Essays on Technology Choice and Spillovers

Erika Färnstrand Damsgaard

Stockholm University

ISSN 0346-6892
ISBN 978-91-7155-669-1

Cover Picture: By an unknown photographer, Wright Aeroplane, Ft. Myer, VA, Sept 1908. Records of the Office of the Chief Signal Officer.

Printed in Sweden by Arkitektkopia AB, Stockholm 2008
Distributor: Institute for International Economic Studies

Doctoral Dissertation
Department of Economics
Stockholm University

Abstract

This thesis consists of three essays on technology choice and spillovers.

Patent Scope and Technology Choice analyzes the effect of an increase in patent scope on investments in R&D and innovation. Patent scope affects incentives for innovation via the research strategies chosen by firms; a broad scope of the patent on the state-of-the-art technology can induce entrant firms to choose to do research on alternative technologies to avoid patent infringement. If the alternative technologies have a lower probability of success, this reduces incentives for investment in R&D by entrant firms and the probability that they innovate. On the other hand, the allocation of total R&D across projects is improved, since there is less wasteful duplication of R&D investments. I present a model where the trade-off induced by patent scope can be analyzed. The model predicts that an increase in patent scope can increase the probability of innovation if the incumbent's increase in profits from innovating is large, and the patented technology has a small advantage relative to the alternative technology. However, when the model is extended to allow for Stackelberg competition or license agreements, the benefit of a broad patent scope to a large extent disappears.

The World Distribution of Productivity: Country TFP Choice in a Nelson-Phelps Economy builds a theory of the shape of the distribution of total-factor productivity (TFP) across countries. The data on productivity suggests vast differences across countries, and it arguably even has "twin peaks". The theory proposed here is consistent with vast differences in long-run productivity, as well as with a twin-peaks outcome, even under the assumption that all countries are ex-ante identical. It is based on the hypothesis that TFP improvements in a given country follows a Nelson-Phelps specification. Thus, they derive from past investments in the country itself and, through a spillover (or catch-up) term, from past investments in other countries. We then construct a stochastic dynamic general equilibrium model of the world with externalities: each country invests in TFP and internalizes the dynamic effects of its own investment, while treating other countries' investments as given. We find that, in the long run, the world distribution of TFP across countries can be asymmetric, i.e., twin-peaked, or bimodