^2 - 2\omega\Theta + \omega^2)}$		$\dfrac{d(\omega\gamma - \Theta + \omega)^2}{(\omega^2\gamma^2 + 2\omega\Theta\gamma + 2\omega^2\gamma + \Theta^2 - 2\omega\Theta + \omega^2)}$

	e_1	Tax regime	e_2
RE_1 $\dfrac{1}{\omega} \leq \Theta \leq \dfrac{1+\omega}{\omega}$	$\dfrac{d(\omega^2\gamma - \omega^2 + 1)}{\gamma\omega^2 + \omega^2 + 2\omega + 1}$		$\dfrac{d(\omega^2\gamma + \omega^2 - 1)}{\gamma\omega^2 + \omega^2 + 2\omega + 1}$
RE_1^a $\dfrac{1+\omega}{\omega} \leq \Theta$	$\dfrac{d(\omega^2\gamma - \omega^2 + 1)}{\gamma\omega^2 + \omega^2 + 2\omega + 1}$		$\dfrac{d(\omega^3\gamma^2 + \omega^3\gamma + 2\omega^2\gamma + \omega\gamma - \gamma\Theta\omega^2 + \Theta\omega^2 - \Theta)}{(\omega^2\gamma + \omega^2 + 2\omega + 1)(\omega\gamma + \Theta)}$
RE_2 $\dfrac{1+\omega}{\omega} \leq \Theta \leq \dfrac{1}{\omega}$	$\dfrac{d(\omega\gamma - \Theta\omega^2 + \Theta)}{\omega\gamma + \Theta\omega^2 + 2\omega\Theta + \Theta}$		$\dfrac{d(\omega\gamma + \Theta\omega^2 - \Theta)}{\omega\gamma + \Theta\omega^2 + 2\omega\Theta + \Theta}$

RE_2^a $0 \leq \Theta \leq \frac{1+\omega}{\omega}$	$\frac{d(\omega\gamma^2 + \gamma\Theta\omega^2 + 2\gamma\omega\Theta + \gamma\Theta - \omega\gamma - \Theta\omega^2 + \Theta)}{(\omega\gamma + \Theta\omega^2 + 2\omega\Theta + \Theta)(\gamma + 1)}$	$\frac{d(\omega\gamma + \Theta\omega^2 - \Theta)}{\omega\gamma + \Theta\omega^2 + 2\omega\Theta + \Theta}$

	e_1	Social optimum	e_2
	$\frac{d(\omega\gamma + 1 + \Theta - \omega - \omega\Theta)}{\omega\gamma + 1 + \Theta + \omega + \omega\Theta}$		$\frac{d(\omega\gamma - 1 - \Theta + \omega + \omega\Theta)}{\omega\gamma + 1 + \Theta + \omega + \omega\Theta}$

	e_1	Nash equilibrium	e_2
	$\frac{d(\omega\gamma + \Theta - \omega)}{\omega\gamma + \Theta + \omega}$		$\frac{d(\omega\gamma - \Theta + \omega)}{\omega\gamma + \Theta + \omega}$

	e_1	Minimax	e_2
1 is minimaxed	$\frac{d(\gamma - 1)}{\gamma + 1}$		d
2 is minimaxed	d		$\frac{d(\gamma\omega - \Theta)}{\gamma\omega + \Theta}$

Note: The symbols imply: $RE_j, j \in \{1, 2\}$, defines parameter regions (Θ and ω). The subscript *a* stands for adjustment. e_1 and e_2 are emissions of country 1 and country 2.

$\Sigma e_k^T < \Sigma e_k^N$

First, assume $t_i = t_j = t$, then $t = t^*$ (see Lemma VIII.1 above). Since a uniform tax rate is efficient (see Section 11.2), $\Sigma e_k^T = \Sigma e_k^S$. Since $\Sigma e_k^S < \Sigma e_k^N$ (see Proposition 9.1), $\Sigma e_k^T = \Sigma e_k^S < \Sigma e_k^N$ follows.

Second, assume $t_i < t_j$ and country i is the bottleneck. Further, note that $t_i > t_i^N \ \forall \ i \in I$ where t_i^N denotes a national tax which brings about e_i^N in country i and hence $e_i^T < e_i^N$. This is evident from comparing the FOC in the (interior) Nash equilibrium (NE) with those when deriving a tax proposal t_i:

$$\frac{\partial \pi_i}{\partial t_i^N} = \frac{\partial \beta_i}{\partial e_i}\frac{\partial e_i}{\partial t_i^N} - \frac{\partial \phi_i}{\partial e_i}\frac{\partial e_i}{\partial t_i^N} = 0; \quad \frac{\partial \pi_i}{\partial t_i} = \frac{\partial \beta_i}{\partial e_i}\frac{\partial e_i}{\partial t_i} - \frac{\partial \phi_i}{\partial e_i}\frac{\partial \Sigma e_k}{\partial t_i} = 0 \quad \text{(VIII.8)}$$

and noting

$$\left|\frac{\partial \phi_i}{\partial e_i}\frac{\partial e_i}{\partial t_i^N}\right| < \left|\frac{\partial \phi_i}{\partial e_i}\frac{\partial \Sigma e_k}{\partial t_i}\right| \ \forall \ t_i \text{ and } \beta_i'' < 0.$$

Then, two cases can be distinguished:

1. $t_i \geq t_j^N \Rightarrow e_j^T \leq e_j^N$. Since $t_i > t_i^N$ implies $e_i^T < e_i^N$, $\Sigma e_j^T < \Sigma e_j^N$ follows.
2. $t_i < t_j^N \wedge$ (a) $\beta_j'(e_j^T) \geq \phi_j'(e_i^T, e_j^T) \Rightarrow$ country j does not adjust.
   (b) $\beta_j'(e_j^T) < \phi_j'(e_i^T, e_j^T) \Rightarrow$ country j adjusts.

2(a) and 2(b) imply $e_j^T > e_j^N$ and hence $\beta_j'(e_j^T) < \beta_j'(e_j^N)$. From the assumption $\beta_j'(e_j^T) \geq \phi_j'(\Sigma e_k^T)$ in 2(a), $\phi_j'(e^T) \ll \phi_j'(\Sigma e_k^N)$ follows, which is only true if $\Sigma e_j^T < \Sigma e_j^N$. In 2(b) country j adjusts according to its reaction function. Since $e_i^T < e_i^N$ and the slopes of the reaction functions are less than 1 in absolute terms, we conclude $|e_i^N - e_i^T| > |e_j^N - e_j^T(e_i^T)|$ and therefore $e_i^T + e_j^T(e_i^T) < \Sigma e_i^N$ is proved.

$\Sigma e_k^T > \Sigma e_k^S$

From Lemma VIII.1 above we have $t^A \leq t^*$ and since $\partial \Sigma e_k / \partial t < 0$, $\Sigma e_k^T > \Sigma e_k^S$ follows.

$\Sigma e_k^Q \gtrless \Sigma e_k^S$

$\Sigma e_k^Q \geq \Sigma e_k^S$ can be shown for the functions in (11.8), using the information of Table VIII.1:

$$\Theta \leq \omega: \ \Sigma e_k^Q - \Sigma e_k^S =$$

$$\frac{2d\omega\gamma(\gamma(\omega^2\Theta + \omega - 2\omega\Theta) + \omega^2\Theta - \omega\Theta^2 + \omega - \Theta + 2\omega\Theta - 2\Theta^2)}{(\gamma\omega + 1 + \Theta + \omega + \omega\Theta)(\gamma^2\omega^2 + 2\gamma\omega\Theta + 2\gamma\omega^2 + \Theta^2 - 2\omega\Theta + \omega^2)} > 0 \quad \text{(VIII.9)}$$

$$\Theta \geq \omega: \Sigma e_k^{\mathrm{Q}} - \Sigma e_k^{\mathrm{S}} =$$
$$\frac{2d\omega\gamma(\gamma(\omega^2\Theta + \omega - 2\omega^2) - \omega^2\Theta + \omega\Theta^2 - \omega + \Theta + 2\omega\Theta - 2\omega^2)}{(\gamma\omega + 1 + \Theta + \omega + \omega\Theta)(\gamma^2\omega^2 + 2\gamma\omega\Theta + 2\gamma\omega^2 + \Theta^2 - 2\omega\Theta + \omega^2)} > 0 \tag{VIII.10}$$

where (VIII.9) and (VIII.10) are positive for $\Theta = \omega \neq 1$ and zero for $\Theta = \omega = 1$.

$\Sigma e_k^{\mathrm{Q}} < \Sigma e_k^{\mathrm{S}}$ can be shown by assuming slightly different net benefit functions than in (11.8). For instance, consider:

$$\pi_1 = b_1 e_1 - \frac{1}{2}e_1^2 - \frac{c}{2}(e_1 + e_2)^2,\ \pi_2 = \omega\left(b_2 e_2 - \frac{1}{2}e_2^2\right) - \frac{c}{2}\Theta(e_1 + e_2)^2 \tag{VIII.11}$$

and assume $b_1 = 5$, $b_2 = 10$, $c = 1$, $\omega = 1$ and $\Theta = 11/5$. Then $e_1^N = 10/7$, $e_2^N = 15/7$, $e_1^{\mathrm{S}} = 0$ and $e_2^{\mathrm{S}} = 50/21$. The proposals under the quota regime are given by $r_1 = 15/29 > 11/32 = r_2$ and hence $r_2 = r^{\mathrm{A}}$, which implies $\Sigma e_i^{\mathrm{Q}} = 75/32 < 50/21 = \Sigma e_i^{\mathrm{S}}$. QED

VIII.4 Appendix 4

Proposition 11.4 claims $\Sigma\pi_k^{\mathrm{Q}} > \Sigma\pi_k^N$, $\Sigma\pi_k^{\mathrm{T}} > \Sigma\pi_k^N$, $t_i \neq t_j \Rightarrow \Sigma\pi_k^{\mathrm{T}} < \Sigma\pi_k^{\mathrm{S}}$, $t_i = t_j \Rightarrow \Sigma\pi_k^{\mathrm{T}} = \Sigma\pi_k^{\mathrm{S}}$, $r_i \neq r_j \Rightarrow \Sigma\pi_k^{\mathrm{Q}} < \Sigma\pi_k^{\mathrm{S}}$ but from $r_i = r_j$ $\Sigma\pi_k^{\mathrm{Q}} = \Sigma\pi_k^{\mathrm{S}}$ does *not* follow.

Proof

$\Sigma\pi_k^{\mathrm{Q}} > \Sigma\pi_k^N$
Since $r^{\mathrm{A}} > 0$, $\pi_i^{\mathrm{Q}} > \pi_i^N\ \forall\ i$ follows by strictly concave payoff functions, which implies $\Sigma\pi_k^{\mathrm{Q}} > \Sigma\pi_k^N$. See also the proof of Proposition 10.4.

$\Sigma\pi_k^{\mathrm{T}} > \Sigma\pi_k^N$
We distinguish the same cases as in the proof in Appendix VIII.3 above (1, 2(a) and 2(b)). For case 1 and 2(a) (no adjustment) $\Sigma e_i^{\mathrm{S}} \leq \Sigma e_i^{\mathrm{T}} < \Sigma e_i^N$ has been shown, which automatically implies $\Sigma\pi_i^{\mathrm{T}} > \Sigma\pi_i^N$ since a tax is an efficient instrument, $\Sigma\pi_i^{\mathrm{T}}$ is strictly concave in the tax rate and $\partial\Sigma\pi_i^{\mathrm{T}}/\partial t^{\mathrm{A}} < 0\ \forall\ t^{\mathrm{A}} \leq t^*$. Only case 2(b) is less straightforward because (11.3) in the text does not hold after adjustment has been conducted. If $\Sigma e_i^{\mathrm{T}} < \Sigma e_i^N$ were to hold *before* adjustment takes place, then $\Sigma\pi_i^{\mathrm{T}} > \Sigma\pi_i^N$ would already be true without adjustment (see the arguments above). Additionally, since the non-bottleneck country j conducts the adjustment voluntarily, this country must gain from such an action and, by $\partial\pi_i/\partial e_j < 0$, country i as well.

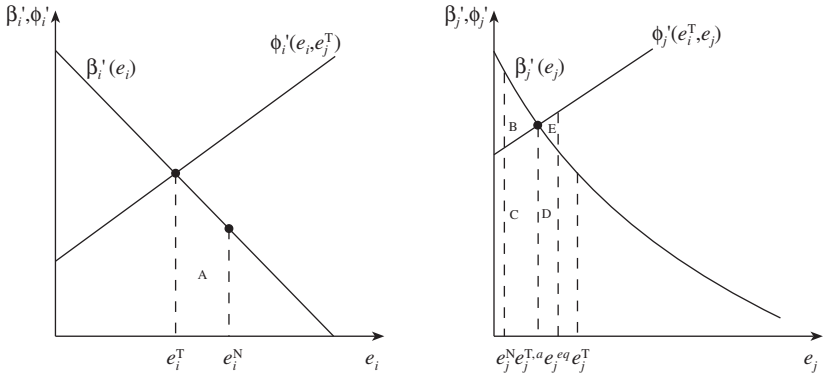


Figure VIII.1 Adjustment under the tax regime

To show $\Sigma\pi_i^T > \Sigma\pi_i^N$ if $\Sigma e_i^T > \Sigma e_i^N$ before adjustment holds (though $\Sigma e_i^T < \Sigma e_i^N$ holds after adjustment has taken place, which is known from Proposition 11.3) is more difficult. First, we establish the following relations $e_j^N < e_j^{T,a} < e_j^{eq} < e_j^T$, such that $\beta_j'(e_j^N) > \beta_j'(e_j^{T,a}) > \beta_j'(e_j^{eq}) > \beta_j'(e_j^T) = \beta_i'(e_i^T) > \beta_i'(e_i^N)$ is true (see Figure VIII.1). The superscript *eq* stands for 'equivalent' with $\Sigma e_i^N - e_i^T = e_j^{eq}$ and *a* stands for adjustment. Since $e_i^T < e_i^N$, e_j^{eq} gives emissions in country *j* which leave aggregate emissions unchanged compared to the NE. $e_j^{T,a} < e_j^{eq}$ is known because $e_i^T + e_j^{T,a} < \Sigma e_i^N$ by Proposition 11.3. $e_j^N < e_j^{T,a}$ follows from the fact that adjustment takes place and that the reaction functions are downward-sloping in emission space with slope less than 1 in absolute terms. Finally, $e_j^{eq} < e_j^T$ must hold, otherwise $\Sigma e_i^T > \Sigma e_i^N$ is not possible and the result for $\Sigma e_i^T < \Sigma e_i^N$ derived above would apply.

In a next step we compare the welfare changes of a move from e_i^N to e_i^T and from e_j^N to $e_j^{T,a}$. The move from e_j^N to $e_j^{T,a}$ is split into two parts: a move from e_j^N to e_j^{eq} and a move from e_j^{eq} to $e_j^{T,a}$. Then:

$$\underset{\text{A}}{\left|\int_{e_i^N}^{e_i^T} \beta_i'(e_i)\, de_i\right|} < \underset{\text{B+C+D}}{\left|\int_{e_j^N}^{e_j^{eq}} \beta_j'(e_j)\, de_j\right|} \text{ and } \underset{\text{D}}{\left|\int_{e_j^{eq}}^{e_j^{T,a}} \beta_j'(e_j)\, de_j\right|} < \underset{\text{D+E}}{\left|\int_{e_j^{eq}}^{e_j^{T,a}} \phi_j'(\Sigma e_k)\, de_j\right|} \tag{VIII.12}$$

is true, where the second relation simply follows from the fact that adjustment is conducted voluntarily and the first is a consequence of the fact that the marginal opportunity costs of abatement 'become more equal' so that cost efficiency potentials are realized.

$t_i \neq t_j \Rightarrow \Sigma\pi_k^T < \Sigma\pi_k^S$; $t_i = t_j \Rightarrow \Sigma\pi_k^T = \Sigma\pi_k^S$
From Lemma VIII.1 $t_i = t_j = t^*$ where t^* is a socially optimal tax and hence $t_i = t_j \Rightarrow \Sigma\pi_k^T = \Sigma\pi_k^S$ is true. Since $t_i \neq t_j$ implies $t^A < t^*$, $\Sigma e_k^T < \Sigma e_k^S$ and $\Sigma\pi_k^T < \Sigma\pi_k^S$ must be true.

$r_i \neq r_j \Rightarrow \Sigma\pi_k^Q < \Sigma\pi_k^S$; $r_i = r_j$ does not imply $\Sigma\pi_k^Q = \Sigma\pi_k^S$
$r_i \neq r_j$ implies $r^A \neq r^*$ and a uniform emission reduction quota can be socially optimal if and only if $r^S = r^* = r^A$. That $r_i = r_j$ does not imply $\Sigma\pi_k^Q = \Sigma\pi_k^S$ is simply demonstrated by example. For instance, choose for the payoff function (11.8), $\Theta = \omega \neq 1$, then $r_1 = r_2 = r^*$. (VIII.9) and (VIII.10) above reveal that $\Sigma e_k^Q > \Sigma e_k^S$ and therefore $\Sigma\pi_k^Q < \Sigma\pi_k^S$ follows. QED

VIII.5 Appendix 5

In Proposition 11.5, assuming country *i* to be the bottleneck in the negotiations, we claim $\pi_i^Q > \pi_i^N \; \forall \; i \in I$, $\pi_j^T > \pi_j^N$, $\pi_i^T > \leq \pi_i^N$, $\pi_i^T \geq \pi_i^M$ and $\pi_i^T < \pi_i^N$ if $I_j(t^A) < 0$.

$\pi_i^Q > \pi_i^N \; \forall \; i \in I$
See the proof of $\Sigma\pi_k^Q > \Sigma\pi_k^N$ above in Appendix VIII.4.

$\pi_j^T > \pi_j^N$
Consider the cases 1, 2(a) and 2(b) in Appendix VIII.3. Case 2(b) implies adjustment by country *j*. Since we show below that $I_j(t^A) < 0$ implies $\pi_i^T < \pi_i^N$ and by Proposition 11.4. $\Sigma\pi_k^T > \Sigma\pi_k^N$ holds, $\pi_j^T > \pi_j^N$ must be true. In case 2(a), where there is no adjustment, $t_i < t_j^N$ implies $e_j^T > e_j^N$. Moreover $t_i > t_i^N$ is known from Appendix VIII.3 which implies $e_i^T < e_i^N$. Since country *j* always has the option to choose $e_j^{Ta} < e_i^T$ and since $\pi_j(e_i^T, e_j^N) > \pi_j(e_i^N, e_j^N)$ holds due to $\partial\pi_j/\partial e_i < 0$, $\pi_j(e_i^T, e_j^T) > \pi_j(e_i^N, e_j^N)$ must be true. Case 1 implies an even higher tax $t_i \geq t_j^N$ and since $t_j \geq t_i$, t_j^N is true by assumption that country *i* is the bottleneck, $\pi_j^T > \pi_j^N$ must hold by the strict concavity of payoff functions with respect to the tax rate and $\pi_j^T > \pi_j^N$ for $t_i < t_j^N$ as demonstrated above.

$\pi_i^T > \leq \pi_i^N$
This statement is simply proven by example. For instance, choose $\Theta = 1/2$, $\omega = 1/4$, $d = 100$, $b = 10$, $c = 1$ for the functions in (11.8), then $\pi_1^T > \pi_1^N$ and $\pi_2^T < \pi_2^N$. For $\Theta = 1$, $\omega = 1$ and *d*, *b*, *c* as assumed above, $\pi_i^T > \pi_i^N \; \forall \; i \in I$.

$\pi_i^T \geq \pi_i^M$
Note that π_i^M is derived from max $\pi_i(e_i, e_j^{max})$ and that $t = 0$ implies $e_j^T = e_j^{max}$. Any proposal t_i by country i, however, is derived with the information $e_j^T < e_j^{max}$ for any $t_i > 0$ and hence $\pi_i^T \geq \pi_i^M$ must hold by $\partial \pi_i / \partial e_j < 0$.

$\pi_i^T < \pi_i^N$ if $I_j(t^A) < 0$
First, note that $e_i^N > e_i^T$ and $e_j^{T,a} := e_j(e_i^T) > e_i^N$ despite adjustment by country j. $e_j^{T,a} > e_i^N$ is shown by contradiction: suppose $e_j^{T,a} < e_i^N$ then $\phi_j'(e_i^T, e_j^{T,a}) < \phi_j'(e_i^N, e_j^N)$, $\beta_j'(e_j^{T,a}) > \beta_j'(e_j^N)$, which cannot be true by $I_i(e_i^T, e_j^{T,a}) = 0$ and $I_i(e_i^N, e_j^N) = 0$. Next, we show that any emission tuple e^T for which $e_j^T > e_i^N$ ($e_j^{T,a} > e_i^N$) and $e_i^T < e_i^N$ is true implies $\pi_i(e^T) < \pi_i(e^N)$. For this, note the following relations:

1. Assumption A_1 implies concave payoff indifference curves in the $e_i \times e_j$ space which do not intersect (see Chapter 9).
2. Any emission tuple in a north-west direction reaction of e^N (for example e^T) must imply a welfare loss to country i compared to the NE, which is evident from Figure VIII.2. QED

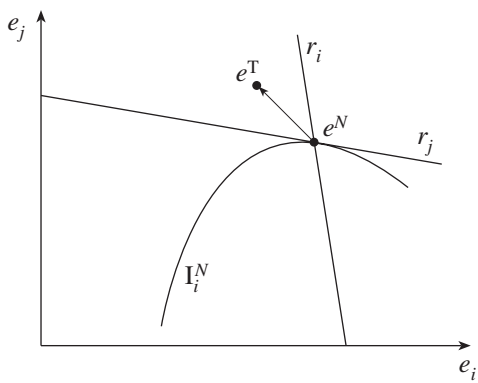


Figure VIII.2 Welfare loss to the bottleneck country if adjustment occurs

VIII.6 Appendix 6

Write down the FOC from which a biased proposal is derived to note that $t_i^{str} < t_i$ and that the implied emissions are exactly those in a Stackelberg equilibrium. Then most claims of Proposition 11.7 follow immediately from Proposition 10.2. Only (1) $\Sigma \pi_k^T(t_i^{str}) < \Sigma \pi_k^T(t_i)$; (2) $\pi_j^T(t^{str}) \geq \pi_j^M$ and (3) $\pi_j^T(t^{str}) < \pi_j^T(t_i)$ have to be shown additionally.

(1): First, note that $\Sigma e_k^{\mathrm{T}}(t_i^{str}) > \Sigma e_k^{N} > \Sigma e_k^{\mathrm{T}}(t_i)$ from Proposition 10.2 and Proposition 11.3. Second, note that after a non-biased proposal by country i either (a) no adjustment takes place by country j or (b) adjustment takes place by country j.[12] Possibility (a) implies $\beta_i'(e_i^{\mathrm{T}}) = \beta_j'(e_j^{\mathrm{T}})$ and a biased proposal will lead to $\beta_i'(e_i^{\mathrm{T}}(t_i^{str})) < \beta_j'(e_j^{\mathrm{T}}(t_i^{str}))$. Possibility (b) implies $\beta_i'(e_i^{\mathrm{T}}) < \beta_j'(e_j^{\mathrm{T},a})$ and a biased proposal will lead to $\beta_i'(e_i^{\mathrm{T}}(t_i^{str})) \ll \beta_j'(e_j^{\mathrm{T}}(t_i^{str}))$. That is, the biased proposal creates (increases) a (the) gap between marginal opportunity costs of abatement in the two countries and therefore leads to a (more) inefficient allocation of emission reduction burdens (see CEA condition (11.2)). Together with higher aggregate emissions resulting from a biased proposal, implying higher damages, global welfare must have decreased through the strategic proposal. (The case of $t_i = t^{\mathrm{A}}$ implying $e_i^{\mathrm{T}} > 0$ and $e_j^{\mathrm{T}} = 0$ or $e_i^{\mathrm{T}} = 0$ and $e_j^{\mathrm{T}} = 0$ (corner solutions) can be discarded because they obviously provide no incentive for the strategic mover i to increase emissions since $e_j^{\mathrm{T}}(t_i^{str}) < 0$ is not possible.)

(2): π_j^{M} follows from $\max \pi_i(e_i^{\max}, e_j)$ whereas $\pi_j^{\mathrm{T}}(e(t^{str})) = \pi_j^{\mathrm{ST}}$ follows from $\max \pi_i(e_i^{\mathrm{ST}}, e_j)$ where $e_i^{\mathrm{ST}} \leq e_i^{\max}$ by the definition of the strategy space and hence $\pi_j^{\mathrm{ST}} \geq \pi_j^{\mathrm{M}}$ by $\partial \pi_j / \partial e_i < 0$.

(3): From above $e_i^{\mathrm{T},str} > e_i^{\mathrm{T}}$, $\Sigma e_k^{str} > \Sigma e_k^{\mathrm{T}}$, $\Sigma e_k^{\mathrm{T},a}$ and $e_j^{\mathrm{T},str} < e_j^{\mathrm{T}}, e_j^{\mathrm{T},a}$ is true. Consequently, $\beta_j(e_j^{str}) < \beta_j(e_j^{\mathrm{T}})$, $\beta_j(e_j^{\mathrm{T},a})$ and $\phi_j(\Sigma e_k^{str}) > \phi_j(\Sigma e_k^{\mathrm{T}})$, $\phi_j(\Sigma e_k^{\mathrm{T},a})$ and the claim is proved. QED

IX CHAPTER 12: APPENDIX

We claim in Sub-section 12.3.3 that for $t_i^{\mathrm{P}} = \infty$ a necessary requirement to find an emission tuple during the punishment phase satisfying equation (12.19) is that payoffs in the cooperative phase are restricted to being smaller than the Stackelberg leader payoff, that is, $\mathrm{C}_i^{\mathrm{C}} := \pi_i^* < \pi_i^{\mathrm{ST}}$.

Proof: Assume country 2 punishes country 1. Then (12.19), using (11.8), may be written as:

$$\pi_2^* \leq b\omega\left(de_2^1 - \frac{1}{2}e_2^{1^2}\right) - \frac{c}{2}\Theta(e_1^1 + e_2^1)^2. \tag{IX.1}$$

Solving (IX.1), using country 1's best reply, $e_1^1 = (bd - ce_2^1)/(b + c)$, gives because of (12.27) in the text:

$$e_2^1 \in [e_{2,\min}^1, e_{2,\max}^1] \tag{IX.2}$$

where:

$$e^1_{2,\min} = \frac{d(b^2\omega^2 + 2bc\Theta\omega + c^2\Theta^2 - bc\omega^2) - (b\omega + c\Theta)\cdot \mathrm{A}}{b^2\omega^2 + 2bc\Theta\omega + c^2\Theta^2 + bc\omega^2}$$

$$e^1_{2,\max} = \frac{d(b^2\omega^2 + 2bc\Theta\omega + c^2\Theta^2 - bc\omega^2) + (b\omega + c\Theta)\cdot \mathrm{A}}{b^2\omega^2 + 2bc\Theta\omega + c^2\Theta^2 + bc\omega^2}$$

$$\mathrm{A} := \sqrt{b(b^3d^2\omega^2 - 2\omega b^2\pi^*_2 + 2cb^2d^2\omega^2 - 3cb^2d^2\omega\Theta - 2c\Theta b\pi^*_2 - 4c\omega b\pi^*_2 + c^2bd^2\omega^2 - 2c^2\omega\pi^*_2)}.$$

In order to obtain a solution for e^1_2 A must be positive, which is true if:

$$\pi^*_2 \leq \frac{\omega bd^2(\omega b^2 + 2\omega cb - 3\Theta cb + \omega c^2)}{2(\omega b^2 + \Theta cb + 2\omega cb + \omega c^2)}. \tag{IX.3}$$

Computation of the Stackelberg payoff according to the procedure described in Section 10.2 shows that this is equal to the RHS expression in (IX.3).

By the same token, it can also be shown that country 1 cannot receive more than its Stackelberg payoff to meet a precondition for an emission tuple to be renegotiation-proof if $t^{\mathrm{P}}_1 = \infty$. QED

X CHAPTER 13: APPENDICES

X.1 Appendix 1

The aim of this appendix is to provide all background information which supports the assertions in Sub-section 13.2.3. The proofs concentrate on interior solutions exclusively. That is, $e^{\mathrm{J}*}_i = \tilde{e}^{\mathrm{J}}_i \geq 0$ and $e^{NJ*}_j = \tilde{e}^{NJ}_j \geq 0$.

(1) Determination of n^* and ξ^*

Using (13.18) and (13.19), then ($\xi = n/N$):

$$\pi^{\mathrm{J}}_i(\xi) - \pi^{NJ}_j(\xi) = \frac{d^2bc^2(c\xi + b\xi N - b - c)(c\xi - b\xi N - b - c)}{2(b^2 + 2bc - 2bc\xi + c^2 2c^2\xi + c^2\xi^2 + b\xi^2 Nc)^2} \tag{X.1}$$

which implies that there are two zero points:

$$\xi_1 = -\frac{\gamma + 1}{N\gamma - 1};\ \xi_2 = \frac{\gamma + 1}{N\gamma + 1}. \tag{X.2}$$

However, $\xi_1 \notin [0,1]$ since for $\gamma N - 1 > 0 \Rightarrow \xi_1 < 0$ and for $\gamma N - 1 < 0$ $\xi_1 = -(\gamma + 1)/(1 - N\gamma) > \gamma + 1 > 1$. Thus:

$$\pi_i^{\mathrm{J}}(\xi)-\pi_j^{NJ}(\xi)>0 \text{ if } 0\leq\xi<\frac{\gamma+1}{N\gamma+1} \text{ and } \pi_i^{\mathrm{J}}(\xi)-\pi_j^{NJ}(\xi)\leq 0 \text{ if } \frac{\gamma+1}{N\gamma+1}\leq\xi\leq 1 \tag{X.3}$$

(regardless of whether $\gamma N-1>0$ or $\gamma N-1\leq 0$) and hence:

$$\xi^*=\frac{\gamma+1}{N\gamma+1},\ n^*=\frac{N(\gamma+1)}{N\gamma+1} \tag{X.4}$$

for which it is easily checked that $\partial\xi^*/\partial N<0$ and $/\partial\xi^*/\partial\gamma<0$ hold as claimed in (13.20).

(2) $n^*<n^{*\mathrm{L}}<N^*\leq n^{*\mathrm{U}}$, $N^*\in[2,N]$

Note the following relations:

$$\tilde{e}_i^{\mathrm{J}}(\xi)-e_i^N=-\frac{d(-\xi+1+\gamma)(-\xi-1+\gamma\xi+\gamma\xi N)}{(\gamma+1)(\gamma^2+2\gamma-2\gamma\xi+1-2\xi+\xi^2+\gamma\xi^2N)^2}<(>)0$$

$$\tilde{e}_i^{NJ}(\xi)-e_j^N=-\frac{d\xi(-\xi-1+\gamma\xi+\gamma\xi N)}{(\gamma^2+2\gamma-2\gamma\xi+1-2\xi+\xi^2+\gamma\xi^2N)^2}>(<)0 \text{ if } \xi>(<)\xi^*$$

$$\begin{aligned}&\Sigma\tilde{e}_k^{NJ}(\xi)+\Sigma\tilde{e}_k^{\mathrm{J}}(\xi)-\Sigma e_k^N\\&=-\frac{\gamma dN\xi(-\xi-1+\gamma\xi+\gamma\xi N)}{(\gamma+1)(\gamma^2+2\gamma-2\gamma\xi+1-2\xi+\xi^2+\gamma\xi^2N)^2}<(>)0\end{aligned} \tag{X.5}$$

from which it is evident that $\pi_j^{NJ}\leq\pi_j^N$ for $\xi\leq\xi^*$ (or $n\leq n^*$) must hold. Together with $\pi_i^{\mathrm{J}}>\pi_i^N\ \forall\ \xi\neq\xi^*$ (or $n\neq n^*$) as shown below in (3), it follows that a non-signatory has an incentive to join the coalition as long as $n\leq n^*$. Moreover, C_3 is a monotone function for which $\mathrm{C}_3(n^*)>0$ and $\partial\mathrm{C}_3/\partial n|_{n^*}<0$ is true and thus, $n^*<n^{*\mathrm{L}}$ is established. Since $n^{*\mathrm{L}}+1=n^{*\mathrm{U}}$, $n^{*\mathrm{L}}<n^{*\mathrm{U}}$ is evident. $N^*\in[n^{*\mathrm{L}},n^{*\mathrm{U}}]$ is evident from Definition 13.2 of internal and external stability in the text. $n^{*\mathrm{L}}<N^*$ follows from two pieces of information: (a) if $n^{*\mathrm{L}}$ is not an integer value, then of course $N^*=[n^{*\mathrm{L}}]+1$; (b) we show below that $\pi_i^{\mathrm{J}}(n^{*\mathrm{U}})>\pi_i^{\mathrm{J}}(n^{*\mathrm{L}})$ is true (see (3) below) and hence if $n^{*\mathrm{L}}$ is an integer value $N^*=n^{*\mathrm{U}}=[n^{*\mathrm{L}}]+1=n^{*\mathrm{L}}+1$.

$N^*\in[2,N]$ follows from $\lim_{\gamma\to 0} n^*=1$, $\lim_{\gamma\to\infty} n^*=N$ and $n^*<N^*$.

(3) $\Sigma e_k^N>\Sigma e_k^*\geq\Sigma e_k^{\mathrm{S}}$, $e_i^N>e_i^{\mathrm{J}*}$, $e_j^N<e_j^{NJ*}$, $\pi_i^N<\pi_i^{\mathrm{J}*}<\pi_j^{NJ*}$, $\Sigma\pi_k^N<\Sigma\pi_k^*\leq\Sigma\pi_k^{\mathrm{S}}$

The first three relations are an immediate implication of the results in (X.5) and $n^*<N^*$ ($\xi^*<\mu^*$). $\pi_i^{\mathrm{J}*}>\pi_i^N$ follows by noting that $\pi_i^{\mathrm{J}}(n)$ is a convex function where the minimum is located at n^* (since $\partial\pi_i^{\mathrm{J}}(n)/\partial n|_{n^*}=0$, $\partial^2\pi_i^{\mathrm{J}}(n)/\partial n^2>0\ \forall\ n\in[0,N]$; see Figure 13.1(a), again, $n^*\leq N^*$ and $\pi_i^{\mathrm{J}}=\pi_i^N$ at n^*. (Hence the claims in (2) above that $\pi_i^{\mathrm{J}}(n^{*\mathrm{U}})>\pi_i^{\mathrm{J}}(n^{*\mathrm{L}})$ and therefore N^*

$= n^{*U}$ if n^{*L} is an integer value and $\pi_i^J > \pi_i^N \; \forall \; \xi \neq \xi^*$ (or $n \neq n^*$) is proved.) Since $e_i^{J*} < e_j^N < e_j^{NJ*}$ and signatories and non-signatories suffer from the *same* global emissions, $\pi_i^{J*} < \pi_j^{NJ*}$ follows from $\partial \beta_j / \partial e_j > 0$ (as long as $e_j < e_j^0 = d$, which is easily confirmed to hold true). Of course, $\pi_i^N < \pi_i^{J*} < \pi_j^{NJ*}$ implies $\Sigma \pi_k^N < \Sigma \pi_k^*$ and $\Sigma \pi_k^* \leq \Sigma \pi_k^S$ follows from the fact that only for $\xi = 1$ $\tilde{e}_i^J(\xi) = e_i^S$ and $\tilde{e}_j^{NJ}(\xi) = e_j^S$ (and that the socially optimal emission allocation is unique for payoff functions of type (3)).

(4) $\dfrac{\partial \mu^*}{\partial \gamma} < 0, \dfrac{\partial I_3}{\partial \gamma} < 0, \dfrac{\partial I_4}{\partial \gamma} < 0$ **for** $N \to \infty$

Compute the difference functions $C_2(n)$ and $C_3(n)$ and let $N \to \infty$. Then it is found that:

$$n_{N\to\infty}^{*L} = \frac{\gamma^2 + \sqrt{\gamma^2(\gamma^2+1)}}{\gamma^2}, \; n_{N\to\infty}^{*U} = \frac{2\gamma^2 + \sqrt{\gamma^2(\gamma^2+1)}}{\gamma^2} = n_{N\to\infty}^{*L} + 1 \quad \text{(X.6)}$$

where $C_2 \geq (<) \; 0$ if $n \leq (>) \; n_{N\to\infty}^{*U}$ $C_3 \leq (>) \; 0$ if $n \geq (<) \; n_{N\to\infty}^{*L}$ and where $n^{*L} \leq N^* \leq n^{*U}$ must hold for the stability of a coalition. It is easily checked that $\partial n_{N\to\infty}^{*L} / \partial \gamma < 0$ and $\partial n_{N\to\infty}^{*U} / \partial \gamma < 0$ hold and therefore $\partial N^* / \partial \gamma < 0$ is true. Since by assumption of $N \to \infty$ N is fixed, $\partial \mu^* / \partial \gamma < 0$ is evident. Since $\lim_{\mu^* \to 0} \big|_{N\to\infty} \Sigma \tilde{e}_k^{NJ}(\xi) + \Sigma \tilde{e}_k^J(\xi) = \Sigma e_k^N$, $\partial I_3 / \partial \gamma = \partial I_1 / \partial \gamma$ and $\partial I_4 / \partial \gamma = \partial I_2 / \partial \gamma$ follow where the signs of $\partial I_1 / \partial \gamma$ and $\partial I_2 / \partial \gamma$ have been established in (13.4) and (13.5). QED

X.2 Appendix 2

In this appendix we prove that the social optimum associated with the transfer scheme described in (13.27) lies in the γ core provided one of the three properties mentioned in Proposition 13.5 holds (see Chander and Tulkens 1995, 1997).[13]

Proof: As preliminary information we need the following lemma (Chander and Tulkens 1997, Proposition 4):

Lemma X.1
The partial-agreement equilibrium (PANE) as described by Definition 14.6 has the following properties: (1) there exists a PANE; (2) the equilibrium emission vector is unique; (3) $\Sigma \tilde{e}_k = \Sigma \tilde{e}_i^J + \Sigma \tilde{e}_j^{NJ} \leq \Sigma e_k^N$; and (4) $e_j^N \leq \tilde{e}_j^{NJ} \; \forall \; j \notin I^J$ in the PANE.

Proof: Properties (1) and (2) are proved by applying the theorems as laid out in Section 9.4. Property (3) is proved by contradiction. Suppose $\Sigma \tilde{e}_k > \Sigma e_k^N$ were true. This implies (a) $\Sigma \phi_i'(\tilde{e}) = \beta_i'(\tilde{e}_i^J) > \beta_i'(e_i^N) = \phi_i'(e^N) \forall$

$i \in \mathrm{I^J}$ which is only possible if $\tilde{e}_i^{\mathrm{J}} < e_i^N \; \forall \; i \in \mathrm{I^J}$ is true; and (b) $\phi_j'(\tilde{e}) = \beta_j'(\tilde{e}_j^{NJ}) > \beta_j'(e_j^N) = \phi_j'(e^N) \; \forall \; j \notin \mathrm{I^J}$ which is only possible if $\tilde{e}_j^{NJ} < e_j^N \; \forall \; j \notin \mathrm{I^J}$ is true. $\tilde{e}_i^{\mathrm{J}} < e_i^N \; \forall \; i \in \mathrm{I^J}$ and $\tilde{e}_j^{NJ} < e_j^N \; \forall \; j \notin \mathrm{I^J}$ obviously imply $\Sigma\tilde{e}_k < \Sigma e_k^N$ which contradicts the initial assumption. Property (4) is proved by noting that $\Sigma\tilde{e}_k < \Sigma e_k^N$ implies $\phi_j'(\tilde{e}) = \beta_j'(\tilde{e}_j^{NJ}) < \beta_j'(e_j^N) = \phi_j'(e^N) \; \forall \; j \notin \mathrm{I^J}$ which is only true if $e_j^N \leq \tilde{e}_j^{NJ}$ by $\beta_i'' < 0$. QED

Case 1: Linear damage cost functions

(a) $\pi^{*\psi} = (\pi_1^{*\psi}, \ldots, \pi_N^{*\psi})$ is an imputation since:

$$\Sigma\pi_i^{*\psi} = \Sigma\pi_i^{\mathrm{S}}(e^{\mathrm{S}}) + \Sigma t_i^* = w(\mathrm{I}) \tag{X.7}$$

and $\Sigma t_i^* = 0$.

(b) Suppose $\pi^{*\psi}$ is not in the core. Then, there would exist a coalition $\mathrm{I^J} \subset \mathrm{I}$ and a PANE $\tilde{e} = (\tilde{e}^{\mathrm{J}}, \tilde{e}^{NJ})$ such that:

$$w^{\gamma}(\mathrm{I^J}) = \sum_{i \in \mathrm{I^J}} \pi_i^{\psi}(\tilde{e}) > \sum_{i \in \mathrm{I^J}} \pi_i^{*\psi}. \tag{X.8}$$

Note first that for linear damage cost functions non-signatories have a dominant strategy and hence $e_J^N = \tilde{e}_j^{NJ} \; j \notin \mathrm{I^J}$. Moreover, $e_i^{\mathrm{S}} \leq \tilde{e}_i^{\mathrm{J}}$ and $e_i^N \geq \tilde{e}_i^{\mathrm{J}}$ $i \in \mathrm{I^J}$ hold.[14]

Consider now an alternative imputation $\hat{\pi}^{\psi}$ where $\hat{\pi}_i^{\psi} = \pi_i(e^{\mathrm{S}}) + \hat{t}_i$ and where transfers are given by:

$$\hat{t}_i = -[\beta_i(e_i^{\mathrm{S}}) - \beta_i(\tilde{e}_i^{\mathrm{J}})] + \frac{\phi_i'(e^{\mathrm{S}})}{\Sigma_{i \in \mathrm{I}}\phi_i'(e^{\mathrm{S}})} \cdot [\Sigma_{i \in \mathrm{I}}\beta_i(e_i^{\mathrm{S}}) - \Sigma_{i \in \mathrm{I}}\beta_i(\tilde{e}_i^{\mathrm{J}})]. \tag{X.9}$$

Now it has been shown that:

$$\Sigma_{i \in \mathrm{I^J}}\hat{\pi}_i^{\psi} > \Sigma_{i \in \mathrm{I^J}}\pi_i(\tilde{e}) \tag{X.10}$$

and

$$\Sigma_{j \notin \mathrm{I^J}}\hat{\pi}_j^{\psi} \geq \Sigma_{j \notin \mathrm{I^J}}\pi_j^{*\psi} \tag{X.11}$$

hold so that the contradiction:

$$\Sigma_{k \in \mathrm{I}}\hat{\pi}_k^{\psi} > \Sigma_{i \in \mathrm{I^J}}\pi_i(\tilde{e}) + \Sigma_{j \notin \mathrm{I^J}}\pi_j^{*\psi} = \Sigma_{i \in \mathrm{I^J}}\pi_i^{*\psi} + \Sigma_{i \notin \mathrm{I^J}}\pi_j^{*\psi} = \Sigma_{k \in \mathrm{I}}\pi_k^{*\psi} \tag{X.12}$$

follows (since $\max \Sigma_{k \in \mathrm{I}}\pi_k = \Sigma_{k \in \mathrm{I}}\pi_k^{*\psi} \geq w(\mathrm{I})$ by definition) where the equality sign in (X.12) is implied by (X.8).

(X.10) is shown by noting that:

$$\Sigma_{i\in \mathrm{I^J}}\hat{\pi}_i^{\psi}=\Sigma_{i\in \mathrm{I^J}}\pi_i^{\mathrm{S}}(e^{\mathrm{S}})+\Sigma_{i\in \mathrm{I^J}}\hat{t}_i \tag{X.13}$$

implies:

$$\Sigma_{i\in \mathrm{I^J}}\hat{\pi}_i^{\psi}=\Sigma_{i\in \mathrm{I^J}}\beta_i(\tilde{e}_i^{\mathrm{J}})-\Sigma_{i\in \mathrm{I^J}}\phi_i(e^{\mathrm{S}})+\frac{\Sigma_{i\in \mathrm{I^J}}\phi_i'(e^{\mathrm{S}})}{\Sigma_{i\in \mathrm{I}}\phi_i'(e^{\mathrm{S}})}[\Sigma_{i\in \mathrm{I}}\beta_i(e_i^{\mathrm{S}})-\Sigma_{i\in \mathrm{I}}\beta_i(\tilde{e}_i^{\mathrm{J}})] \tag{X.14}$$

or

$$\Sigma_{i\in \mathrm{I^J}}\hat{\pi}_i^{\psi}=\Sigma_{i\in \mathrm{I^J}}\beta_i(\tilde{e}_i^{\mathrm{J}})-\Sigma_{i\in \mathrm{I^J}}\phi_i(\tilde{e})+\frac{\Sigma_{i\in \mathrm{I^J}}\phi_i'(e^{\mathrm{S}})}{\Sigma_{i\in \mathrm{I}}\phi_i'(e^{\mathrm{S}})}\cdot \mathrm{A};$$

$$\mathrm{A}:=[\Sigma_{i\in \mathrm{I}}\phi_i'(e^{\mathrm{S}})\cdot(\Sigma\tilde{e}_i-\Sigma e_i^{\mathrm{S}})+(\Sigma_{i\in \mathrm{I}}\beta_i(e_i^{\mathrm{S}})-\Sigma_{i\in \mathrm{I}}\beta_i(\tilde{e}_i^{\mathrm{J}}))] \tag{X.15}$$

where (X.15) has been obtained by adding to and subtracting $\Sigma_{i\in \mathrm{I^J}}\phi_i(\tilde{e})$ from (X.14) and use has been made of the fact that the damage function is linear by assumption. Since $\Sigma_{i\in \mathrm{I^J}}\pi_i(\tilde{e})=\Sigma_{i\in \mathrm{I^J}}\beta_i(\tilde{e}_i^{\mathrm{J}})-\Sigma_{i\in \mathrm{I^J}}\phi_i(\tilde{e})$ and term A is positive,[15] (X.10) is established.

(X.11) is shown by noting that:

$$\hat{\pi}_j^{\psi}=\beta_j(e_j^{\mathrm{S}})-\phi_j(e^{\mathrm{S}})+\hat{t}_j=$$

$$\beta_j(\tilde{e}_j^{NJ})-\phi_j(e^{\mathrm{S}})+\frac{\phi_j'(e^{\mathrm{S}})}{\Sigma_{j\in \mathrm{I}}\phi_j'(e^{\mathrm{S}})}[\Sigma_{j\in \mathrm{I}}\beta_j(e_j^{\mathrm{S}})-\Sigma_{j\in \mathrm{I}}\beta_j(\tilde{e}_j^{NJ})] \tag{X.16}$$

and:

$$\pi_j^{*\psi}=\beta_j(e_j^{\mathrm{S}})-\phi_j(e^{\mathrm{S}})+t_j^*=$$

$$\beta_j(e_j^{N})-\phi_j(e^{\mathrm{S}})+\frac{\phi_j'(e^{\mathrm{S}})}{\Sigma_{j\in \mathrm{I}}\phi_j'(e^{\mathrm{S}})}[\Sigma_{j\in \mathrm{I}}\beta_j(e_j^{\mathrm{S}})-\Sigma_{j\in \mathrm{I}}\beta_j(e_j^{N})] \tag{X.17}$$

which implies:

$$\beta_j(\tilde{e}_j^{NJ})+\frac{\phi_j'(e^{\mathrm{S}})}{\Sigma_{j\in \mathrm{I}}\phi_j'(e^{\mathrm{S}})}[\Sigma_{j\in \mathrm{I}}\beta_j(e_j^{\mathrm{S}})-\Sigma_{j\in \mathrm{I}}\beta_j(\tilde{e}_j^{NJ})]$$

$$\geq \beta_j(e_j^{N})+\frac{\phi_j'(e^{\mathrm{S}})}{\Sigma_{j\in \mathrm{I}}\phi_j'(e^{\mathrm{S}})}[\Sigma_{j\in \mathrm{I}}\beta_j(e_j^{\mathrm{S}})-\Sigma_{j\in \mathrm{I}}\beta_j(e_j^{N})] \tag{X.18}$$

according to (X.11). It easily checked that (X.18) in fact holds by noting that $e_j^N=\tilde{e}_j^{NJ}\ \forall\ j\notin \mathrm{I^J}$ due to linear damage cost functions and $e_i^N\geq\tilde{e}_i^{\mathrm{J}}\ i\in \mathrm{I^J}$ as established above, and hence $\beta_j(\tilde{e}_j^{NJ})=\beta_j(e_j^N)$ and $\Sigma_{j\in \mathrm{I}}\beta_j(\tilde{e}_j^{NJ})\leq\Sigma_{j\in \mathrm{I}}\beta_j(e_j^N)$ is true.

Case 2: $\forall\, \mathrm{I^J} \subset \mathrm{I}, |\, \mathrm{I^J}\, | \geq 2, \Sigma_{k \in \mathrm{I^J}} \phi_k'(e^{\mathrm{S}}) \geq \phi_i'(e^{\mathrm{N}}) \;\forall\, \mathrm{i} \in \mathrm{I^J}$

For the subsequent proof we need the following lemma (see Tulkens and Chander 1997, Section 6, Proposition 5):

Lemma X.2

Under the assumption of case 2, $\Sigma \tilde{e}_i \geq \Sigma e_i^{\mathrm{S}}\ \forall\ i \in \mathrm{I}$ and $e_i^N \geq \tilde{e}_i^{\mathrm{J}}\ \forall\ i \in \mathrm{I^J}$.

Proof: First, we show $\Sigma \tilde{e}_i \geq \Sigma e_i^{\mathrm{S}}$. Suppose instead $\Sigma \tilde{e}_i \leq \Sigma e_i^{\mathrm{S}}$ were true. Then:

$$\forall\ i \in \mathrm{I^J}\colon \beta_i'(\tilde{e}_i^{\mathrm{J}}) = \Sigma_{k \in \mathrm{I^J}} \phi_k'(\tilde{e}) \leq \Sigma_{k \in \mathrm{I^J}} \phi_k'(e^{\mathrm{S}}) < \Sigma_{k \in \mathrm{I}} \phi_k'(e^{\mathrm{S}}) = \beta_i'(e_i^{\mathrm{S}})$$

$$\forall\ j \notin \mathrm{I^J}\colon \beta_j'(\tilde{e}_j^{NJ}) = \phi_j'(\tilde{e}) \leq \phi_j'(e^{\mathrm{S}}) < \beta_j'(e_j^{\mathrm{S}})$$

which implies $\beta_k'(\tilde{e}_k) < \beta_k'(e_k^{\mathrm{S}})\ \forall\ k \in \mathrm{I}$. Therefore, $\tilde{e}_k > e_k^{\mathrm{S}}\ \forall\ k \in \mathrm{I}$ by $\beta_k'' < 0$ and consequently $\Sigma \tilde{e}_i \geq \Sigma e_i^{\mathrm{S}}$, which contradicts the initial assumption.

Second, from $\Sigma \tilde{e}_i \geq \Sigma e_i^{\mathrm{S}}$, as established above, $\Sigma_{k \in \mathrm{I^J}} \phi_k'(\tilde{e}) \geq \Sigma_{k \in \mathrm{I^J}} \phi_k'(e^{\mathrm{S}})$ follows by $\phi_k'' \geq 0$. Hence:

$$\beta_i'(\tilde{e}_i^{\mathrm{J}}) = \Sigma_{k \in \mathrm{I^J}} \phi_k'(\tilde{e}_k) \geq \Sigma_{k \in \mathrm{I^J}} \phi_k'(e^{\mathrm{S}}) \geq \phi_i'(e^N) = \beta_i'(e_i^N) \qquad \text{(X.19)}$$

where the second inequality is an implication of the central assumption in case 2. Obviously, (X.19) can only hold provided $e_i^N \geq \tilde{e}_i^{\mathrm{J}}\ \forall\ i \in \mathrm{I^J}$ due to $\beta_i'' < 0$. QED

Next we have to show that $\pi^{*\psi}$ lies in the core. As in case 1, this is shown by contradication. Thus, exactly the same arguments which have been presented when going from (X.8) to (X.12) apply. Consequently, it has to be shown that (X.10) and (X.11) hold in the present case 2.

(X.10) implies:

$$\Sigma_{i \in \mathrm{I^J}} \hat{t_i} + \Sigma_{i \in \mathrm{I^J}} \beta_i(e_i^{\mathrm{S}}) - \Sigma_{i \in \mathrm{I^J}} \phi_i(e^{\mathrm{S}}) > \Sigma_{i \in \mathrm{I^J}} \tilde{t}_i + \Sigma_{i \in \mathrm{I^J}} \beta_i(\tilde{e}_i^{\mathrm{J}}) - \Sigma_{i \in \mathrm{I^J}} \phi_i(\tilde{e}) \qquad \text{(X.20)}$$

where $\Sigma_{i \in \mathrm{I^J}} \tilde{t}_i = 0$ by definition. Note that (with a slight abuse of notation, for example, $\Sigma \tilde{e} = \tilde{e}$ and so on):

$$\begin{aligned}
&\Sigma_{i \in \mathrm{I^J}} \beta_i(\tilde{e}_i^{\mathrm{J}}) - \Sigma_{i \in \mathrm{I^J}} \phi_i(\tilde{e}) \\
&\leq \Sigma_{i \in \mathrm{I^J}} \beta_i(\tilde{e}_i^{\mathrm{J}}) - \Sigma_{i \in \mathrm{I^J}} \phi_i(\tilde{e}) + \Sigma_{i \in \mathrm{I^J}} \phi_i(\tilde{e}) - \Sigma_{i \in \mathrm{I^J}} \phi_i(e^{\mathrm{S}}) - \Sigma_{i \in \mathrm{I^J}} \phi_i'(e^{\mathrm{S}})(\tilde{e} - e^{\mathrm{S}}) \\
&= \Sigma_{i \in \mathrm{I^J}} \beta_i(\tilde{e}_i^{\mathrm{J}}) - \Sigma_{i \in \mathrm{I^J}} \phi_i(e^{\mathrm{S}}) - \Sigma_{i \in \mathrm{I^J}} \phi_i'(e^{\mathrm{S}})(\tilde{e} - e^{\mathrm{S}}) \qquad \text{(X.21)}
\end{aligned}$$

since $\phi_i(\tilde{e}) - \phi_i(e^{\mathrm{S}})/(\tilde{e} - e^{\mathrm{S}}) > \phi_i'(e^{\mathrm{S}})$ by the convexity of the damage cost function and $\tilde{e} > e^{\mathrm{S}}$ by Lemma X.2. Using (X.21) and recalling $\Sigma_{i \in \mathrm{I^J}} \tilde{t}_i = 0$, (X.20) reads:

$$\Sigma_{i\in \mathrm{I}^{\mathrm{J}}}\hat{t}_i+\Sigma_{i\in \mathrm{I}^{\mathrm{J}}}\beta_i(e_i^{\mathrm{S}})-\Sigma_{i\in \mathrm{I}^{\mathrm{J}}}\phi_i(e^{\mathrm{S}})\geq\Sigma_{i\in \mathrm{I}^{\mathrm{J}}}\beta_i(\tilde{e}_i^{\mathrm{J}})-\Sigma_{i\in \mathrm{I}^{\mathrm{J}}}\phi_i(e^{\mathrm{S}})-\Sigma_{i\in \mathrm{I}^{\mathrm{J}}}\phi_i'(e^{\mathrm{S}})(\tilde{e}-e^{\mathrm{S}}) \tag{X.22}$$

or:

$$\Sigma_{i\in \mathrm{I}^{\mathrm{J}}}\hat{t}_i+\Sigma_{i\in \mathrm{I}^{\mathrm{J}}}\beta_i(e_i^{\mathrm{S}})\geq\Sigma_{i\in \mathrm{I}^{\mathrm{J}}}\beta_i(\tilde{e}_i^{\mathrm{J}})-\frac{\Sigma_{i\in \mathrm{I}^{\mathrm{J}}}\phi_i'(e^{\mathrm{S}})}{\Sigma_{i\in \mathrm{I}}\phi_i'(e^{\mathrm{S}})}\Sigma_{i\in \mathrm{I}}\phi_i'(e^{\mathrm{S}})(\tilde{e}-e^{\mathrm{S}}) \tag{X.23}$$

or, using $\hat{t}_i$ as defined in (X.9):

$$-\Sigma_{i\in \mathrm{I}^{\mathrm{J}}}\beta_i(e_i^{\mathrm{S}})+\Sigma_{i\in \mathrm{I}^{\mathrm{J}}}\beta_i(\tilde{e}_i^{\mathrm{J}})+\sum_{i\in \mathrm{I}^{\mathrm{J}}}\frac{\phi_i'(e^{\mathrm{S}})}{\Sigma_{i\in \mathrm{I}}\phi_i'(e^{\mathrm{S}})}(\Sigma_{i\in \mathrm{I}}\beta_i(e_i^{\mathrm{S}})-\Sigma_{i\in \mathrm{I}}\beta_i(\tilde{e}_i))$$

$$\geq\Sigma_{i\in \mathrm{I}^{\mathrm{J}}}\beta_i(\tilde{e}_i^{\mathrm{J}})-\frac{\Sigma_{i\in \mathrm{I}^{\mathrm{J}}}\phi_i'(e^{\mathrm{S}})}{\Sigma_{i\in \mathrm{I}}\phi_i'(e^{\mathrm{S}})}\Sigma_{i\in \mathrm{I}}\phi_i'(e^{\mathrm{S}})(\tilde{e}-e^{\mathrm{S}}) \tag{X.24}$$

after further manipulation one derives:

$$\Sigma_{i\in \mathrm{I}}\phi_i'(e^{\mathrm{S}})(\tilde{e}-e^{\mathrm{S}})\geq\Sigma_{i\in \mathrm{I}}(\beta_i(e_i^{\mathrm{S}})-\beta_i(\tilde{e}_i)). \tag{X.25}$$

Using the FOC $\Sigma_{i\in \mathrm{I}}\phi_i'(e^{\mathrm{S}})=\beta_i'(e_i^{\mathrm{S}})$ (and the original notation) we have:

$$\beta_i'(e_i^{\mathrm{S}})(\Sigma_{i\in \mathrm{I}}\tilde{e}_i-\Sigma_{i\in \mathrm{I}}e_i^{\mathrm{S}})\geq\Sigma_{i\in \mathrm{I}}(\beta_i(e_i^{\mathrm{S}})-\beta_i(\tilde{e}_i)). \tag{X.26}$$

A sufficient condition for (X.26) to hold is:

$$\beta_i'(e_i^{\mathrm{S}})(\tilde{e}_i-e_i^{\mathrm{S}})\geq\beta_i(\tilde{e}_i)-\beta_i(e_i^{\mathrm{S}})\forall\ i\in \mathrm{I}. \tag{X.27}$$

It is easily checked that (X.27) holds regardless whether $\tilde{e}_i>e_i^{\mathrm{S}}$ or $\tilde{e}_i\leq e_i^{\mathrm{S}}$ by the concavity of the benefit function.

Next we have to show that (X.11) holds. Hence:

$$\Sigma_{j\notin \mathrm{I}^{\mathrm{J}}}\beta_j(e_j^{\mathrm{S}})+\Sigma_{j\notin \mathrm{I}^{\mathrm{J}}}\hat{t}_j\geq\Sigma_{i\notin \mathrm{I}^{\mathrm{J}}}\beta_j(e_j^{\mathrm{S}})+\Sigma_{j\notin \mathrm{I}^{\mathrm{J}}}\hat{t}_j^{*} \tag{X.28}$$

must be true ($\Sigma_{i\notin \mathrm{I}^{\mathrm{J}}}\phi_i'(e_j^{\mathrm{S}})$ on both sides has been dropped). Using (X.9), the LHS term may be written as:

$$\Sigma_{j\notin \mathrm{I}^{\mathrm{J}}}\beta_j(e_j^{\mathrm{S}})+\Sigma_{j\notin \mathrm{I}^{\mathrm{J}}}\hat{t}_j$$

$$=\Sigma_{j\notin \mathrm{I}^{\mathrm{J}}}\beta_j(\tilde{e}_j^{NJ})+\sum_{j\notin \mathrm{I}^{\mathrm{J}}}\frac{\phi_j'(e^{\mathrm{S}})}{\Sigma_{j\in \mathrm{I}}\phi_j'(e^{\mathrm{S}})}(\Sigma_{j\in \mathrm{I}}\beta_j(e_j^{\mathrm{S}})-\Sigma_{j\in \mathrm{I}}\beta_j(\tilde{e}_j))$$

$$\Leftrightarrow \quad \Sigma_{j\notin \mathrm{I^J}}\beta_j(e_j^N)+\sum_{j\notin \mathrm{I^J}}\frac{\phi_j'(e^{\mathrm{S}})}{\Sigma_{j\in \mathrm{I}}\phi_j'(e^{\mathrm{S}})}(\Sigma_{j\in \mathrm{I}}\beta_j(e_j^N)-\Sigma_{j\in \mathrm{I}}\beta_j(e_j^S))$$

$$+\Sigma_{j\notin \mathrm{I^J}}\beta_j(\tilde{e}_j^{NJ})-\Sigma_{j\notin \mathrm{I^J}}\beta_j(e_j^N)+\sum_{j\notin \mathrm{I^J}}\frac{\phi_j'(e^{\mathrm{S}})}{\Sigma_{j\in \mathrm{I}}\phi_j'(e^{\mathrm{S}})}(\Sigma_{j\in \mathrm{I}}\beta_j(e_j^N)-\Sigma_{j\in \mathrm{I}}\beta_j(\tilde{e}_j))$$

$$\Leftrightarrow \quad \Sigma_{j\notin \mathrm{I^J}}\beta_j(e_j^{\mathrm{S}})+\Sigma_{j\notin \mathrm{I^J}}t_j^*+\Sigma_{j\notin \mathrm{I^J}}\beta_j(\tilde{e}_j^{NJ})-\Sigma_{j\in \mathrm{I}}\beta_j(\tilde{e}_j^N)$$

$$+\sum_{j\notin \mathrm{I^J}}\frac{\phi_j'(e^{\mathrm{S}})}{\Sigma_{j\in \mathrm{I}}\phi_j'(e^{\mathrm{S}})}(\Sigma_{j\in \mathrm{I}}\beta_j(e_j^N)-\Sigma_{j\in \mathrm{I}}\beta_j(\tilde{e}_j))$$

$$\Leftrightarrow \quad \Sigma_{j\notin \mathrm{I^J}}\beta_j(e_j^{\mathrm{S}})+\Sigma_{j\notin \mathrm{I^J}}t_j^*+\Sigma_{j\notin \mathrm{I}}\beta_j(\tilde{e}_j^{NJ})-\Sigma_{j\notin \mathrm{I^J}}\beta_j(e_j^N)$$

$$+\sum_{j\notin \mathrm{I^J}}\frac{\phi_j'(e^{\mathrm{S}})}{\Sigma_{j\in \mathrm{I}}\phi_j'(e^{\mathrm{S}})}(\Sigma_{j\notin \mathrm{I^J}}\beta_j(e_j^N)-\Sigma_{j\notin \mathrm{I^J}}\beta_j(\tilde{e}_j^{NJ}))$$

$$+\sum_{j\notin \mathrm{I^J}}\frac{\phi_j'(e^{\mathrm{S}})}{\Sigma_{j\in \mathrm{I}}\phi_j'(e^{\mathrm{S}})}(\Sigma_{j\notin \mathrm{I^J}}\beta_j(e_j^N)-\Sigma_{j\notin \mathrm{I^J}}\beta_j(\tilde{e}_j^{\mathrm{J}}))$$

$$\Leftrightarrow \quad \Sigma_{j\notin \mathrm{I^J}}\beta_j(e_j^{\mathrm{S}})+\Sigma_{j\notin \mathrm{I^J}}t_j^*+\left(1-\sum_{j\notin \mathrm{I^J}}\frac{\phi_j'(e^{\mathrm{S}})}{\Sigma_{j\in \mathrm{I}}\phi_j'(e^{\mathrm{S}})}\right)(\Sigma_{j\notin \mathrm{I^J}}\beta_j(\tilde{e}_j^{NJ})-\Sigma_{j\notin \mathrm{I^J}}\beta_j(e_j^N))$$

$$-\sum_{j\in \mathrm{I^J}}\frac{\phi_j'(e^{\mathrm{S}})}{\Sigma_{j\in \mathrm{I}}\phi_j'(e^{\mathrm{S}})}(\Sigma_{j\in \mathrm{I^J}}\beta_j(\tilde{e}_j^{\mathrm{J}})-\Sigma_{j\in \mathrm{I^J}}\beta_j(e_j^N)). \qquad \text{(X.29)}$$

Note that the first term in brackets is obviously positive; the second term in brackets as well, since $\tilde{e}_j^{NJ} \geq e_j^N \; \forall \, j \notin \mathrm{I^J}$ by Lemma X.1, and the third term in brackets is negative since $\tilde{e}_j^{\mathrm{J}} \geq e_j^N \; \forall \; j \in \mathrm{I^J}$, as has been established in Lemma X.2 above. Taken together, (X.10) ((X.20)) and (X.11) ((X.28)) have been shown to be true in case 2, which completes the proof. QED

Case 3: Symmetric countries
Proof is obvious and therefore omitted. QED

X.3 Appendix 3

In this appendix we prove that for the transfer scheme (13.30) in the text the imputation as defined in Proposition 13.6 lies in the core at each stage t.

Proof: We first have to establish that the iterative process is individually rational; that is, each country gains along this process: $\Delta\pi_{i,t} \geq 0 \; \forall \; t$ and $i \in \mathrm{I}$ (Germain *et al.* 1995, Theorem 2).

Note that:

$$\Delta\Sigma\pi_{i,t} = \Sigma\pi^*_{i,t+1} - \Sigma\pi^*_{i,t} \tag{X.30}$$

which, assuming a linear damage cost function, that is, $\phi_i = d_i \Sigma e_k$ (implying $\Delta\phi_{i,t} = d_i \Delta\Sigma e_{k,t}$), implies:

$$\Delta\Sigma\pi_{i,t} = \Sigma\Delta\beta_{i,t} - \Sigma d_i \Sigma\Delta e^*_{k,t} \Leftrightarrow \Sigma[\Delta\beta_{i,t} - d\Delta e^*_{k,t}] =: \Sigma\Delta g_{i,t},\ d = \Sigma d_i \tag{X.31}$$

where *g* stands for gains or surplus from cooperation. From (X.8) it follows that:

$$g_{i,t} = \beta_i(e^*_{i,t}) - d \cdot e^*_{i,t}. \tag{X.32}$$

Due to the concavity of the function g_i (since β_i is a concave function and ϕ_i is a linear function) we can write:

$$\Delta g_{i,t} = g_i(e^*_{i,t+1}) - g_i(e^*_{i,t}) \geq g_i'(e^*_{i,t+1})[e^*_{i,t+1} - e^*_{i,t}] \tag{X.33}$$

where:

$$g_i'(e^*_{i,t+1}) = \beta_i'(e^*_{i,t+1}) - d. \tag{X.34}$$

Since by definition of a local optimum either $g_i'(e^*_{i,t+1}) \leq 0$ if $e^*_{i,t} = e^*_{i,t+1}$ or $g_i'(e^*_{i,t+1}) = 0$ if $e^*_{i,t} \neq e^*_{i,t+1}$ hold, $\Delta g_{i,t} = \Delta\pi_{i,t} \geq 0\ \forall\ i$ and t has been established.

Second, we have to show that, given the transfer (13.30) in the text, the imputation is individually rational. This is straightforward since:

$$\Delta\pi^{\psi}_{i,t} = \Delta\beta_i - \Delta\phi_i + \Delta t_i \tag{X.35}$$

which, given:

$$\Delta t_{i,t} = -(\Delta\beta_{i,t} - \Delta\phi_{i,t}) + \Sigma\psi_i \Delta g_{j,t} \tag{X.36}$$

reduces to $\Delta\pi^{\psi}_{i,t} = \Sigma\psi_i \Delta g_{j,t} \geq 0$ where $\psi_i = \phi_i'/\Sigma\phi_k'$ is proposed by Chander and Tulkens, which implies $\psi_i = d_i/d$ for linear damage cost functions.

Third, it has to be shown that the imputation as defined in Proposition 13.6 lies in the core. Since the proof proceeds exactly along the same lines as the proof for the linear damage cost function as given in Appendix X.2, case 1, the proof is omitted here. The only difference is that the socially optimal emission vector e^S has to be replaced by the locally optimal emis-

sion vector e_t^* and the PANE with respect to the coalition I^J has to be indexed by time t. Both PANE and the local global optimum are defined with respect to constraint (13.29) in the text. QED

XI CHAPTER 14: APPENDICES

XI.1 Appendix 1

In this appendix we provide all the background information referred to in Sections 14.2 and 14.4.

For payoff functions (14.1), emissions in the Nash equilibrium are given by:

$$e_i^N = \frac{d(2\gamma + N + 1 - 2i)}{2\gamma + N + 1},\ \Sigma e_i^N = \frac{2dN\gamma}{2\gamma + N + 1} \tag{XI.1}$$

with associated payoffs:

$$\pi_i^N = \frac{bd^2(4b^2 + 4bcN + 4bc - 4bciN + c^2N^2 + 2c^2N + c^2 - 4c^2i^2)}{2(2b + cN + c)^2},$$

$$\Sigma\pi_i^N = \frac{bd^2N(12b^2 + 12bc + 6bcN - 6bcN^2 + c^2 - c^2N^2)}{6(2b + cN + c)^2}. \tag{XI.2}$$

A sufficient condition to ensure positive emissions is:

$$\text{NNC}_1\text{: } \gamma \geq \frac{N-1}{2}, \tag{XI.3}$$

where NNC stands for 'non-negativity constraint'. This condition is considered in all simulations in Section 13.3. For the social optimum we find:

$$e_i^S = \frac{2d\gamma}{2\gamma + N^2 + N},\ \Sigma e_i^S = \frac{2dN\gamma}{2\gamma + N^2 + N}, \tag{XI.4}$$

$$\pi_i^S = \frac{2b^2d^2(b + cN^2 + cN - icN)}{(2b + cN^2cN)^2},\ \Sigma\pi_i^S = \frac{b^2d^2N}{2b + cN^2 + cN}, \tag{XI.5}$$

for which it is easily checked that $\partial e_i^N/\partial i < 0$, $\partial(\pi_i^S - \pi_i^N)/\partial i > 0$ and $\pi_i^S - \pi_i^N <, >, = 0$ hold, as stated in Section 14.2. Using (XI.1)–(XI.5) the result stated in (14.2) in the text can be confirmed.

According to the procedure outlined in Chapter 11, the proposal of country i under the quota and tax regime can be derived. We find:

$$r_i = \frac{2i(2N\gamma - 2\gamma - N - 1 + 2i)}{4\gamma^2 + 4N\gamma + 4\gamma - 8i\gamma + 4iN\gamma + N^2 + 2N - 4Ni + 1 - 4i + 4i^2}$$

$$t_i = \frac{iNdbc}{b + icN} \tag{XI.6}$$

for which $\partial r_i/\partial i > 0$ and $\partial t_i/\partial i > 0$ is true. Hence country 1 is the bottleneck under the LCD decision rule and by substituting $i = 1$ in (XI.6), the agreement according to the LCD decision rule, that is, $r^{\mathrm{A}} = r_1$ and $t^{\mathrm{A}} = t_1$ is determined. From this *global* emission reduction (based on initial emissions $\Sigma e_i^{\mathrm{I}} = \Sigma e_i^N$ as referred to in Table 14.1), $r^{\mathrm{Q}1}$ and $r^{\mathrm{T}1}$, can be computed to be ($r^{\mathrm{Q}1} = r_1$ and $\Sigma e_i^{\mathrm{T}} = (1 - r^{\mathrm{T}1})$. Σe_i^N where $e_i^{\mathrm{T}} = e_i(t_1)$):

$$r^{\mathrm{Q}1} = \frac{2(2N\gamma - 2\gamma - N - 1)}{4\gamma^2 + 8N\gamma - 4\gamma + N^2 - 2N + 1}, \; r^{\mathrm{T}1} = \frac{N - 1}{2(\gamma + N)}. \tag{XI.7}$$

A globally optimal solution, given the constraint of a uniform application of the instrument, leads to:

$$r^{\mathrm{Q}^*} = \frac{(N^2 - 1)(6\gamma + 1)}{12\gamma^2 + 6\gamma N + N^2 - 1 + 6\gamma N^2}, \; r^{\mathrm{T}^*} = \frac{N^2 - 1}{2\gamma + N^2 + N} \tag{XI.8}$$

where $r^{\mathrm{T}^*} = r^{\mathrm{S}^*}$. Substitution of $i = (N+1)/2$ into (XI.6) leads to the median country proposal for which it is easily checked that this implies global emission reductions as given in (XI.8).

For a sub-coalition the proposal of a potential signatory i is derived as follows. In the quota regime, two pieces of information are used. First, country i knows that countries 1 to N_0 choose their initial emission level e_i^N. Therefore, emissions of all non-signatories are given by $\Sigma e_j^{N\mathrm{J}} = \Sigma_{j=1}^{N_0} e_j^N$ where $j \in \mathrm{I}^{N\mathrm{J}} = \{1, 2, ..., N_0\}$. Second, if countries $N_0 + 1$ to N comply with country i's proposal, r_i, then signatories' emissions will be $\Sigma e_k^{\mathrm{J}} = \Sigma_{k=N_0+1}^{N} e_k^N \cdot (1 - r_i)$ where $i, k \in \mathrm{I}^{\mathrm{J}} = \{N_0 + 1, \ldots, N\}$. Consequently, aggregate emissions are given by $\Sigma e_k^{\mathrm{Q}} = \Sigma e_k^{N\mathrm{J}} + \Sigma e_k^{\mathrm{J}}$. Using payoff function (14.1) in the text, it is straightforward to compute a signatory i's proposal to be:

$$r_i = \frac{2N\gamma i(2N\gamma - 2N_0\gamma + N_0^2 - NN_0 - N + 2i - 2\gamma)}{\mathrm{A}} \tag{XI.9}$$

$$\begin{aligned}\mathrm{A}: = {} & 4\gamma^3 N - 8i\gamma^2 N - 4iN^2\gamma + \gamma N - 4i\gamma N + 4\gamma^2 N + 2\gamma N^2 - 8i\gamma^2 NN_0 + 4i\gamma^2 N^2 \\ & - 4i\gamma N^2 N_0 + 4i\gamma^2 N_0^2 - 2iNN_0^3 - 4i\gamma N_0^3 + 4i^2\gamma N + 8i\gamma NN_0^2 + iN^2 N_0^2 \\ & + \gamma N^3 + 4\gamma^2 N^2 + iN_0^4\end{aligned}$$

where $\partial r_i/\partial i > 0$ holds. That is, the signatory with the lowest index number is the bottleneck country according to the LCD decision rule, which is country $N_0 + 1$.

According to a similar procedure, a signatory's tax proposal:

$$t_i = \frac{bcdj(N - N_0)(2Nb + N^2c - N_0^2c + Nc - N_0c)}{(2b + Nc + c)(Nb + N^2jc + jN_0^2c - 2jN_0Nc)} \qquad \text{(XI.10)}$$

is derived where, again $\partial t_i/\partial i > 0$ holds, implying that the proposal of the signatory with the lowest index is accepted within the IEA. In order to rule out corner solutions, that is, $e_i^T \geq 0 \; \forall \; i \in I^J = \{N_0 + 1, ..., N\}$ (recall non-signatories emit e_j^N as given in (XI.1) for which $e_j^N \geq 0$ holds by NNC_1 as given in (XI.3)), a sufficient condition is:

$$NNC_2: \gamma > N^2/6 \qquad \text{(XI.11)}$$

which is considered in all simulations in Section 14.4 and which is a more restrictive condition than NNC_1 in (XI.3).

XI.2 Appendix 2

To test whether an emission tuple $e^* = (e_1^*, e_2^*, ..., e_N^*)$ of the grand coalition is a WRPE if $\delta \to 1$ the following algorithm is used:

A. Set $i = 1$.
B. Determine the amount Δe_{-i} by which all countries $-i$ have to increase their emissions during the punishment of country i in order to fulfill (12.1) in the text.
   - B.I Determine the reaction function of country i, $e_i(e_{-i})$.
   - B.II Insert $e_{-i}^i = e_{-i}^* + \Delta e_{-i}$ in (12.1) and solve for $\Delta e_{-i} > \Delta e_{\min}$. Define $\Delta e_{-i} := \Delta e_{\min} + \varepsilon$ where ε is a sufficiently small positive number.
C. Check whether (12.2) in the text can be satisfied for all countries $-i$ if they jointly increase emissions by the amount Δe_{-i}.
   - C.I Set $e_i^i = 0$ in (12.2) to determine the outer boundaries of the WRPE emission space.[16] Hence, $e_{-i}^i = e_{-i}^* + \Delta e_{-i}$ where Δe_{-i} has been determined in step B.II.
   - C.II For all $j \neq i$ let $e_j^i := e_j^* + \Delta e_j$ with Δe_j the amount by which country j increases its emission to punish i. Further, define $\Phi_{\min} := 0$ and $\Phi_{\max} := 0$.
   - C.III Repeat for all $j \neq i$:
     Insert e_{-i}^i in (12.2). Solve (12.2) for Δe_j to have $\Delta e_j \in [\Delta e_{j,\min}, \Delta e_{j,\max}]$ or $\Delta e_j \in \varnothing$. In the first case define $\Phi_{\min} := \Phi_{\min} + \Delta e_{j,\min}$ and $\Phi_{\max} := \Phi_{\max} + \Delta e_{j,\max}$. In the second case stop C.III and define $\Phi_{\min} := 0$ and $\Phi_{\max} := 0$.
   - C.IV If, and only if $\Delta e_{-i} \in [\Phi_{\min}, \Phi_{\max}]$, then country i can be punished by WRPE strategies. If this is the case continue with C.V; otherwise stop.

C.V Define $i := i+1$. Repeat all steps starting from B. If the algorithm runs for all i, $i \in \{1, \ldots, N\}$, then e^* is a WRPE.

XI.3 Appendix 3

For country i the determination of $\delta_i^{\min}$ (in the case of a grand coalition) proceeds as follows:

A. Set $\delta_i = 1$.
B. Set $\delta_i := \delta_i - \kappa$ where κ is a suitable step size (for example, $\kappa = 0.01$ or $\kappa = 0.001$).
C. Determine the amount $\Delta e^a_{-i,\min}$ by which all countries $-i$ have to increase their emissions during the punishment of i in order to satisfy (12.24) in the text.
   C.I Determine the best deviation in the normal phase of country i, $e_i(e^*_{-i})$, to have π_i^D.
   C.II Insert $e^i_{-i} = e^*_{-i} + \Delta e_{-i}$ in (12.24) and solve for $\Delta e_{-i} \geq \Delta e^a_{-i,\min}$.
D. Check whether there exist e^i_i and $\Delta e_{-i} \geq \Delta e^a_{-i,\min}$, so that (12.23) and (12.2) in the text can be satisfied jointly.
   D.I Define $e^i_i := 0$ in (12.23) and (12.2) in the text.
   D.II Solve (12.23) for $\Delta e_{-i} \geq \Delta e^b_{-i,\min}$. Define $\Delta e_{-i} := \max\{\Delta e^a_{-i,\min}, \Delta e^b_{-i,\min}\}$.
   D.III Let $e^i_j := e^*_j + \Delta e_j$ where Δe_j is the amount by which country j increases its emission to punish i. Moreover, let $\Sigma e^i_k = e^i_i + e^i_{-i} = e^i_i + e^*_{-i} + \Delta e_{-i}$ and define $\Phi_{\min} := 0$ and $\Phi_{\max} := 0$.
   D.IV Repeat for all $j \neq i$:
   Evaluate Σe^i_k and insert this in (12.2). Solve (12.2) for Δe_j to have $\Delta e_j \in [\Delta e_{j,\min}, \Delta e_{j,\max}]$ or $\Delta e_j \in \varnothing$. In the first case $\Phi_{\min} := \Phi_{\min} + \Delta e_{j,\min}$ and $\Phi_{\max} := \Phi_{\max} + \Delta e_{j,\max}$. In the second case stop D.IV and define $\Phi_{\min} := 0$, $\Phi_{\max} := 0$.
   D.V (a) If $\Delta e_{-i} \in [\Phi_{\min}, \Phi_{\max}]$ then i can be punished using WRPE strategies. In this case start again with step B in order to test if δ_i may be further reduced.
   (b) If $\Delta e_{-i} \notin [\Phi_{\min}, \Phi_{\max}]$ and $e^i_i < e^*_i$ define $e^i_i := e^i_i + h$, with h a suitable step size. Go back to D.II.
   (c) If $\Delta e_{-i} \notin [\Phi_{\min}, \Phi_{\max}]$ and $e^i_i \geq e^*_i$, a punishment using WRPE strategies is not possible. Then $\delta_i^{\min} := \delta_i + \kappa$. Stop algorithm.

XI.4 Appendix 4

The test of stability of a sub-coalition if $\delta \to 1$ proceeds along the same lines as laid out in Appendix XI.2. First, the amount Δe_{-i} by which the signatories have to increase their emissions to punish a non-signatory to satisfy (12.1) is determined. Then, it is checked whether the signatories can provide such a Δe_{-i} under the restriction that (12.2) is met for each signatory. Accordingly, the same procedure is applied in the case where a signatory breaches the contract.

XII CHAPTER 15: APPENDICES

XII.1 Appendix 1

In Proposition 15.1 we claim that a symmetric global emission game with payoff functions (13.1) satisfies conditions C_1–C_4.

Proof: The payoff function $\pi_i = \beta(e_i) - \phi(\Sigma e_j)$ may be written as $\pi_i = \beta(e_i) - f(\Sigma e_j)$ for a linear damage cost function where $f > 0$ denotes constant marginal damages. The FOC of a coalition c_i of size $|c_i|$, maximizing aggregate payoffs of its members is given by:

$$\beta'(e_i) = |c_i| \cdot f \tag{XII.1}$$

from which it is evident that $\partial e_i / \partial |c_i| < 0$ (since $\beta'' < 0$). That is, members of larger coalitions emit less than members of smaller coalitions. Since all countries suffer equally from damages, members of smaller coalitions receive a higher payoff than those of larger coalitions (C_2). Next consider that two coalitions i and j merge. Then for payoff funtion (13.1) and symmetric countries – with slight abuse of notation – equilibrium emissions are given in the initial situation by:

$$e_i(|c_i|) = \frac{bdN - |c_i| \cdot c}{Nb},\ e_j(|c_j|) = \frac{bdN - |c_j| \cdot c}{Nb} \tag{XII.2}$$

and after they have merged ($c_k = \{\{c_i\} \cup \{c_j\}\}$) by:

$$e_k(|c_k|) = \frac{bdN - (|c_i| + |c_j|) \cdot c}{Nb}. \tag{XII.3}$$

Since equilibrium emissions of countries outside both coalitions are not affected by the merger, a comparison of aggregate emissions *before* and *after* the merger is straightforward. We find:

$$|c_i|\cdot e_i(|c_i|)+|c_j|\cdot e_j(|c_j|)-(|c_i|+|c_j|)\cdot e_k(|c_k|)=\frac{(|c_i|\cdot|c_j|)\cdot c}{Nb}>0. \quad \text{(XII.4)}$$

Thus a merger reduces global emissions and hence outsiders benefit from a merger (C_1). Next consider what happens if a member of coalition *i* leaves to join an equal-sized or larger coalition *j*, that is, $|c_i|\leq|c_j|$. Then equilibrium emissions are given by:

$$e_i(|c_i|-1)=\frac{bdN-(|c_i|-1)\cdot c}{Nb},\ e_j(|c_j|+1)=\frac{bdN-(|c_j|+1)\cdot c}{Nb}. \quad \text{(XII.5)}$$

Since, again, equilibrium emissions of countries outside both coalitions are not affected by the merger, a comparison of aggregate emissions *before* and *after* the change of membership is easily computed. We find:

$$\begin{aligned}|c_i|\cdot e_i(|c_i|)+|c_j|\cdot e_j(|c_j|)-(|c_i|-1)\cdot e_i(|c_i|-1)-(|c_j|+1)\cdot e_j(|c_j|+1)\\ =2c\frac{|c_j|-|c_i|+1}{Nb}>0,\ |c_i|\leq|c_j|.\end{aligned} \quad \text{(XII.6)}$$

This has two implications. First, members of the 'old' coalition *i* are better off if a member leaves to join coalition *j*. On the one hand, global emissions decrease through such a move; on the other hand, members of coalition *i* increase emissions after the move. Thus, the remaining members in coalition *i* must be better off (C_3). For the deviator two effects must be considered. On the one hand, s/he has to carry a higher abatement burden in the new coalition which decreases his/her benefits. On the other hand, global emissions decrease, implying lower damage. To sign the overall effect, we compute payoffs of country *k* which left coalition *i* before and after the accession to coalition *j*. We find:

$$\pi_k(|c_i|,c)-\pi_k(|c_j|+1,c')=\frac{c^2(|c_i|+|c_j|-3)(|c_j|-|c_i|+1)}{2N^2b}>0,|c_i|\leq|c_j|. \quad \text{(XII.7)}$$

Hence, the deviator loses through accession to coalition *j* (C_4). QED

Remark: Note in addition the following relations which are useful for establishing the incentive profile in Appendix XII.2:

$$\pi_j(|c_j|,c)-\pi_j(|c_j|+1,c')=\frac{c^2(4\cdot|c_i|-2\cdot|c_j|-3)}{2N^2b}>\leq 0. \quad \text{(XII.8)}$$

The members of the coalition *j* may gain or lose if a member of coalition *i* joins. The larger coalition *j* is compared to coalition *i*, the more likely will a member of coalition *j* gain from the accession. If the country which joins

coalition *j* belonged to a singleton coalition, then coalition *j* will undoubtedly gain.

The following result looks at the effect of a merger of two coalitions:

$$\pi_i(|c_i|,c)-\pi_i(|c_i|+|c_j|,c')=\frac{c^2\cdot|c_j|\cdot(|c_j|-2\cdot|c_i|)}{2N^2b}>\leq 0 \qquad \text{(XII.9)}$$

$$\pi_j(|c_j|,c)-\pi_j(|c_i|+|c_j|,c')=\frac{c^2\cdot|c_i|\cdot(|c_i|-2\cdot|c_j|)}{2N^2b}<0. \qquad \text{(XII.10)}$$

That is, members of the larger or equal-sized coalition *j* always gain from a merger, the members of the smaller coalition *i* gain provided they belong to a coalition of at least half the size of the larger coalition *j*.

XII.2 Appendix 2

In this appendix we derive the incentive profile of (*ex ante* symmetric) countries with payoff function (13.1). The profile is derived for $N \in [3, 4, 5]$. Note that for symmetric countries only the coalition structure as such and not the membership matters. (That is, for instance, {{1, 2},{3}} is identical to {{1}, {2, 3}} and{{1, 3},{2}}.)

Three countries

There are three permutations to be considered: {{1, 2, 3}}, {{1, 2},{3}} and {{1},{2},{3}} (Table XII.1).

*Table XII.1 Incentive profile of countries (*N*=3)*

Change of permutation		1	2	3	Reason
{{1, 2, 3}}	→{{1, 2},{3}}	−	−	0	1, 2: (XII.10); 3: (XII.9)
{{1, 2, 3}}	→{{1}, {2},{3}}	−	−	−	1, 2, 3: repeated appl. of (XII.9) and (XII.10)
{{1, 2}, {3}}	→{{1}, {2},{3}}	−	−	−	1, 2: (XII.10); 3: C_1

From the profile it is evident that all three permutations are stand-alone stable.

Four countries

There are five permutations to be considered: {{1, 2, 3, 4}}, {{1, 2, 3},{4}}, {{1, 2},{3, 4}}, {{1, 2},{3},{4}} and {{1},{2},{3},{4}}. Except for the first permutation, all permutations are stand-alone stable (Table XII.2).

*Table XII.2 Incentive profile of countries (*N=*4)*

Change of permutation		1	2	3	4	Reason
{{1, 2, 3, 4}}	→ {{1, 2, 3},{4}}	−	−	−	+	1, 2, 3: (XII.10); 4: (XII.9)
{{1, 2, 3, 4}}	→ {{1, 2}, {3, 4}}	−	−	−	−	1–4: (XII.10)
{{1, 2, 3, 4}}	→ {{1, 2}, {3},{4}}	−	−	−	−	1, 2: C_1; 3, 4: (XII.10)
{{1, 2, 3, 4}}	→ {{1}, {2}, {3},{4}}	−	−	−	−	1–4: repeated appl. of (XII.9) and (XII.10)
{{1, 2, 3}, {4}}	→ {{1, 2}, {3, 4}}	+	+	+	−	1, 2: (XII.8); 3: C_4; 4: C_3
{{1, 2, 3}, {4}}	→ {{1, 2}, {3},{4}}	−	−	0	−	1, 2: (XII.10); 3: (XII.9), 4: C_1
{{1, 2, 3}, {4}}	→ {{1}, {2}, {3},{4}}	−	−	−	−	1, 2, 3: (XII.9), (XII.10), C_1; 4: C_1
{{1, 2}, {3, 4}}	→ {{1, 2}, {3},{4}}	−	−	−	−	1, 2: C_1; 3, 4: (XII.10)
{{1, 2}, {3, 4}}	→ {{1}, {2}, {3},{4}}	−	−	−	−	1–4: (XII.10), C_1
{{1, 2}, {3}, {4}}	→ {{1}, {2}, {3},{4}}	−	−	−	−	1, 2: (XII.10); 3, 4: C_1

Five countries
There are seven permutations to be considered: {{1, 2, 3, 4, 5}}, {{1, 2, 3, 4},{5}}, {{1, 2, 3}, {4, 5}}, {{1, 2, 3},{4},{5}}, {{1, 2},{3, 4},{5}}, {{1, 2},{3},{4},{5}} and {{1},{2},{3},{4},{5}} (Table XII.3).

R_1: Payoffs to 1 and 2 are the same as to 3 and 4 one row above. Payoffs to 3 and 4 are the same as to 1, 2 and 3 one row above. The implication for 5 follows from: $\pi_5^*(\{\{1, 2, 3, 4\},\{5\}\}) > \pi_5^*(\{\{1, 2\},\{3, 4\},\{5\}\}) = \pi_5^*(\{\{1, 2\},\{3, 4, 5\}\})$ where the inequality sign follows from C_1 and the equality sign from (XII.9).

R_2: Payoffs to 1, 2, 4 are the same as one row above. Payoff to 5 is worse than one row above by C_2. The implication for 3 follows from $\pi_3^*(\{\{1, 2\},\{3\},\{4, 5\}\}) = \pi_3^*(\{\{1, 2, 3\},\{4, 5\}\}) < \pi_3^*(\{\{1, 2, 3, 4\},\{5\}\})$ where the equality sign follows from (XII.9) and the inequality sign from (XII.8).

R_3: The coalition structure is the same as one row above. Payoffs to 1 and 2 are the same as to 3 one row above. Payoff to 3 is the same as one row above. Countries 4 and 5 are worse off than one row above by C_2.

The permutations {{1, 2, 3},{4, 5}}, {{1, 2, 3},{4},{5}}, {{1, 2},{3, 4},{5}}, {{1, 2},{3},{4},{5}} and {{1},{2},{3},{4},{5}} are stand-alone stable.

XII.3 Appendix 3

In this appendix we derive the NE, CPNE and SNE in the open-membership game. Each country makes an announcement (number) m_i (m is the collection of announcements) and countries with the same announcement form a coalition.

***N*=3**
$m=(1, 1, 1) \Rightarrow \{\{1, 2, 3\}\}$: NE. $m=(1, 1, 2) \Rightarrow \{\{1, 2\},\{3\}\}$: NE since 3 is indifferent to announcing $m=1$. $m=\{1, 2, 3\}$: no NE since by announcing $m_1=2$, 1 can improve upon its payoff.

Both NE – {{1, 2, 3}} and {{1, 2},{3}} – are Pareto-efficient and are therefore an SNE. From $C^{SNE} \subseteq C^{CPNE} \subseteq C^{NE}$ it follows that these two SNE coalition structures are also a CPNE.

***N*=4**
$m=(1, 1, 1, 1) \Rightarrow \{\{1, 2, 3, 4\}\}$: no NE since it is not stand-alone stable.
$m=(1, 1, 1, 2) \Rightarrow \{\{1, 2, 3\},\{4\}\}$: no NE since a country i, $i \in \{1, 2, 3\}$,

*Table XII.3 Incentive profile of countries (*N = *5)*

Change of permutation		1	2	3	4	5	Reason
{{1, 2, 3, 4, 5}}	→ {{1, 2, 3, 4},{5}}	−	−	−	−	+	1–4: (XII.10); 5: (XII.9)
{{1, 2, 3, 4, 5}}	→ {{1, 2, 3},{4, 5}}	−	−	−	−	−	1, 2, 3: (XII.10); 4, 5: (XII.9)
{{1, 2, 3, 4, 5}}	→ {{1, 2, 3},{4},{5}}	−	−	−	−	−	1, 2, 3: C_1 ({4}∪{5}) and (XII.10) ({1, 2, 3}∪{4, 5}); 4, 5: twice appl. of (XII.9)
{{1, 2, 3, 4, 5}}	→ {{1, 2},{3, 4},{5}}	−	−	−	−	−	1–4: (XII.10), ({1, 2}∪{3, 4} and ({1, 2, 3, 4}∪{5}); 5: {3, 4}∪{5} ((XII.9) and {1, 2}∪{3, 4, 5} ((XII.10))
{{1, 2, 3, 4, 5}}	→ {{1, 2},{3},{4},{5}}	−	−	−	−	−	1, 2: C_1 ({3}∪{4}∪{5}) and (XII.9) ({1, 2}∪{3, 4, 5}); 3–5: (XII.10) and C_1 ({3}∪{4}), (XII.9) and (XII.10) ({3, 4}∪{5}), (XII.10) ({1, 2}∪{3, 4, 5})
{{1, 2, 3, 4, 5}}	→ {{1},{2},{3},{4},{5}}	−	−	−	−	−	1–5: repeated appl. of (XII.9), (XII.10) and C_1
{{1, 2, 3, 4}, {5}}	→ {{1, 2, 3},{4, 5}}	+	+	+	+	−	1–3: (XII.8); 4: C_4; 5: C_3
{{1, 2, 3, 4},{5}}	→ {{1, 2},{3, 4, 5}}	+	+	+	+	−	R_1
{{1, 2, 3, 4},{5}}	→ {{1, 2, 3},{4},{5}}	−	−	−	+	−	1–3: (XII.10); 4: (XII.9); 5: C_1
{{1, 2, 3, 4},{5}}	→ {{1, 2},{3,4},{5}}	−	−	−	−	−	1–4: (XII.10); 5: C_1
{{1, 2, 3, 4},{5}}	→ {{1, 2},{3},{4, 5}}	−	−	−	−	−	R_2

{{1, 2, 3, 4},{5}}	→ {{1, 2},{3},{4},{5}}	−	−	−	−	−	1,2: C_1 ({3}∪{4}) and (XII.10) ({1, 2}∪{3, 4}); 3, 4: (XII.10) ({3}∪{4} and ({1, 2}∪{3, 4}); 5: C_1
{{1, 2, 3, 4},{5}}	→ {{1},{2},{3},{4},{5}}	−	−	−	−	−	1–4: repeated appl. of (XII.9) and (XII.10); 5: C_1
{{1, 2, 3}, {4, 5}}	→ {{1, 2, 3},{4},{5}}	−	−	−	−	−	1, 2, 3: C_1; 4, 5: (XII.10)
{{1, 2, 3}, {4, 5}}	→ {{1, 2},{3},{4, 5}}	−	−	0	−	−	1, 2: (XII.10); 3: (XII.9); 4, 5: C_1
{{1, 2, 3}, {4, 5}}	→ {{1}, {2}, {3},{4, 5}}	−	−	−	−	−	1–3: (XII.10) and C_1 ({1}∪{2}), (XII.9) and (XII.10) ({1, 2}∪{3}); 4, 5: C_1
{{1, 2, 3}, {4, 5}}	→ {{1, 2},{3},{4},{5}}	−	−	−	−	−	1–5: (XII.9),(XII.10) and C_1
{{1, 2, 3}, {4, 5}}	→ {{1},{2},{3},{4},{5}}	−	−	−	−	−	1–5: (XII.9),(XII.10) and C_1
{{1, 2, 3},{ 4},{5}}	→ {{1, 2},{3, 4},{5}}	+	+	+	−	−	1, 2: (XII.8); 3: C_4; 4: C_3; 5: by computation
{{1, 2, 3}, {4},{5}}	→ {{1, 2},{3},{4},{5}}	−	−	0	−	−	1, 2: (XII.10); 3: (XII.9); 4, 5: C_1
{{1, 2, 3}, {4},{5}}	→ {{1},{2},{3},{4, 5}}	0	0	0	−	−	R_3
{{1, 2},{3, 4},{5}}	→ {{1, 2},{3},{4, 5}}	0	0	+	0	−	1, 2, 4: same coal. structure 3, 5: C_2
{{1, 2},{3, 4},{5}}	→ {{1, 2},{3},{4},{5}}	−	−	−	−	−	1, 2, 5: C_1; 3, 4: (XII.10)
{{1, 2},{3, 4},{5}}	→ {{1}, {2},{3},{4},{5}}	−	−	−	−	−	1–4: repeated appl. of (XII.10) and C_1; 5: C_1
{{1, 2},{3},{4},{5}}	→ {{1},{2},{3},{4},{5}}	−	−	−	−	−	1, 2: (XII.10); 3–5: C_1

improves upon its payoff by announcing $m_i = 2$. $m = (1, 1, 2, 2) \Rightarrow \{\{1, 2\},\{3,4\}\}$: NE since announcing $m_i = 2$ or $m_i = 3$, $i \in \{1, 2\}$, implies a payoff loss to i. By symmetry, the same holds for a country j, $j \in \{3, 4\}$. $m = (1, 1, 2, 3)$: no NE since country 3 has an incentive to announce $m_3 = 3$. $m = (1, 2, 3, 4)$: no NE since country 1 has an incentive to announce $m_1 = 2$.

There is no SNE since the only NE, that is, $\{\{1, 2\},\{3, 4\}\}$, is not Pareto-efficient ($m_3 = m_4 = 1$ improves upon 3 and 4's payoff).

$\{\{1, 2\},\{3, 4\}\}$ is a CPNE since the only sensible deviation would be to form a grand coalition which is subject to a further deviation. (3 and 4 announce $m_3 = m_4 = 1$ but 4 is better off by announcing $m_4 = 2$.)

$N = 5$

$m = (1, 1, 1, 1, 1)$ and $m = (1, 1, 1, 1, 2)$ are not stand-alone stable and therefore no NE. $m = (1, 1, 1, 2, 2) \Rightarrow \{\{1, 2, 3\},\{4, 5\}\}$: NE since for 1, 2 and 3 announcing $m_i = 2$ or $m_i = 3$, $i \in \{1, 2, 3\}$, implies a payoff loss or leaves these countries indifferent; for 4 and 5 announcing $m_j = 1$ or $m_j = 3$ implies a payoff loss. $m = (1, 1, 1, 2, 3)$: no NE since $m_4 = 3$ improves upon 4's payoff. $m = (1, 1, 2, 2, 3) \Rightarrow \{\{1, 2\},\{3, 4\},\{5\}\}$: NE since $m_i = 2$ or $m_i = 4$, $i \in \{1, 2\}$ implies a payoff loss to i, $m_i = 3$ leaves i indifferent; by symmetry the same holds for $j \in \{3, 4\}$ and $m_5 = 1$ leaves 5 indifferent. $m = (1, 1, 2, 3, 4)$: no NE since $m_4 = 2$ improves upon 4's payoff. $m = (1, 2, 3, 4, 5)$: no NE since $m_2 = 1$ improves upon 2's payoff.

There is no SNE since both NE are Pareto-inefficient.

$\{\{1, 2, 3\},\{4, 5\}\}$ is a CPNE since forming smaller coalitions is not profitable. 1, 2 and 3 are indifferent to joining $\{4, 5\}$. A member of $\{4, 5\}$ has no incentive to join $\{1, 2, 3\}$. If both coalitions merge to form the coalition, further deviations will occur. $\{\{1, 2\},\{3, 4\},\{5\}\}$ is no CPNE since $\{3, 4\}$ would like to merge with $\{5\}$, which is a CPNE as laid out above.

XII.4 Appendix 4

In this appendix we derive the NE, CPNE and SNE in the exclusive membership Γ game. Each country makes an announcement regarding a coalition. A coalition only forms by unanimous agreement. If one country deviates, the coalition breaks apart.

Since a coalition structure can emerge from different proposals, we shall only investigate below whether an equilibrium combination of announcements exists leading to a particular coalition structure.

$N = 3$

$\{\{1, 2, 3\}\}$: NE since a deviation leads to $\{\{1\},\{2\},\{3\}\}$ and leaves all countries worse off. $\{\{1, 2\},\{3\}\}$: NE since if 1 and 2 make different proposals

this leads to {{1},{2},{3}} and leaves both countries worse off; 3 has no incentive to make a different proposal. {1, 2, 3}: NE by definition: if each country proposes a coalition by itself, no country can form another coalition by the unanimity rule.

The NE coalition structures {{1, 2, 3}} and {{1, 2},{3}} are SNE since they are Pareto-efficient.

From $C^{SNE} \subseteq C^{CPNE} \subseteq C^{NE}$ it follows that these two SNE coalition structures are also a CPNE. However, {1, 2, 3} is not a CPNE since it is Pareto-dominated by the other two CPNE.

$N=4$

It is easily checked that all permutations constitute an NE (see the discussion of $N=3$). That is, {{1, 2, 3, 4}}, {{1, 2, 3},{4}}, {{1, 2},{3, 4}}, {{1, 2},{3},{4}}, {{1},{2},{3},{4}} are all NE. Only {{1, 2, 3, 4}}, {{1, 2, 3},{4}} are SNE since they are efficient NE.

These coalitions structures are therefore also a CPNE. Since all other permutations are Pareto-dominated by these coalition structures no other CPNE exists.

$N=5$

It is easily checked that all permutations constitute an NE. That is, {{1, 2, 3, 4, 5}}, {{1, 2, 3, 4},{5}}, {{1, 2, 3},{4, 5}}, {{1, 2, 3},{4},{5}}, {{1, 2},{3, 4},{5}}, {{1, 2},{3},{4},{5}}, {{1},{2},{3},{4},{5}} are an NE.

Only {{1, 2, 3, 4, 5}}, {{1, 2, 3, 4},{5}} are SNE. All other NE are Pareto-dominated by these two SNE.

Accordingly (see the discussion of $N=4$ above), these are also the only CPNE.

XII.5 Appendix 5

In this appendix we derive the NE, CPNE and SNE in the exclusive membership Δ game. Each country makes an announcement regarding a coalition in which it is a member. A coalition forms of those countries which have made the same proposal even though not all countries which have proposed it may eventually join the coalition. If one country deviates, the remaining coalition members remain in the coalition.

Since a coalition structure can emerge from different proposals, we shall only investigate below whether an equilibrium combination of proposals exists leading to a particular coalition structure.

$N=3$

{{1, 2, 3}}: NE since a deviation by 3 leads to {{1, 2},{3}} and leaves 3 indifferent. {{1, 2},{3}}: NE since 1 and 2 have no incentive to be in

singleton coalitions and 3 has no incentive to merge with {1, 2}. {1, 2, 3}: NE by definition: if each country proposes a coalition by itself, then any other proposal leaves the coalition structure unaffected.

The NE coalition structures {{1, 2, 3}} and {{1, 2},{3}} are an SNE since they are Pareto-efficient.

From $C^{SNE} \subseteq C^{CPNE} \subseteq C^{NE}$ it follows that these two SNE coalition structures are also a CPNE. {1, 2, 3} is not a CPNE since it is Pareto-dominated by the other two CPNE.

***N*=4**

It is easily checked that all permutations ({{1, 2, 3},{4}}, {{1, 2},{3, 4}}, {{1, 2},{3},{4}}, {{1},{2},{3},{4}}, are an NE except {{1, 2, 3, 4}} which is not stand-alone stable.

Only {{1, 2, 3},{4}} is an SNE since all other coalition structures are Pareto-dominated by the grand coalition.

Thus {{1, 2, 3},{4}} is a CPNE too. Moreover, {{1, 2},{3, 4}} is a CPNE. No country has an incentive to propose a coalition of three countries since the proposed joining country and the coalition accepting the new member would be worse off. The grand coalition, though it would raise all countries' payoff, is subject to a further deviation. Any smaller coalitions would imply a payoff loss. {{1, 2},{3},{4}} and {{1},{2},{3},{4}} are no CPNE since these coalition structures are Pareto-dominated by the CPNE {{1, 2},{3, 4}}.

***N*=5**

It is easily checked that all permutations ({{1, 2, 3},{4, 5}}, {{1, 2, 3},{4},{5}}, {{1, 2},{3, 4},{5}}, {{1, 2},{3},{4},{5}}, {{1},{2},{3},{4},{5}}) constitute an NE except {{1, 2, 3, 4, 5}}, {{1, 2, 3, 4},{5}} which are not stand-alone stable. For instance, {{1, 2, 3},{4},{5}} is an NE since 1, 2 and 3 are indifferent to leaving the coalition and to operating as singletons. A merger of 1, 2 or 3 with countries 4 or 5, as well as a merger of 4 and 5 requires a simultaneous change in at least two proposals which is not considered by the NE concept.

There is no SNE since all NE are Pareto-dominated by the grand coalition.

{{1, 2, 3},{4, 5}} is CPNE since larger coalitions are not stable and there is no incentive to form smaller coalitions. {{1, 2, 3},{4},{5}} is no CPNE since it is Pareto-dominated by the CPNE {{1, 2, 3},{4, 5}}. {{1, 2},{3, 4},{5}} is CPNE since country 5 is indifferent to accession of a coalition of two countries. Any coalition of four or five countries is not stable. {{1, 2},{3},{4},{5}} and {{1},{2},{3},{4},{5}} are not a CPNE since they are Pareto-dominated by the CPNE {{1, 2},{3, 4},{5}}.

XII.6 Appendix 6

In this appendix the core-stable, α- and β-stable coalition structures are derived. For the definition of the concept, see Sub-section 15.4.3.

***N*=3**

{{1, 2, 3}}: CS since no player is better off in another coalition. {{1, 2},{3}}: CS since 1 and 2 can only jointly deviate with 3. However, 3 is indifferent to such an offer. {{1},{2},{3}}: no CS since all players prefer a grand coalition.

{{1, 2, 3}} and {{1, 2},{3}} are AS (BS) by $C_c \subseteq C_\beta = C_\alpha$. {{1},{2},{3}}: no AS (BS) since all players prefer a grand coalition.

***N*=4**

{{1, 2, 3, 4}}: no CS since 4 prefers {{1, 2, 3},{4}}. {{1, 2, 3},{4}}: no CS since 1 and 2 prefer {{1, 2},{3, 4}}. {{1, 2},{3, 4}}: no CS since all prefer {{1, 2, 3, 4}}. {{1, 2},{3},{4}}, {{1},{2},{3},{4}}: no CS since all players prefer {{1, 2, 3, 4}}.

{{1, 2, 3, 4}}: AS (BS) since any deviation may lead eventually to {{1}, {2}, {3}, {4}} which is Pareto-dominated. {{1, 2, 3}, {4}}: AS (BS). Though 1 and 2 prefer {{1, 2},{3, 4}}, they have to reckon with {{1, 2},{3},{4}} which is a Pareto-inferior coalition structure for them. {{1, 2},{3, 4}}, {{1, 2},{3},{4}} and {{1},{2},{3},{4}}: no AS (BS) since the grand coalition is a Pareto-superior coalition structure.

***N*=5**

{{1, 2, 3, 4, 5}} and {{1, 2, 3, 4},{5}} are not stand-alone stable and hence no CS. All other permutations ({{1, 2, 3}, {4, 5}}, {{1, 2, 3}, {4}, {5}}, {{1, 2}, {3, 4}, {5}}, {{1, 2}, {3}, {4}, {5}} and {{1},{2},{3},{4},{5}}) are Pareto-dominated by the grand coalition and therefore no CS. Similar arguments as laid out for $N=4$ establish that {{1, 2, 3, 4, 5}} and {{1, 2, 3, 4},{5}} are AS (BS).

XII.7 Appendix 7

In this appendix the coalition structures are derived by applying Chwe's far-sighted coalitional stability concept. For the definition of the largest consistent set, C^{LCS}, see Sub-section 15.4.4. An inequality sign below implies strict Pareto-dominance.

***N*=3**

$\{\{1\},\{2\},\{3\}\} < \{\{1, 2\},\{3\}\} \Leftrightarrow \{\{1, 2, 3\}\}$. Hence, $C^{LCS} = \{\{1, 2, 3\}\},\{\{1, 2\},\{3\}\}$.

***N*=4**

$\{\{1, 2, 3, 4\}\} < \{\{1, 2, 3\},\{4\}\}$. $\{\{1, 2, 3\},\{4\}\}$ is not Pareto-dominated by any other coalition structure and is therefore stable. $\{\{1, 2\},\{3, 4\}\}$ is directly dominated by the grand coalition and hence indirectly by $\{\{1, 2, 3\},\{4\}\}$. Since, however, $\{\{1, 2, 3\},\{4\}\}$ is not preferable by all members of the original coalition, a deviation is not deemed feasible and $\{\{1, 2\},\{3, 4\}\}$ is stable. $\{\{1, 2\},\{3\},\{4\}\}$ and $\{\{1\},\{2\},\{3\},\{4\}\}$ are not stable, since they are Pareto-dominated by $\{\{1, 2\},\{3, 4\}\}$. Hence, $C^{LCS} = \{\{\{1, 2, 3\},\{4\}\}, \{\{1, 2\},\{3, 4\}\}\}$.

***N*=5**

A deviation from $\{\{1, 2, 3, 4, 5\}\}$ to $\{\{1, 2, 3, 4\},\{5\}\}$ is subject to a further deviation to $\{\{1, 2, 3\},\{4\},\{5\}\}$, which is not in the interest of all original coalition members. A deviation from $\{\{1, 2, 3, 4, 5\}\}$ to $\{\{1, 2, 3\}, \{4, 5\}\}$ is also not in the interest of any country. Smaller coalitions are Pareto-dominated by $\{\{1, 2, 3, 4, 5\}\}$.

$\{\{1, 2, 3, 4\},\{5\}\} \notin C^{LCS}$ since $\pi_4^*(\{\{1, 2, 3, 4\},\{5\}\}) < \pi_4^*(\{\{1, 2, 3\},\{4\},\{5\}\}) < \pi_4^*(\{\{1, 2, 3\},\{4, 5\}\}) < \pi_4^*(\{\{1, 2, 3, 4, 5\}\})$. $\{\{1, 2, 3\},\{4, 5\}\}$ and $\{\{1, 2, 3, 4, 5\}\}$ deliver a higher payoff to country 4 than $\{\{1, 2, 3, 4\},\{5\}\}$. That is, country 4 leaves coalition $\{\{1, 2, 3, 4\},\{5\}\}$ since it expects that the grand coalition will eventually form. All other coalition structures are Pareto-dominated by the grand coalition and therefore do not belong to the largest consistent set. Hence, the $C^{LCS} = \{\{1, 2, 3, 4, 5\}\}$.

XII.8 Appendix 8

In this appendix the coalition structure is derived by applying the concept of equilibrium binding agreements. For the definition see Sub-section 15.5.1.

***N*=3**

$\{\{1\},\{2\},\{3\}\}$ is EBA by definition. $\{\{1, 2\},\{3\}\}$ is EBA since it is not blocked by $\{\{1\},\{2\},\{3\}\}$. $\{\{1, 2, 3\}\}$ is also EBA since it is not blocked by $\{\{1, 2\},\{3\}\}$ and $\{\{1\},\{2\},\{3\}\}$. By applying $\text{Eff}(C^{EBA})$ (extension to the original concept), $\text{Eff}(C^{EBA}) = \{\{\{1, 2\},\{3\}\}, \{\{1,2,3\}\}\}$.

***N*=4**

$\{\{1\},\{2\},\{3\},\{4\}\}$ is stable by definition. $\{\{1, 2\},\{3\},\{4\}\}$ and $\{\{1, 2\},\{3, 4\}\}$ are stable since a further partition implies a payoff loss to all countries. $\{\{1, 2, 3\},\{4\}\}$ is stable since country 3 is indifferent to forming a singleton coalition. $\{\{1, 2, 3, 4\}\}$ is not stable since it is blocked by $\{\{1, 2, 3\},\{4\}\}$. Hence, $\text{Eff}(C^{EBA}) = \{\{\{1, 2, 3\},\{4\}\}, \{\{1, 2\},\{3, 4\}\}\}$.

***N*=5**

{{1},{2},{3},{4},{5}} is stable by definition. {{1, 2},{3},{4},{5}}, {{1, 2},{3, 4},{5}}, {{1, 2, 3},{4},{5}}, {{1, 2, 3},{4, 5}} are stable since a partition does not pay a country. {{1, 2, 3, 4},{5}} is not stable since country 4 has an incentive to form a singleton. {{1, 2, 3, 4, 5}} is stable since a deviation from {{1, 2, 3, 4, 5}} to {{1, 2, 3, 4},{5}} leads eventually to {{1, 2, 3},{4},{5}} which is Pareto-dominated by the grand coalition. A deviation by two players to {{1, 2, 3},{4, 5}} is also a Pareto-dominated partition. Any smaller partitions are also Pareto-dominated and hence the grand coalition is an efficient equilibrium binding agreement. Eff(C^{EBA})={{1, 2, 3, 4, 5}}.

XII.9 Appendix 9

In this appendix the coalition structure under the sequential coalition formation process is derived. We distinguish between Bloch's original concept, Ray and Vohra's extension which assumes that if a country is indifferent between two coalitions it accepts the larger coalition and Finus's extension of trembling-hand sequential equilibrium (see Sub-section 15.5.2 for details).

***N*=3**

{{1},{2},{3}} is no sequential equilibria (SE) since if country 1 proposes itself, country 2 will propose {2, 3}. {{1},{2, 3}} is an SE by the previous argument. {{1, 2, 3}} is also SE since no country has an incentive to reject the proposal.

Since country 1 is indifferent between {{1, 2, 3}} and {{1},{2, 3}} it proposes {{1, 2, 3}} according to Ray and Vohra's extension.

Since country 1 reckons that its proposal may be rejected by coincidence and it may end up in {{1, 3},{2}} (since 2 proposes {2}), it proposes itself and {{1},{2, 3}} is the equilibrium according to Finus's extension.

***N*=4**

From the incentive profile it is known that if country 3 is in the position to make a proposal (that is, {1, 2} has formed), it will suggest {3, 4}. If it is country 2's turn to make a proposal (that is, {1} has formed), it can choose between {{1},{2, 3, 4}}, {{1},{2},{3, 4}} and {{1},{2, 3},{4}}. Since $\pi_2^*(\{\{1\},\{2,3\},\{4\}\}) < \pi_2^*(\{\{1\},\{2\},\{3,4\}\}) = \pi_2^*(\{\{1\},\{2,3,4\}\})$, country 2 will propose either {2} or {2, 3, 4}. Thus, country 1 has the option between (1) {1, 2, 3, 4}⇒{1, 2, 3, 4}; (2) {1, 2, 3}⇒{{1, 2, 3},{4}}; (3) {1, 2}⇒{{1, 2},{3, 4}}; (4) {1}⇒(a) {{1},{2, 3, 4}} or (b) {{1},{2},{3, 4}}. From country 1's perspective option 1 dominates option 2. Option 1 dominates

option 3. Option 1 is dominated by 4(a) but dominates 4(b). Hence, there are two sequential equilibria: {{1, 2, 3, 4}} which is supported by country 2 choosing {2} if it is its turn to make a proposal and {{1},{2, 3, 4}} which is supported if 2 proposes {2, 3, 4} if it is its turn to make a proposal.

For Ray and Vohra's extension the same arguments as developed above apply except that option 4(b) is dropped. Hence, country 1 realizes that option 1 dominates options 2 and 3 but is dominated by option 4. Therefore, country 1 proposes {1} and {{1},{2, 3, 4}} forms in equilibrium.

According to Finus's extension, country 1 proposes the grand coalition since it fears that if it proposes {1} it may end up {{1},{2},{3, 4}} instead of the expected {{1},{2, 3, 4}}.

***N*=5**

A similar procedure as laid out for $N=4$ establishes that {{1, 2, 3, 4},{5}} and {{1, 2, 3, 4, 5}} are sequential equilibria. Ray and Vohra's extension leads to the grand coalition and Finus's extension to {{1, 2, 3, 4},{5}}.

NOTES

1. A proof for asymmetric payoffs proceeds along exactly the same lines. Since the notation is rather messy and confirms Proposition 3.1, the proof is not reproduced here.
2. Multiple deviations by more than one player are ignored and not punished.
3. We assume that players also receive payoffs at time T.
4. Of course, for the entire game $2u_i(a_{i1},a_{i2})-(1-\delta_i)2u_i(c_{i1},c_{i2})\geq 0 \forall\ i \in \mathrm{I}$ must hold, but upon dividing through by 2 this is equivalent to (V.3).
5. More precisely, we subtract the LHS term of (V.2)+(V.23) from (V.3).
6. ϕ_i will be an increasing function in any case since $-1<k_i<0$ implies that if the Stackelberg leader increases emissions, e_i, global emissions, Σe_k, will increase, though e_j diminishes.
7. This last relation can also be used to prove that $e_i^{\mathrm{ST}}<e_i^N$ is irrational since this would imply $\Sigma e_k^{\mathrm{ST}}\geq\Sigma e_k^N$, $e_j^{\mathrm{ST}}>e_j^N$ and hence $\pi_i^{\mathrm{ST}}<\pi_i^N$. By choosing $e_i^{\mathrm{ST}}=e_i^N$ the leader could at least guarantee him or herself $\pi_i^{\mathrm{ST}}=\pi_i^N$.
8. $\Sigma e_i^N<\Sigma e_i^{\mathrm{S}}$ can also be shown by proceeding as in Section 9.6. Assume $\Sigma e_i^N>\Sigma e_i^{\mathrm{S}}$ is true, then marginal damage in the NE would be higher than in the social optimum since $\phi_i'(\Sigma e_i^N)>\phi_i'(\Sigma e_i^{\mathrm{S}})+\phi_j'(\Sigma e_i^{\mathrm{S}})\ \forall\ i \in \mathrm{I}$ due to $\phi_j'<0$ ($\phi_j'=\partial\phi_j/\partial e_i$). Then the FOC in the Nash equilibrium would require higher marginal benefits and hence $e_i^N<e_i^{\mathrm{S}}\ \forall\ i \in \mathrm{I}$ must be true, which contradicts the initial assumption $\Sigma e_i^N>\Sigma e_i^{\mathrm{S}}$. The extension to the cases of corner equilibria would confirm this result.
9. It is easy to see that SOC in the NE are satisfied. Since $\partial^2\phi_j/\partial e_i^2>0$ by assumption, this is also true in the social optimum.
10. $(\partial\phi_i/\partial e_i)-(\partial\phi_i/\partial e_j)(\partial e_j/\partial e_i)>0$ is a sufficient condition to ensure that the FOC can be satisfied. Since Proposition 10.2 assumes $|\partial\phi_i/\partial e_i|>|\partial\phi_i/\partial e_j|$ and $\partial e_j/\partial e_i<1$ has been established above, this condition is satisfied. With respect to the SOC similar arguments to those developed in Appendix VII.1 apply. That is, the SOC may not be generally satisfied and must therefore be assumed to hold.
11. From country j's perspective (VII.7) implies

$$-\frac{\partial\beta_j}{\partial e_j}\cdot e_j^{\mathrm{I}}+\frac{\partial\phi_j}{\partial e_j}\cdot\left(\frac{1}{\mu}\cdot e_i^{\mathrm{I}}+e_j^{\mathrm{I}}\right)=0$$

which leads to the same conclusions as above.

12. It is only sensible to consider the possibility that the bottleneck country i (assuming a non-biased proposal) makes a strategic proposal. If country j made a strategic proposal this would imply two Stackelberg followers for which no equilibrium exists. To see this, note that for a proposal t_i^{str} to be effective $\beta_j'(e_j^T(t_i^{str})) < \phi_j'(\Sigma e_k^T(t_i^{str}))$ and $\beta_i'(e_i^T(t_i^{str})) > \phi_i'(\Sigma e_k^T(t_i^{str}))$ must hold before adjustment is conducted so that the neighboring country j adjusts and country i is in the lead. If country i is already the bottleneck without strategic considerations $\beta_j'(e_j^T(t_i)) < \phi_j'(\Sigma e_k^T(t_i))$ holds for country j before it conducts adjustment. Hence, any $t_j^{str} < t_i$ would lead to $\beta_j'(e_j^T(t_j^{str})) \ll \phi_j'(\Sigma e_k^T(t_j^{str}))$ and if the proposal were to be effective $\beta_i'(e_i^T(t_j^{str})) < \phi_i'(\Sigma e_k^T(t_j^{str}))$ would hold. This would basically imply two Stackelberg followers, which is not possible.
13. All proofs related to Section 13.3 assume an interior Nash equilibrium and social optimum.
14. Comparing the FOC in the PANE with that in the social optimum, assuming marginal damages in country i to be given by d_i and the number of countries in I^J to be given by N^*, we find: $\beta_i'(\tilde{e}_i^J) = \Sigma_{k\in I^J}\phi_k'(\tilde{e}) = \Sigma_{i=1}^{N^*} d_i \leq \Sigma_{i=1}^{N} d_i = \Sigma_{k\in I}\phi_k'(e^S) = \beta_i'(e_i^S)$ and hence $e_i^S \leq \tilde{e}_i^J \, i \in I^J$ follows from $\beta_i'' < 0$. Comparing the FOC in the PANE with that in the Nash equilibrium, we find: $\beta_i'(\tilde{e}_i^J) = \Sigma_{k\in I^J}\phi_k'(\tilde{e}) = \Sigma_{i=1}^{N^*} d_i \geq d_i = \phi_k'(e^N) = \beta_i'(e_i^N)$ and hence $e_i^N \geq \tilde{e}_i^J \forall \, i \in I^J$ by $\beta_i'' < 0$.
15. To see this, note that from the FOC in the social optimum we have $\Sigma_{i\in I}\phi_i'(e^S) = \beta_i'(e_i^S)$. Rearranging terms gives $\beta_i'(e_i^S)(\Sigma_{i\in I}\tilde{e}_i - \Sigma_{i\in I}e_i^S) \geq \Sigma_{i\in I}(\beta_i(e_i^S) - \beta_i(\tilde{e}_i))$ which is satisfied since $\beta_i'(e_i^S) \geq [\beta_i(\tilde{e}_i^J) - \beta_i(e_i^S)]/[\tilde{e}_i - e_i^S]$ by the concavity of the benefit function (and $\tilde{e}_i \geq e_i^S$ $\forall \, i \in I$).
16. For the appropriateness of this assumption in the case of $\delta \longrightarrow 1$, see Chapter 12.

References

Abreu, D. (1986), 'Extremal equilibrium of oligopolistic supergames', *Journal of Economic Theory*, **39**, 191–225.

Abreu, D. (1988), 'On the theory of infinitely repeated games with discounting', *Econometrica*, **56**, 383–96.

Abreu, D., D. Pearce and E. Stacchetti (1986), 'Optimal cartel equilibrium with imperfect monitoring', *Journal of Economic Theory*, **39**, 251–69.

Abreu, D., D. Pearce and E. Stacchetti (1993), 'Renegotiation and symmetry in repeated games', *Journal of Economic Theory*, **60**, 217–40.

Alho, K. (1992), 'Bilateral transfers and lending in international environmental cooperation', *Environmental and Resource Economics*, **2**, 201–20.

Andersson, T. (1991), 'Government failure: the cause of global environmental mismanagement', *Ecological Economics*, **4**, 215–36.

Arnold, V. (1984), 'Umweltschutz als internationales öffentliches Gut: Komparative Kostenvorteile und Verhandlungsgewinne', *Zeitschrift für Wirtschafts- und Sozialwissenschaften*, **104**, 111–29.

Arnold, V. (1992), *Theorie der Kollektivgüter*, Munich: Franz Vahlen.

Aronson, A.L. (1993), 'From "cooperator's loss" to cooperative gain: negotiating greenhouse gas abatement', *Yale Law Journal*, **102**, 2143–74.

Asheim, G.B. (1991), 'Extending renegotiation-proofness to infinite games', *Games and Economic Behavior*, **3**, 278–94.

Aumann, R.J. (1959), 'Acceptable points in general cooperative *N*-person games', in A.N. Tucker and R.D. Luce (eds), *Contributions to the Theory of Games, Volume IV (Annals of Mathematics Studies 40)*, Princeton, NJ: Princeton University Press, pp. 287–324.

Aumann, R. and L. Shapley (1976), 'Long-term competition: a game-theoretic analysis', mimeo.

Ausubel, J.H. and D.G. Victor (1992), 'Verification of international environmental agreements', *Annual Review of Energy and Environment*, **17**, 1–43.

Avenhaus, R. (1992), 'Monitoring the emissions of pollutants by means of the inspector leadership method', in R. Pethig (ed.), *Conflicts and Cooperation in Managing Environmental Resources: Microeconomic Studies*, Berlin: Springer, pp. 241–69.

Axelrod, R. (1984), *The Evolution of Cooperation*, New York: Basic Books.

Bac, M. (1996), 'Incomplete information and incentives to free-ride on

international environmental resources', *Journal of Environmental Economics and Management*, **30**, 301–15.
Barratt-Brown, E.P. (1991), 'Building a monitoring and compliance regime under the Montreal Protocol', *Yale Journal of International Law*, **16**, 519–70.
Barrett, S. (1990), 'The problem of global environmental protection', *Oxford Review of Economic Policy*, **6**, 68–79.
Barrett, S. (1991a), 'Economic instruments for climate change policy', in OECD, *Responding to Climate Change: Selected Economic Issues*, Paris: OECD, pp. 53–108.
Barrett, S. (1991b), 'Economic analysis of international environmental agreements: lessons for a global warming treaty', in OECD, *Responding to Climate Change: Selected Economic Issues*, Paris: OECD, pp. 111–49.
Barrett, S. (1991c), 'The paradox of international environmental agreements', London: London Business School, mimeo.
Barrett, S. (1992a), 'Alternative instruments for negotiating a global warming convention', in OECD (ed.), *Convention on Climate Change: Economic Aspects of Negotiations*, Paris: OECD, pp. 11–48.
Barrett, S. (1992b), 'Side payments in a global warming convention', in OECD (ed.), *Convention on Climate Change: Economic Aspects of Negotiations*, Paris: OECD, pp. 49–71.
Barrett, S. (1992c), 'Free-rider deterrence in a global warming convention', in OECD (ed.), *Convention on Climate Change: Economic Aspects of Negotiations*, Paris: OECD, pp. 73–97.
Barrett, S. (1992d), 'International environmental agreements as games', in R. Pethig (ed.), *Conflicts and Cooperation in Managing Environmental Resources: Microeconomic Studies*, Berlin: Springer, pp. 11–37.
Barrett, S. (1992e), 'Cooperation and competition in international environmental protection', invited paper, Economics of the Environment Session, International Economic Association Meeting, 24–28 August, Moscow.
Barrett, S. (1994a), 'The biodiversity supergame', *Environmental and Resource Economics*, **4**, 111–22.
Barrett, S. (1994b), 'Self-enforcing international environmental agreements', *Oxford Economic Papers*, **46**, 804–78.
Barrett, S. (1997a), 'Toward a theory of international environmental cooperation', in C. Carraro and D. Siniscalco (eds), *New Directions in the Economic Theory of the Environment*, Cambridge: Cambridge University Press, pp. 239–80.
Barrett, S. (1997b), 'Heterogeneous international agreements', in C. Carraro (ed.), *International Environmental Negotiations: Strategic Policy Issues*, Cheltenham, UK: Edward Elgar, pp. 9–25.

Barrett, S. (1997c), 'The strategy of trade sanctions in international environmental agreements', *Resource and Energy Economics*, **19**, 345–61.
Basar, T. and G. Olsder (1982), *Dynamic Non-cooperative Game Theory*, London: Academic Press.
Bauer, A. (1992), 'International cooperation over greenhouse gas abatement', Seminar für empirische Wirtschaftsforschung, University of Munich, mimeo.
Baumol, W.J. and W.E. Oates (1971), 'The use of standards and prices for protection of the environment', *Swedish Journal of Economics*, **73**, 42–54.
Baumol, W. and W.E. Oates (1990), *The Theory of Environmental Policy*, 2nd edn, Cambridge: Cambridge University Press.
Benedick, R.E. (1991), 'Protecting the ozone layer: new directions in diplomacy', in J.T. Mathews (ed.), *Preserving the Global Environment: The Challenge of Shared Leadership*, New York and London: W.W. Norton, pp. 112–53.
Benedick, R.E. and R. Pronove (1992), 'Atmosphere and outer space', in P.H. Sand (ed.), *The Effectiveness of International Environmental Agreements*, Cambridge: Grotius, pp. 123–48.
Benoît, J.-P. and V. Krishna (1985), 'Finitely repeated games', *Econometrica*, **53**, 890–904.
Benoît, J.-P. and V. Krishna (1993), 'Renegotiation in finitely repeated games', *Econometrica*, **61**, 303–23.
Bergesen, H.O. and G. Parmann (eds) (1997), *Green Globe Yearbook 1997*, New York: Oxford University Press.
Bergin, J. and W.B. MacLeod (1993), 'Efficiency and renegotiation in repeated games', *Journal of Economic Theory*, **61**, 42–73.
Bernheim, D. and D. Ray (1985), 'Pareto-perfect Nash equilibria', Stanford University, mimeo.
Bernheim, D. and D. Ray (1989), 'Collective dynamic consistency in repeated games', *Games and Economic Behavior*, **1**, 295–326.
Bernheim, D. and M.D. Whinston (1987), 'Coalition-proof Nash equilibria. II: Applications', *Journal of Economic Theory*, **42**, 13–29.
Bernheim, D., B. Peleg and M.D. Whinston (1987), 'Coalition-proof Nash equilibria. I: Concepts', *Journal of Economic Theory*, **42**, 1–12.
Beyerlin, U. (1996), 'State community interests and institution-building in international environmental law', *Zeitschrift für ausländisches öffentliches Recht und Völkerrecht*, **56**, 602–27.
Binmore, K. (1990), *Fun and Games: A Text on Game Theory*, Lexington, VA: Heath.
Binmore, K. and P. Dasgupta (eds) (1987), *The Economics of Bargaining*, Oxford and New York: Basil Blackwell.

Binmore, K., A. Rubinstein and A. Wolinsky (1986), 'The Nash bargaining solution in economic modelling', *RAND Journal of Economics*, **17**, 176–88.
Birchenhall, C. and P. Grout (1984), *Mathematics for Modern Economics*, New York: Barnes & Noble.
Black, J., M.D. Levi and D. de Meza (1992), 'Creating a good atmosphere: minimum participation for tackling the "greenhouse effect"', *Economica*, **60**, 281–93.
Blackhurst, R. and A. Subramanian (1992), 'Promoting multilateral cooperation on the environment', in K. Anderson and R. Blackhurst (eds), *The Greening of the World Trade Issue*, New York: Harvester Wheatsheaf, pp. 247–68.
Bloch, F. (1995), 'Endogenous structures of associations in oligopolies', *RAND Journal of Economics*, **26**, 537–56.
Bloch, F. (1996), 'Sequential formation of coalitions in games with externalities and fixed payoff division', *Games and Economic Behavior*, **14**, 90–123.
Bloch, F. (1997), 'Non-cooperative models of coalition formation in games with spillovers', in C. Carraro and D. Siniscalco (eds), *New Directions in the Economic Theory of the Environment*, Cambridge: Cambridge University Press, pp. 311–52.
Boadway, R. and N. Bruce (1993), *Welfare Economics*, Oxford and Cambridge, MA: Basil Blackwell.
Bohm, P. (1993), 'Incomplete international cooperation to reduce CO_2 emissions: alternative policies', *Journal of Environmental Economics and Management*, **24**, 258–71.
Bohm, P. (1994), 'Making carbon emission quota agreements more efficient: joint implementation versus quota tradability', in G. Klaassen and F.R. Førsund (eds), *Economic Instruments for Air Pollution Control*, Dordrecht: Kluwer, pp. 187–208.
Bohm, P. and B. Larsen (1993), 'Fairness in a tradeable-permit treaty for carbon emissions reductions in Europe and the former Soviet Union', *Environmental and Resource Economics*, **4**, 219–39.
Bothe, M. (1996), 'The evaluation of enforcement mechanisms in international environmental law', in R. Wolfrum (ed.), *Enforcing Environmental Standards: Economic Mechanisms as Viable Means?*, Berlin: Springer, pp. 13–38.
Botteon, M. and C. Carraro (1997), 'Burden-sharing and coalition stability in environmental negotiations with asymmetric countries', in C. Carraro (ed.), *International Environmental Negotiations: Strategic Policy Issues*, Cheltenham, UK: Edward Elgar, pp. 26–55.
Botteon, M. and C. Carraro (1998), 'Strategies for environmental negotiations: issue linkage with heterogeneous countries', in N. Hanley and H.

Folmer (eds), *Game Theory and the Global Environment*, Cheltenham, UK: Edward Elgar, pp. 180–200.
Brandenburger, A. and E. Dekel (1989), 'The role of common knowledge assumptions in game theory', in F. Hahn (ed.), *The Economics of Missing Markets, Information, and Games*, Oxford: Clarendon Press, pp. 46–61.
Brown Weiss, E. and H.K. Jacobson (1997), 'Compliance with international environmental accords', in M.H. Rolen, H. Sjöberg and U. Svedin (eds), *International Governance on Environmental Issues*, Dordrecht: Kluwer, pp. 78–110.
Brunner, R.D. (1991), 'Global climate change: defining the policy problem', *Policy Sciences*, **24**, 291–311.
Buchholz, W. and K.A. Konrad (1994), 'Global environmental problems and the strategic choice of technology', *Journal of Economics*, **60**, 299–321.
Burbidge, J., J.A. DePater, G.M. Myers and A. Sengupta (1997), 'A coalition-formation approach to equilibrium federations and trading blocs', *American Economic Review*, **87**, 940–56.
Burger, C. (1994), 'Spieltheoretische Analyse umweltrelevanten Verhaltens', in R. Bartel (ed.), *Einführung in die Umweltpolitik*, Vahlen, Munich: WiSo-Kurzlehrbücher, pp. 119–38.
Caldwell, L.K. (1984), *International Environmental Policy: Emergence and Dimensions*, Durham, NC: Duke University Press.
Canning, D. (1989), 'Bargaining theory', in F. Hahn (ed.), *The Economics of Missing Markets, Information, and Games*, Oxford: Clarendon Press, pp. 163–87.
Cansier, D. (1991), 'Bekämpfung des Treibhauseffektes aus ökonomischer Sicht', in M. Kloepfer (ed.), *Studien zum Umweltstaat*, Berlin: Springer.
Carraro, C. (1997), *The Structure of International Environmental Agreements*, Working Paper, Milan: Fondazione Eni Enrico Mattei.
Carraro, C. and F. Moriconi (1997), *International Games on Climate Change Control*, Working Paper, Milan: Fondazione Eni Enrico Mattei.
Carraro, C. and D. Siniscalco (1991), *Strategies for the International Protection of the Environment*, Working Paper, Milan: Fondazione Eni Enrico Mattei.
Carraro, C. and D. Siniscalco (1992), 'The international dimension of environmental policy', *European Economic Review*, **36**, 379–87.
Carraro, C. and D. Siniscalco (1993), 'Strategies for the international protection of the environment', *Journal of Public Economics*, **52**, 309–28.
Carraro, C. and D. Siniscalco (1997), 'R&D cooperation and the stability of international environmental agreements', in C. Carraro (ed.), *International Environmental Negotiations: Strategic Policy Issues*, Cheltenham, UK: Edward Elgar, pp. 71–96.

Carraro, C. and D. Siniscalco (1998), 'International environmental agreements: incentives and political economy', *European Economic Review*, **42**, 561–72.

Cesar, H. and A. de Zeeuw (1996), 'Issue linkage in global environmental problems', in A. Xepapadeas (ed.), *Economic Policy for the Environment and Natural Resources: Techniques for the Management and Control of Pollution*, Cheltenham, UK and Brookfield, USA: Edward Elgar, pp. 158–73.

Chander, P. and H. Tulkens (1991), *Strategically Stable Cost Sharing in an Economic–Ecological Negotiation Process*, CORE Discussion Paper no. 9135, Louvain: Center for Operations, Research and Econometrics, Université Catholique de Louvain.

Chander, P. and H. Tulkens (1992), 'Theoretical foundations of negotiations and cost sharing in transfrontier pollution problems', *European Economic Review*, **36**, 388–98.

Chander, P. and H. Tulkens (1995), 'A core-theoretic solution for the design of cooperative agreements on transfrontier pollution', *International Tax and Public Finance*, **2**, 279–93.

Chander, P. and H. Tulkens (1997), 'The core of an economy with multilateral environmental externalities', *International Journal of Game Theory*, **26**, 379–401.

Chapman, D. and T. Drennen (1990), 'Equity and effectiveness of possible CO_2 treaty proposals', *Contemporary Policy Issues*, **8**, 16–28.

Chatterjee, K., B. Dutta, D. Ray and K. Sengupta (1993), 'A noncooperative theory of coalitional bargaining', *Review of Economic Studies*, **60** (2), 463–77.

Chayes, A.H. and A. Chayes (1993), 'On compliance', *International Organization*, **47**, 175–205.

Chayes, A.H. and A. Chayes (1995), *The New Sovereignty*, Cambridge, MA and London: Harvard University Press.

Chen, Z. (1997), 'Negotiating an agreement on global warming: a theoretical analysis', *Journal of Environmental Economics and Management*, **32**, 170–88.

Chiang, A.C. (1984), *Fundamental Methods of Mathematical Economics*, Singapore: McGraw-Hill.

Chichilnisky, C., G. Heal and A. Vercelli (1998) (eds), *Sustainability: Dynamics and Uncertainty*, Dordrecht: Kluwer.

Chillemi, O. (1996), *International Environmental Agreements and Asymmetric Information*, Working Paper: Economics, Energy Environment, no. 69.96, Milan: Fondazione Eni Enrico Mattei.

Chwe, M.S.-Y. (1994), 'Farsighted coalitional stability', *Journal of Economic Theory*, **63**, 299–325.

Clemhout, S. and H.Y. Wan Jr (1994), 'Differential games: economic applications', in R. Aumann and S. Hart (eds), *Handbook of Game Theory with Economic Applications*, Amsterdam: Elsevier, pp. 801–25.
Cline, W.R. (1992a), *Global Warming: The Economic Stakes*, Washington, DC: Institute for International Economics.
Cline, W.R. (1992b), *The Economics of Global Warming*, Washington, DC: Institute for International Economics.
Coase, R. (1960), 'The problem of social cost', *Journal of Law and Economics*, **3**, 1–44.
Compte, O. and P. Jehiel (1997), 'International negotiations and dispute resolution mechanisms: the case of environmental negotiations', in C. Carraro (ed.), *International Environmental Negotiations: Strategic Policy Issues*, Cheltenham, UK: Edward Elgar, pp. 56–70.
Congelton, R.D. (1992), 'Institutions for internalizing international environmental externalities', *Review of Economics and Statistics*, **74**, 412–21.
Congleton, R.D. (1994), *International Institutions for Environmental Protection: Transactions Costs and Environmental Treaties*, Working Paper: Economics, Energy, Environment, no. 39.94, Milan: Fondazione Eni Enrico Mattei.
Congleton, R.D. (1996), 'Political institutions and pollution control', in R.D. Congleton (ed.), *The Political Economy of Environmental Protection: Analysis and Evidence*, Ann Arbor, MI: University of Michigan Press, pp. 273–89.
Cornes, R. and T. Sandler (1983), 'On commons and tragedies', *American Economic Review*, **73**, 787–92.
Cornes, R. and T. Sandler (1984a), 'Easy riders, joint production, and public goods', *Economic Journal*, **94**, 580–98.
Cornes, R. and T. Sandler (1984b), 'The theory of public goods: non-Nash behaviour', *Journal of Public Economics*, **23**, 367–79.
Cornes, R. and T. Sandler (1985a), 'On the consistency of conjectures with public goods', *Journal of Public Economics*, **27**, 125–9.
Cornes, R. and T. Sandler (1985b), 'The simple analytics of pure public good provision', *Economica*, **52**, 103–16.
Cornes, R. and T. Sandler (1986), *The Theory of Externalities, Public Goods and Club Goods*, Cambridge: Cambridge University Press.
Cournot, A. (1938), *Recherches sur les principes mathématiques de la théorie des richesses*. *Librairie des sciences politiques et sociales*, Paris: M. Rivière & cie.
Crocker, T.D. (1984), 'Scientific truths and policy truths in acid deposition research', in T.D. Crocker (ed.), *Economic Perspectives on Acid Rain Deposition Control*, Acid Precipitation Series, vol. 8, London and Boston, MA: Butterworth, pp. 65–79.

Crosson, P.R. (1989), 'Climate change: problems of limits and policy responses', in N.J. Rosenberg, W.E. Easterling III, P.R. Crosson and J. Darmstadter (eds), *Greenhouse Warming: Abatement and Adaption. Proceedings of a Workshop, June 14 to 15, 1988*, Washington, DC: Resources for the Future.

Dasgupta, P.S. (1990), 'The environment as a commodity', *Oxford Review of Economic Policy*, **6**, 51–67.

D'Aspremont, C. and J.J. Gabszewicz (1986), 'On the stability of collusion', in G.F. Matthewson and J.E. Stiglitz (eds), *New Developments in the Analysis of Market Structure*, New York: Macmillan, pp. 243–64.

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

DeSombre, E.R. and J. Kauffman (1996), 'The Montreal Protocol Multilateral Fund: partial success story', in R.O. Keohane and M.A. Levy (eds), *Institutions for Environmental Aid: Pitfalls and Promise*, Cambridge, MA and London: MIT Press, pp. 89–126.

Dockner, E.J. and N. Van Long (1993), 'International pollution control: cooperative versus noncooperative strategies', *Journal of Environmental Economics and Management*, **24**, 13–29.

Driffill, J. and C. Schultz, (1995), 'Renegotiation in a repeated Cournot duopoly', *Economic Letters*, **47**, 143–8.

Dutta, P.K. and R.K. Sundaram (1993), 'The tragedy of the commons?', *Economic Theory*, **3**, 413–26.

Ecchia, G. and M. Mariotti (1997), 'The stability of international environmental coalitions with farsighted countries: some theoretical observations', in C. Carraro (ed.), *International Environmental Negotiations: Strategic Policy Issues*, Cheltenham, UK: Edward Elgar, pp. 172–92.

Eichberger, J. (1993), *Game Theory for Economists*, San Diego, CA: Academic Press.

Endres, A. (1993), 'Internationale Vereinbarungen zum Schutz der globalen Umweltressourcen – der Fall proportionaler Emmissionsreduktion', *Aussenwirtschaft – The Swiss Review of International Economic Relations*, **48**, 51–76.

Endres, A. (1994), *Umweltökonomie: Eine Einführung*, Darmstadt: Wissenschaftliche Buchgesellschaft.

Endres, A. (1995), 'Zur Ökonomie Internationaler Umweltschutzvereinbarungen', *Zeitschrift für Umweltpolitik*, **8**, 143–78.

Endres, A. (1996a), 'Designing a greenhouse treaty: some economic problems', in E. Eide and R. van den Bergh (eds), *Law and Economics of the Environment*, Oslo: Juridisk Forlag, pp. 201–24.

Endres, A. (1996b), 'Negotiating a climate convention: the role of prices and quantities', *International Review of Law and Economics*, **17**, 201–24.
Endres, A. (1997), 'Increasing environmental awareness to protect the global commons: a curmudgeon's view', *Kyklos*, **50**, 3–27.
Endres, A. and M. Finus (1996a), 'Umweltpolitische Zielbestimmung im Spannungsfeld gesellschaftlicher Interessengruppen: Theorie und Empire', in H. Siebert (ed.), *Elemente einer rationalen Umweltpolitik. Expertisen zur umweltpolitischen Neuorientierung*, Tübingen: J.C.B. Mohr.
Endres, A. and M. Finus (1996b), 'Zur Neuen Politischen Ökonomie der Umweltgesetzgebung. Umweltschutzinstrumente im politischen Prozeß', *Zeitschrift für angewandte Umweltforschung (ZAU)*, **8**, 88–103.
Endres, A. and M. Finus (1998a), 'Renegotiation-proof equilibria in a bargaining game over global emission reductions: does the instrumental framework matter?', in N. Hanley and H. Folmer (eds), *Game Theory and the Global Environment*, Cheltenham, UK: Edward Elgar, pp. 135–64.
Endres, A. and M. Finus (1998b), 'Quotas may beat taxes in a global emission game', preliminary draft, University of Hagen, Hagen.
Endres, A. and M. Finus (1998c), 'Playing a better global emission game: does it help to be green?', *Swiss Journal of Economics and Statistics*, **134**, 21–40.
Endres, A. and M. Finus (1999), 'International environmental agreements: how the policy instrument affects equilibrium emissions and welfare', *Journal of Institutional and Theoretical Economics*, **155**, 527–50.
Endres, A. and K. Holm-Müller (1998), *Die Bewertung von Umweltschäden. Theorie und Praxis sozioökonomischer Verfahren*, Stuttgart: Kohlhammer.
Endres, A. and C. Ohl (1998a), 'Globaler Umweltschutz im Spannungsfeld von Risiko- und Kooperationsbereitschaft', *Gaia*, **7**, 279–85.
Endres, A. and C. Ohl (1998b), 'Die Begrenzung globaler Umweltrisiken: eine risikostrategische Betrachtung', *Zeitschrift für Umweltpolitik und Umweltrecht*, **4**, 511–17.
Evans, R. and E. Maskin (1989), 'Efficient renegotiation-proof equilibria in repeated games', *Games and Economic Behavior*, **1**, 361–9.
Eyckmans, J. (1997), 'Nash implementation of a proportional solution to international pollution control problems', *Journal of Environmental Economics and Management*, **33**, 314–30.
Fankhauser, S. (1995), *Valuing Climate Change: The Economics of the Greenhouse*, London: Earthscan.
Fankhauser, S. and S. Kverndokk (1996), 'The global warming game: simulations of a CO_2 reduction agreement', *Resource and Energy Economics*, **18**, 83–102.

Farrell, J. (1983), 'Credible repeated game equilibrium' unpublished manuscript.
Farrell, J. and E. Maskin (1989a), 'Renegotiation in repeated games', *Games and Economic Behavior*, **1**, 327–60.
Farrell, J. and E. Maskin (1989b), 'Renegotiation-proof equilibrium: reply', *Journal of Economic Theory*, **49**, 376–8.
Feeny, D., F. Berkes, B. McCray, J.M. Acheson (1990), 'The tragedy of the commons: twenty-two years later', *Human Ecology*, **18**, 1–19.
Fees, E. (1995), *Umweltökonomie und Umweltpolitik*, Munich: Franz Vahlen.
Felder, S. and T.F. Rutherford (1993), 'Unilateral reductions and carbon leakage: the consequences of international trade in oil and basic materials', *Journal of Environmental Economics and Management*, **25**, 162–76.
Feldman, A.M. (1980), *Welfare Economics and Social Choice Theory*, Boston, MA: Martinus Nijhoff Publishing.
Finus, M. (1992a), 'Ansätze sur Messung des Wertes von Umweltgütern in der Landwirtschaft: Methodische Grundlagen', *Agrarwirtschaft*, **41**, 367–74.
Finus, M. (1992b), *Der Reisekostenansatz – Darstellung, methodische Probleme und Anwendungsmöglichkeiten*, Materialien des Zentrums für regionale Entwicklungsforschung der Justus-Liebig-Universität Giessen, vol. 24, Giessen.
Finus, M. (1997), 'Eine spieltheoretische Betrachtung internationaler Umweltprobleme: eine Einführung', in P. Weise (ed.), *Nachhaltigkeit in der Ökonomischen Theorie. Ökonomie und Gesellschaft. Jahrbuch 14*, Frankfurt and New York: Campus, pp. 239–300.
Finus, M. and B. Rundshagen (1997), *Toward a Positive Theory of Coalition Formation and Endogenous Instrumental Choice in Global Pollution Control*, Discussion Paper no. 239, University of Hagen, Hagen.
Finus, M. and B. Rundshagen (1998a), 'Toward a positive theory of coalition formation and endogenous instrumental choice in global pollution control', *Public Choice*, **96**, 145–86.
Finus, M. and B. Rundshagen (1998b), 'Renegotiation-proof equilibria in a global emission game when players are impatient', *Environmental and Resource Economics*, **12**, 275–306.
Finus, M. and B. Rundshagen (1999), 'Strategic links between environmental and trade policies if plant location is endogenous', preliminary draft, University of Hagen, Hagen.
Finus, M. and S. Tjøtta (1998), *The Oslo Agreement on Sulfur Reduction in Europe: The Great Leap Forward?*, Working Paper 1898, University of Bergen, Bergen.

Foley, D. (1970), 'Lindahl solution and the core of an economy with public goods', *Econometrica*, **38**, 66–72.
Folmer, H., P. von Mouche and S. Ragland (1993), 'Interconnected games and international environmental problems', *Environmental and Resource Economics*, **3**, 313–35.
Førsund, F.R. and E. Naevdal (1994), 'Trading sulphur emissions in Europe', in G. Klaasen and F.R. Førsund (eds), *Economic Instruments for Air Pollution Control, Economy and Environment*, vol. 9, pp. 231–48.
Foster, B.A. (1993), *The Acid Rain Debate: Science and Special Interests in Policy Formation*, Ames, IA: State University Press.
Friedman, A. (1994), 'Differential games', in R. Aumann and S. Hart (eds), *Handbook of Game Theory with Economic Applications*, Amsterdam: Elsevier, pp. 781–99.
Friedman, J.W. (1971), 'A non-cooperative equilibrium for supergames', *Review of Economic Studies*, **38**, 1–12.
Friedman, J.W. (1985), 'Cooperative equilibria in finite horizon noncooperative supergames', *Journal of Economic Theory*, **35**, 390–98.
Friedman, J.W. (1986), *Game Theory with Applications to Economics*, London and New York: Oxford University Press.
Fudenberg, D. and D. Levine (1992), 'Maintaining a reputation when strategies are imperfectly observable', *Review of Economic Studies*, **59**, 561–79.
Fudenberg, D. and E. Maskin (1986), 'The folk theorem in repeated games with discounting or with incomplete information', *Econometrica*, **54**, 533–54.
Fudenberg, D. and J. Tirole (1996), *Game Theory*, 5th edn, Cambridge, MA and London: MIT Press.
Fudenberg, D., D. Levine and E. Maskin (1994), 'The folk theorem with imperfect public information', *Econometrica*, **62**, 997–1039.
Gardner, R. and E. Ostrom (1991), 'Rules and games', *Public Choice*, **70**, 121–49.
Garrod, G. and K.G. Willis (1999), *Economic Valuation of the Environment*, Cheltenham, UK: Edward Elgar.
General Accounting Office (GAO) (1992), *International Environmental Agreements Are Not Well Monitored*, Washington, DC: United States General Accounting Office, RCED-92-43.
Germain, M., P.L. Toint and H. Tulkens (1995), *International Negotiations on Acid Rains in Northern Europe: A Discrete Time Iterative Process*, CORE Discussion Paper no. 9556, Louvain: Center for Operations Research and Econometrics, Université Catholique de Louvain.
Germain, M., P.L. Toint and H. Tulkens (1996), 'Calcul économique

itératif et stratégique pour les négociations internationales sur les pluies acides entre la Finlande, la Russie et l'Estonie', *Annales d'économie et statistique*, **43**, 101–27.
Germain, M., P.L. Toint, H. Tulkens and A. de Zeeuw (1998), *Transfers to Sustain Core-theoretic Cooperation in International Stock Pollutant Control*, CORE Discussion Paper no. 9832, Louvain: Center for Operations Research and Econometrics, Université Catholique de Louvain.
Gibbons, R. (1992), *A Primer in Game Theory*, New York: Harvester Wheatsheaf.
Gibbons, R. (1997), 'An introduction of applicable game theory', *Journal of Economic Perspectives*, **11**, 127–49.
Golombek, R., C. Hagem and M. Hoel (1995), 'Efficient incomplete international agreements', *Resource and Energy Economics*', **17**, 25–46.
Green, E. and R. Porter (1984), 'Noncooperative collusion under imperfect price information', *Econometrica*, **52**, 87–100.
Grubb, M. (1989), *The Greenhouse Effect: Negotiating Targets*, London: Royal Institute of International Affairs.
Gündling, L. (1996), 'Compliance assistance in international environmental law: capacity building through financial and technology transfer', *Zeitschrift für ausländisches öffentliches Recht und Völkerrecht*, **56**, 796–811.
Güth, W. and R. Pethig (1992), 'Illegal pollution and monitoring of unknown quality: a signaling game approach', in R. Pethig (ed.), *Conflicts and Cooperation in Managing Environmental Resources: Microeconomic Studies*, Berlin: Springer, pp. 275–330.
Guttman, J.M. (1987), 'A non-Cournot model of voluntary collective action', *Economica*, **54**, 1–19.
Guttman, J.M. (1991), 'Voluntary collective action', in A.L. Hillman (ed.), *Markets and Politicians*, Dordrecht: Kluwer, pp. 1–10.
Guttman, J.M. and A. Schnytzer (1992), 'A solution of the externality problem using strategic matching', *Social Choice Welfare*, **8**, 73–88.
Haas, P.M., R.O. Keohane and M.A. Levy (eds) (1993), *Institutions for the Earth*, Cambridge, MA and London: MIT Press.
Hahn, R.W. (1987), 'Jobs and environmental quality: some implications for instrument choice', *Policy Sciences*, **20**, 289–306.
Hahn, R.W. (1989), *A Primer on Environmental Policy Design*, Chur: Harwood Academic Publishers.
Hamburger, H. (1973), '*N*-person prisoners' dilemma', *Journal of Mathematical Sociology*, **3**, 27–48.
Hanley, N. and H. Folmer (eds) (1998), *Game Theory and the Global Environment*, Cheltenham, UK: Edward Elgar.

Hanley, N., J.F. Shogren and B. White (1997), *Environmental Economics in Theory and Practice*, London: Macmillan.

Harsanyi, J. (1973), 'Games with randomly disturbed payoffs: a new rationale for mixed-strategy equilibrium points', *International Journal of Game Theory*, **2**, 1–23.

Hart, S. and M. Kurz (1983), 'Endogenous formation of coalitions', *Econometrica*, **51**, 1047–64.

Heal, G. (1994), 'Formation of international environmental agreements', in C. Carraro (ed.), *Trade, Innovation, Environment*, Dordrecht: Kluwer, pp. 301–22.

Heister, J. (1997), *Der internationale CO_2-Vertrag: Strategien zur Stabilisierung multilateraler Kooperation zwischen souveränen Staaten*, Tübingen: J.C.B. Mohr.

Heister, J. (1998), 'Who will win the ozone game? On building and sustaining cooperation in the Montreal Protocol on Substances that Deplete the Ozone Layer', in P. Michaelis and F. Stähler (eds), *Recent Policy Issues in Environmental and Resource Economics*, Heidelberg and New York: Physica, pp. 121–54.

Heister, J., E. Mohr, W. Plesmann, F. Stähler, T. Stoll and R. Wolfrum (1995), *Economic and Legal Aspects of International Agreements: The Case of Enforcing and Stabilizing an International CO_2 Agreement*, Working Paper 711, Kiel: Kieler Institut für Weltwirtschaft.

Heister, J., E. Mohr, T. Stoll and R. Wolfrum (1997), 'Strategies to enforce compliance with an international CO_2 treaty', *International Environmental Affairs*, **9**, 22–53.

Hicks, J.R. (1940), 'The valuation of social income', *Economica*, **7**, 105–24.

Hoel, M. (1991), 'Global environmental problems: the effects of unilateral actions taken by one country', *Journal of Environmental Economics and Management*, **20**, 55–70.

Hoel, M. (1992a), 'International environment conventions: the case of uniform reductions of emissions', *Environmental and Resource Economics*, **2**, 141–59.

Hoel, M. (1992b), 'Emission taxes in a dynamic international game of CO_2 emissions', in R. Pethig (ed.), *Conflicts and Cooperation in Managing Environmental Resources: Microeconomic Studies*, Berlin: Springer, pp. 39–68.

Hoel, M. (1992c) 'Carbon taxes: an international tax or harmonized domestic taxes?', *European Economic Review*, **36**, 400–406.

Hoel, M. (1994), 'Efficient climate policy in the presence of free riders', *Journal of Environmental Economics and Management*, **27**, 259–74.

Hoel, M. and K. Schneider (1997), 'Incentives to participate in an

international environmental agreement', *Environmental and Resource Economics*, **9**, 153–70.
Holler, M.J. and G. Illing (1993), *Einführung in die Spieltheorie*, 2nd edn, Berlin: Springer.
Holler, M.J. and G. Illing (1996), *Einführung in die Spieltheorie*, 3rd edn, Berlin: Springer.
Hübner, M. and M. Dröttboom (1999), *Environmental Consciousness and Moral Hazard in International Agreements to Protect the Environment: A Note*, Working Paper no. 264, University of Hagen, Hagen.
Intergovernmental Panel on Climate Change (IPCC) (1996a), *Climate Change 1995. Impacts, Adaptions and Mitigation of Climate Change: Scientific–Technical Analyses*, Contribution of Working Group II, Cambridge: Cambridge University Press.
Intergovernmental Panel on Climate Change (IPCC) (1996b), *Climate Change 1995. Economic and Social Dimensions of Climate Change*, Contribution of Working Group III, Cambridge: Cambridge University Press.
Jeppesen, T. and P. Andersen (1998), 'Commitment and fairness in environmental games', in N. Hanley, and H. Folmer (eds), *Game Theory and the Environment*, Cheltenham, UK: Edward Elgar, pp. 65–83.
Johansson, P.O. (1990), 'Valuing environmental damage', *Oxford Review of Economic Policy*, **6**, 34–50.
Jordan, A. and J. Werksman (1996), 'Financing global environmental protection', in J. Cameron, J. Werksman and P. Roderick (eds), *Improving Compliance with International Environmental Law*, London: Earthscan, pp. 247–55.
Just, R.E., D.L. Hueth and A. Schmitz (1982), *Applied Welfare Economics and Public Policy*, Englewood Cliffs, NJ: Prentice-Hall.
Kaitala, V., K.-G. Mäler and H. Tulkens (1995), 'The acid rain game as a resource allocation process with an application to international cooperation among Finland, Russia and Estonia', *Scandinavian Journal of Economics*, **97**, 325–43.
Kaitala, V., M. Pohjola and O. Tahvonen (1991), 'An analysis of SO_2 negotiations between Finland and the Soviet Union', *Finnish Economic Papers*, **4**, 104–12.
Kaitala, V., M. Pohjola and O. Tahvonen (1992), 'Transboundary air pollution and soil acidification: a dynamic analysis of an acid rain game between Finland and the USSR', *Environmental and Resource Economics*, **2**, 161–81.
Kaldor, N. (1939), 'Welfare propositions of economics and interpersonal comparisons of utility', *Economic Journal*, **49**, 549–51.

Kamien, M.I. and N.L. Schwartz (1991), *Dynamic Optimization: The Calculus of Variations and Optimal Control in Economics and Management*, 2nd edn, New York: North-Holland.

Kaneko, M. (1997), 'The ratio equilibrium and a voting game in a public goods economy', *Journal of Economic Theory*, **16**, 123–36.

Katsoulacos, Y. (1997), 'R&D spillovers, cooperation, subsidies and international agreements', in C. Carraro (ed.), *International Environmental Negotiations: Strategic Policy Issues*, Cheltenham UK: Edward Elgar, pp. 97–109.

Kelsen, H. and R.W. Tucker (1967), *Principles of International Law*, 2nd edn, New York: Holt, Rinehart & Winston.

Keohane, R.O. (1995), 'Compliance with international standards: environmental case studies', in J.L. Hargrove (ed.), *Proceedings of the Eighty-ninth Annual Meeting of the American Society of International Law* (ASIL Proceedings), Buffalo, NY: Hein & Co., pp. 206–24.

Kneese, A.V. (1988), 'Environmental stress and political conflicts: salinity in the Colorado River', paper presented at the Conference on Environmental Stress and Security, Stockholm.

Kölle, C. (1995), *Ökonomische Analayse internationaler Umweltkooperationen*, Heidelberg: Physica.

Konishi, H., M. Le Breton and S. Weber (1997), 'Group formation in games without spillovers', in C. Carraro and D. Siniscalco (eds), *New Directions in the Economic Theory of the Environment*, Cambridge: Cambridge University Press, pp. 281–309.

Koutsoyiannis, A. (1991), *Modern Microeconomics*, 2nd edn, London: Macmillan.

Kreps, D.M. (1989), 'Out-of-equilibrium beliefs and out-of-equilibrium behaviour', in F. Hahn (ed.), *The Economics of Missing Markets, Information, and Games*, Oxford: Clarendon Press, pp. 7–45.

Kreps, D.M. (1990), *Game Theory and Economic Modelling*, Oxford: Clarendon Press.

Kroeze-Gil, J. and H. Folmer (1998), 'Linking environmental problems in an international setting: the interconnected games approach', in N. Hanley and H. Folmer (eds), *Game Theory and the Global Environment*, Cheltenham, UK: Edward Elgar, pp. 165–80.

Krutilla, J.V. (1975), 'The International Columbia River Treaty: an economic evaluation', in A.V. Kneese and S.C. Smith (eds), *Water Research*, Baltimore, MD: Johns Hopkins University Press, pp. 68–97.

Kuhl, H. (1987), *Umweltressourcen als Gegenstand internationaler Verhandlungen. Eine theoretische Transaktionskostenanalyse*, Staatliche Allokationspolitik im marktwirtschaftlichen System, no. 23, Frankfurt: Peter Lang.

Kummer, K. (1994), 'Providing incentives to comply with multilateral environmental agreements: an alternative to sanctions?', *European Environmental Law Review*, **3**, 256–63.

Kverndokk, S. (1993), 'Global CO_2 agreements: a cost-effective approach', *Energy Journal*, **14**, 91–112.

Ladenburger, F. (1996), *Durchsetzungsmechanismen im Umweltvölkerrecht. 'Enforcement' gegenüber den Staaten*, Tübingen: Zeeb-Druck.

Laffont, J.-J. (1993), *Regulation of Pollution with Asymmetric Information*, Working Paper: Economics, Energy, Environment, no. 10.93, Milan: Fondazione Eni Enrico Mattei.

Lenschow, A. (1996), 'Der umweltpolitische Entscheidungsprozeß in der Europäischen Union am Beispiel der Klimapolitik', in H.G. Brauch (ed.), *Klimapolitik*, Berlin: Springer, pp. 89–104.

Luce, R. and Raiffa, H. (1957), *Games and Decisions*, New York: John Wiley.

Machina, M.J. (1989), 'Decision-making in the presence of risk', in F. Hahn (ed.), *The Economics of Missing Markets, Information, and Games*, Oxford: Clarendon Press, pp. 278–94.

Makowski, L. (1983), '"Rational conjectures" aren't rational, "reasonable conjectures" aren't reasonable', Economic Theory Discussion Paper, Cambridge.

Mäler, K.-G. (1989), 'The acid rain game', in H. Folmer and E. van Ierland (eds), *Valuation Methods and Policy Making in Environmental Economics*, Amsterdam: Elsevier, pp. 231–52.

Mäler, K.-G. (1990), 'International environmental problems', *Oxford Review of Economic Policy*, **6**, 80–108.

Mäler, K.-G. (1991), 'Incentives in international environmental problems', in H. Siebert (ed.), *Environmental Scarcity: The International Dimension*, Tübingen: J.C.B. Mohr, pp. 75–93.

Mäler, K.-G. (1992), 'Critical loads and international environmental cooperation', in R. Pethig (ed.), *Conflicts and Cooperation in Managing Environmental Resources: Microeconomics Studies*, Berlin: Springer, pp. 71–81.

Mäler, K.-G. (1994), 'Acid rain in Europe: a dynamic perspective on the use of economic incentives', in E.C. van Ierland (ed.), *International Environmental Economics*, Developments in Environmental Economics no. 4, Amsterdam: Elsevier, pp. 351–72.

Malinvaud, E. (1985), *Lectures in Microeconomic Theory*, rev. edn, Amsterdam: North-Holland.

Manne, A.S. and R.G. Richels (1991), 'Global CO_2 emission reductions: the impacts of rising energy costs', *Energy Journal*, **12**, 87–107.

Margolis, H. (1982), *Selfishness, Altruism, and Rationality*, Cambridge: Cambridge University Press.

Marino, A.N. (1988), 'Monopoly, liability and regulation', *Southern Economic Journal*, **54**, 913–27.
Martin, W.E., R.H. Patrick and B. Tolwinski (1993), 'A dynamic game of a transboundary pollutant with asymmetric players', *Journal of Environmental Economics and Management*, **24**, 1–12.
Mas-Colell, A. and J. Silvester (1989), 'Cost sharing equilibria: a Lindahlian approach', *Journal of Economic Theory*, **47**, 229–56.
McLean, I. (1981), 'The social contract and the prisoners' dilemma supergame', *Political Studies*, **29**, 339–51.
Michaelis, P. (1992), 'Global warming: efficient policies in the case of multiple pollutants', *Environmental and Resource Economics*, **2**, 61–77.
Mohr, E. (1988), 'On the incredibility of perfect threats in repeated games: note', *International Economic Review*, **29**, 551–5.
Mohr, E. (1991), 'Global warming: economic policy in the face of positive and negative spillovers', in H. Siebert (ed.), *Environmental Scarcity: The International Dimension*, Tübingen: J.C.B. Mohr, pp. 186–212.
Mohr, E. (1995), *Green Policy Persuasion: Towards a Positive Theory of Discounting the Climate Future*, IWÖ Discussion Paper no. 25, University of St Gallen, St Gallen.
Mohr, E. and J.P. Thomas (1998), 'Pooling sovereign risks: the case of environmental treaties and international debt', *Journal of Development Economics*, **55**, 173–90.
Moreno, D. and J. Wooders (1993), *Coalition-proof Equilibrium*, Discussion Paper 93–7, Department of Economics, University of Arizona, Tucson, AZ.
Moulin, H. (1986), *Game Theory for the Social Sciences: Studies in Game Theory and Mathematical Economics*, 2nd edn, New York: New York University Press.
Moulin, H. (1988), *Axioms of Cooperative Decision Making*, Cambridge: Cambridge University Press.
Moulin, H. (1995), *Cooperative Microeconomics: A Game-theoretic Introduction*, Englewood Cliffs, NJ: Prentice-Hall, and Hemel Hempstead, Herts: Harvester Wheatsheaf.
Mueller, D.C. (1989), *Public Choice II*, Cambridge: Cambridge University Press.
Müller-Fürstenberger, G. and G. Stephan (1997), 'Environmental policy and cooperation beyond the nation state: an introduction and overview', *Structural Change and Economic Dynamics*, **8**, 99–114.
Murdoch, J.C. and T. Sandler (1997), 'Voluntary cutbacks and pretreaty behavior: the Helsinki Protocol and sulfur emissions', *Public Finance Review*, **25**, 139–62.
Musgrave, P. (1994), 'Pure global externalities: international efficiency and

equity', in L. Bovenberg and S. Cnossen (eds), *Public Economics and the Environment in an Imperfect World*, Boston, MA: Kluwer, pp. 237–59.
Myerson, R.B. (1991), *Game Theory: Analysis of Conflict*, Cambridge, MA and London: Harvard University Press.
Myerson, R.B. (1997), 'Game-theoretic models of bargaining: an introduction for economists studying the transnational commons', in P. Dasgupta, K.-G. Mäler and A. Vercelli (eds), *The Economics of Transnational Commons*, Oxford: Clarendon Press, pp. 17–34.
Na, S. and H.S. Shin (1998), 'International environmental agreements under uncertainty', *Oxford Economic Papers*, **50**, 173–85.
Nash, J. (1950a), 'Equilibrium points in *N*-person games', *Proceedings of the National Academy of Sciences*, **36**, 48–9.
Nash, J. (1950b), 'The bargaining problem', *Econometrica*, **18**, 155–62.
Nentjes, A. (1994), 'Control of reciprocal transboundary pollution and joint implementation', in G. Klaasen and F.R. Førsund (eds), *Economic Instruments for Air Pollution Control*, Economy and Environment, vol. 9, Dordrecht: Kluwer, pp. 209–30.
Newbery, D. (1990), 'Acid rain', *Economic Policy*, **5**, 287–346.
Nitze, W.A. (1990), *The Greenhouse Effect: Formulating a Convention*, London: Royal Institute of International Affairs.
Nordhaus, W.D. (1991a), 'A sketch on the economics of the greenhouse effect', *American Economic Review*, **81**, 146–50.
Nordhaus, W.D. (1991b), 'To slow or not to slow: the economics of the greenhouse effect', *Economic Journal*, **101**, 920–37.
Nordhaus, W.D. (1991c), 'The cost of slowing climate change: a survey', *Energy Journal*, **12**, 37–64.
Nordhaus, W.D. (1993), 'Rolling the "DICE": an optimal transition path for controlling greenhouse gases', *Resource and Energy Economics*, **15**, 27–50.
Olson, M. (1965), *The Logic of Collective Action*, Cambridge, MA: Harvard University Press.
Ordeshook, P.C. (1986), *Game Theory and Political Theory: An Introduction*, Cambridge: Cambridge University Press.
Ordeshook, P.C. (1992), *A Political Theory Primer*, New York and London: Routledge Press.
Osborne, M.J. and A. Rubinstein (1990), *Bargaining and Markets*, San Diego, CA: Academic Press.
Osborne, M.J. and A. Rubinstein (1994), *A Course in Game Theory*, Cambridge, MA and London: MIT Press.
Ostrom, E., R. Gardner and J. Walker (1990), 'The nature of common-pool resource problems', *Rationality and Society*, **2**, 335–58.
Ostrom, E., R. Gardner and J. Walker (1994), *Rules, Games, and Common-pool Resources*, Ann Arbor, MI: University of Michigan Press.

Owen, G. (1982), *Game Theory*, 2nd edn, New York: Academic Press.
Oye, K.A. (1986), 'Explaining cooperation under anarchy: hypotheses and strategies', in K.A. Oye (ed.), *Cooperation Under Anarchy*, Princeton, NJ: Princeton University Press, pp. 1–24.
Pearce, D., A. Markandya and E.B. Barbier (1989), *Blueprint for a Green Economy*, London: Earthscan.
Pearce, D.G. (1992), 'Repeated games: cooperation and rationality', in J.-J. Laffont (ed.), *Advances in Economic Theory: Sixth World Congress of the Econometric Society*, Cambridge: Cambridge University Press, pp. 132–74.
Pethig, R. (1982), 'Reciprocal transfrontier pollution', in H. Siebert (ed.), *Global Environmental Resources*, Frankfurt: Verlag Peter Lang, pp. 57–93.
Pethig, R. (ed.) (1992), *Conflicts and Cooperation in Managing Environmental Resources: Microeconomic Studies*, Berlin: Springer.
Petrakis, E. and A. Xepapadeas (1996), 'Environmental consciousness and moral hazard in international agreements to protect the environment', *Journal of Public Economics*, **60**, 95–110.
Ragland, S.E., L.L. Bennett and P. Yolles (1996), 'Unidirectional problems with international river basins: a game theoretic approach', paper presented at the European Meeting of Resource and Environmental Economists, Lisbon, 26–29 June.
Raiffa, H. (1982), *The Art and Science of Negotiation*, Cambridge, MA: Harvard University Press.
Rasmusen, E. (1989), *Games and Information: An Introduction*, Oxford and New York: Basil Blackwell.
Rasmusen, E. (1995), *Games and Information. An Introduction to Game Theory*, 2nd edn, Oxford and Cambridge, MA: Basil Blackwell.
Ray, D. (1994), 'Internally renegotiation-proof equilibrium sets: limit behavior with low discounting', *Games and Economic Behavior*, **6**, 162–77.
Ray, D. and R. Vohra (1997), 'Equilibrium binding agreements', *Journal of Economic Theory*, **73**, 30–78.
Ray, D. and R. Vohra (1999), 'A theory of endogenous coalition structure', *Games and Economic Behavior*, **26**, 286–336.
Richer, J. and J.K. Stranlund (1997), 'Threat positions and the resolution of environmental conflicts', *Land Economics*, **73**, 58–71.
Rosen, J.B. (1965), 'Existence and uniqueness of equilibrium points for concave *n*-person games', *Econometrica*, **33**, 520–34.
Roth, A. (1979), *Axiomatic Models of Bargaining*, Berlin: Springer.
Roth, A. (ed.) (1985), *Game-theoretic Models of Bargaining*, Cambridge: Cambridge University Press.

Rotillon, G. and T. Tazdait (1996), 'International bargaining in the presence of global environmental change', *Environmental and Resource Economics*, **8**, 293–314.
Rubinstein, A. (1976), 'Equilibrium in supergames', Centre for Mathematical Economics and Game Theory, Hebrew University, Jerusalem, mimeo.
Rubinstein, A. (1980), 'Strong perfect equilibrium in supergames', *International Journal of Game Theory*, **9**, 1–12.
Rubinstein, A. (1982), 'Perfect equilibrium in a bargaining model', *Econometrica*, **50**, 97–109.
Rubinstein, A. (1987), 'A sequential strategic theory of bargaining', in T.F. Bewley (ed.), *Advances in Economic Theory: Fifth World Congress of the Econometric Society*, Cambridge: Cambridge University Press, pp. 197–225.
Russell, C.S. (1990), 'Game models for structuring monitoring and enforcement systems', *Natural Resource Modeling*, **4**, 143–73.
Russell, C.S. (1992), 'Monitoring and enforcement of pollution control laws in Europe and the United States', in R. Pethig (ed.), *Conflicts and Cooperation in Managing Environmental Resources: Microeconomics Studies*, Berlin: Springer, pp. 195–213.
Russell, C.S., W. Harrington, and W. Vaughan (1986), *Enforcing Pollution Control Laws*, Washington, DC: Resources for the Future.
Sabourian, H. (1989), 'Repeated games: a survey', in F. Hahn (ed.), *The Economics of Missing Markets, Information, and Games*, Oxford: Clarendon Press, pp. 62–105.
Sand, P.H. (ed.) (1992), *The Effectiveness of International Environmental Agreements*, Cambridge: Grotius Publications.
Sand, P.H. (1994), 'The present state of research', in S. Subedi (ed.), *Environmental Policy: From Regulation to Economic Instruments: International Legal Aspects of Eco-labels*, The Hague: Institute of Social Studies.
Sand, P.H. (1996), 'Compliance with international environmental obligations: existing international legal agreements', in J. Cameron, J. Werksman and P. Roderick (eds), *Improving Compliance with International Environmental Law*, London: Earthscan, pp. 48–82.
Sand, P.H. (1997), 'Commodity or taboo? International regulation of trade in endangered species', in H.O. Bergensen and G. Parmann (eds), *The Green Globe Year Book 1997*, New York: Oxford University Press, pp. 19–36.
Sandler, T. (1992), *Collective Action: Theory and Application*, Ann Arbor, MI: University of Michigan Press.
Sandler, T. (1996), 'A game-theoretic analysis of carbon emissions', in R.D.

Congleton (ed.), *The Political Economy of Environmental Protection: Analysis and Evidence*, Ann Arbor, MI: University of Michigan Press, pp. 251–72.
Schelling, T.C. (1960), *The Strategy of Conflict*, Cambridge, MA: Harvard University Press.
Schelling, T.C. (1991), 'Economic responses to global warming: prospects for cooperative approaches', in R. Dornbusch and J.M. Poterba (eds), *Global Warming: Economic Policy Responses*, Cambridge, MA and London: MIT Press, pp. 197–231.
Schotter, A. and G. Schwödiauer (1980), 'Economics and theory of games: a survey', *Journal of Economic Literature*, **18**, 479–527.
Schultz, C. (1994), 'A note on strongly renegotiation-proof equilibria', *Games and Economic Behavior*, **6**, 469–73.
Seierstad, A. and K. Sydsater (1987), *Optimal Control Theory with Applications*, Amsterdam: North-Holland.
Selten, R. (1965), 'Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit', *Zeitschrift für die gesamte Staatswissenschaften*, **12**, 301–24.
Selten, R. (1975), 'Reexamination of the perfectness concept of equilibrium points in extensive games', *International Journal of Game Theory*, **4**, 25–55.
Selten, R. (1982), 'Einführung in die Theorie der Spiele bei unvollständiger Information', in *Information in der Wirtschaft, Schriften des Vereins für Sozialpolitik*, Neue Folge, vol. 126, Berlin: Duncker & Humbolt.
Sen, A.K. (1984), *Collective Choice and Social Welfare*, Amsterdam: Elsevier.
Shapley, L.S. (1971), 'Cores and convex games', *International Journal of Game Theory*, **1**, 11–26.
Shenoy, P. (1979), 'On coalition formation: a game theoretical approach', *International Journal of Game Theory*, **8**, 133–64.
Shogren, J.F. and T.D. Crocker (1991), 'Cooperative and noncooperative protection against transferable and filterable externalities', *Environmental and Resource Economics*, **1**, 195–214.
Shogren, J.F., K.H. Baik and T.D. Crocker (1992), 'Environmental conflicts and strategic commitment', in R. Pethig (eds), *Conflicts and Cooperation in Managing Environmental Resources: Microeconomic Studies*, Berlin: Springer.
Siebert, H. (1985), 'Spatial aspects of environmental economics', in A.V. Kneese and J.L. Sweeney (eds), *Handbook of Natural Resource and Energy Economics*, vol. I, Amsterdam: North-Holland, pp. 125–64.
Siebert, H. (1992), *Economics of the Environment: Theory and Policy*, 3rd edn, Berlin: Springer.

Simma, B. (1970), 'Reflections on Article 60 of the Vienna Convention on the Law of Treaties and its background in general international law', *Österreichische Zeitschrift für öffentliches Recht*, **20**, 5–83.
Snidal, D. (1985), 'The game theory of international politics', *World Politics*, **38**, 25–57.
Snidal, D. (1988), 'Coordination versus prisoners' dilemma: implications for international cooperation and regimes', *American Political Science Review*, **79**, 923–42.
Snyder, G.H. (1971), '"Prisoners' dilemma" and "chicken" models in international politics', *International Studies Quarterly*, **15**, 66–103.
Spagnolo, G. (1996), *Issue Linkage, Delegation and International Policy Coordination*, Working Paper: Economics, Energy, Environment, no. 49.96, Milan: Fondazione Eni Enrico Mattei.
Stähler, F. (1996), 'Reflections on multilateral environmental agreements', in A. Xepapadeas (ed.), *Economic Policy for the Environment and Natural Resources: Techniques for the Management and Control of Pollution*, Cheltenham, UK and Brookfield, USA: Edward Elgar, pp. 174–96.
Stähler, F. (1998a), 'On the economics of international environmental agreements', in P. Michaelis and F. Stähler (eds), *Recent Policy Issues in Environmental and Resource Economics*, Heidelberg and New York: Physica, pp. 155–73.
Stähler, F. (1998b), *Economic Games and Strategic Behaviour: Theory and Application*, Cheltenham, UK and Northampton, USA: Edward Elgar.
Stein, A. (1982), 'Coordination and collaboration in an anarchic world', *International Organization*, **36**, 299–324.
Steiner, U. (1997), *Signalling in International Environmental Agreements: Using Pre-agreement Emission Levels as a Signalling Device*, Working Paper 97–9, Department of Economics, Aarhus School of Business, Aarhus.
Sudgen, R. (1984), 'Reciprocity: the supply of public goods through voluntary contributions', *Economic Journal*, **94**, 772–87.
Tahvonen, O. (1994), 'Carbon dioxide abatement as a differential game', *European Journal of Political Economy*, **10**, 685–705.
Tahvonen, O., V. Kaitala and M. Pohjola (1993), 'A Finnish–Soviet acid rain game: noncooperative equilibria, cost efficiency, and sulphur agreements', *Journal of Environmental Economics and Management*, **24**, 87–100.
Taylor, M. (1987), *The Possibility of Cooperation*, Cambridge: Cambridge University Press.
Taylor, M. and H. Ward (1982), 'Chickens, whales, and lumpy goods: alternative models of public goods provision', *Political Studies*, **30**, 350–70.
Tulkens, H. (1979), 'An economic model of international negotiations

relating to transfrontier pollution', in K. Krippendorff (ed.), *Communication and Control in Society*, New York: Gordon & Breach, pp. 199–212.

Tulkens, H. (1998), 'Cooperation versus free-riding in international environmental affairs: two approaches', in N. Hanley and H. Folmer (eds), *Game Theory and the Environment*, Cheltenham, UK: Edward Elgar, pp. 30–44.

Turner, K.R., D. Pearce and I. Bateman (1994), *Environmental Economics: An Elementary Introduction*, New York: Harvester Wheatsheaf.

Underdal, A. (1998a), 'Introduction', in A. Underdal (ed.), *The Politics of International Environmental Management*, Dordrecht: Kluwer, pp. 1–12.

Underdal, A. (1998b), 'Leadership in international environmental negotiations: designing feasible solutions', in A. Underdal (ed.), *The Politics of International Environmental Management*, Dordrecht: Kluwer, pp. 101–28.

United Nations (1972), *Report of the United Nations Conference on the Human Environment*, Document A/Conf/48/14/Rev1, New York: United Nations.

United States International Trade Commission (USITC) (1991), *International Agreements to Protect the Environment and Wildlife*, USITC Publication 2351, Washington, DC: United States International Trade Commission.

Ursprung, H.W. (1992), *The Political Economy of Environmental Decision Making*, Discussion Paper II-176, Faculty of Economics, University of Konstanz, Konstanz.

Van Damme, E. (1989), 'Renegotiation-proof equilibria in repeated prisoners' dilemma', *Journal of Economic Theory*, **47**, 206–17.

Van Damme, E. (1991), *Stability and Perfection of Nash Equilibria*, 2nd edn, Heidelberg: Springer.

Van der Lecq, F. (1996), 'Conventions and institutions in coordination problems', *De Economist*, **144**, 397–428.

Van der Ploeg, F. and A.J. Zeeuw (1992), 'International aspects of pollution control', *Environmental and Resource Economics*, **2**, 117–39.

Varian, H.R. (1984), *Microeconomic Analysis*, 2nd edn, New York and London: W.W. Norton.

Victor, D.G., K. Raustiala and E.B. Skolnikoff (eds) (1998), *The Implementation and Effectiveness of International Environmental Commitments*, Cambridge, MA and London: MIT Press.

Vives, X. (1984), 'Duopoly information equilibrium: Cournot and Bertrand', *Journal of Economic Theory*, **34**, 71–94.

von Neumann, J. and Morgenstern, O. (1944), *The Theory of Games and Economic Behavior*, New York: John Wiley.

von Stackelberg, H. (1934), *Marktform und Gleichgewicht*, Vienna: Julius Springer.
Ward, H. (1987), 'The risks of a reputation for toughness: strategy in public goods provision problems modelled by chicken super games', *British Journal of Political Science*, **17**, 23–52.
Ward, H. (1993), 'Game theory and the politics of global commons', *Journal of Conflict Resolution*, **37**, 203–35.
Weimann, J. (1995), *Umweltökonomik*, 3rd edn, Berlin: Springer.
Welsch, H. (1990), 'Cost-effective control strategies for energy-related transboundary air pollution in Western Europe', *Energy Journal*, **11**, 87–104.
Welsch, H. (1993), 'An equilibrium framework for global pollution problems', *Journal of Environmental Economics and Management*, **25**, 64–79.
Welsch, H. (1995), 'Incentives for forty-five countries to join various forms of carbon reduction agreements', *Resource and Energy Economics*, **17**, 213–37.
Werksman, J. (1997), *Five MEAs, Five Years since Rio: Recent Lessons on the Effectiveness of Multilateral Agreements*, Special Focus Report, London: FIELD (Foundation for International Environmental Law and Development).
Willett, T.D. (ed.) (1988), *Political Business Cycles: The Political Economy of Money, Inflation and Unemployment*, Durham, NC: Duke University Press.
Willis, K.G., K. Button and P. Nijkamp (eds) (1999), *Environmental Valuation*, Cheltenham, UK: Edward Elgar.
Wirl, F. (1994), 'Global warming and carbon taxes: dynamic and strategic interactions between energy consumers and producers', *Journal of Policy Modeling*, **16**, 577–96.
Xepapadeas, A. (1997), *Advanced Principles in Environmental Policy*, Cheltenham, UK and Northampton, USA: Edward Elgar.
Yi, S.-S. (1996), 'Endogenous formation of customs unions under imperfect competition: open regionalism is good', *Journal of International Economics*, **41**, 153–77.
Yi, S.-S. (1997), 'Stable coalition structures with externalities', *Games and Economic Behavior*, **20**, 201–37.
Yi, S.-S. and H. Shin (1995), 'Endogenous formation of coalitions in oligopoly', Department of Economics, Dartmouth College, mimeo.
Young, O.R. and K. von Moltke (1994), 'The consequences of international environmental regimes: report from the Barcelona Workshop', *International Environmental Affairs*, **6**, 348–70.

Index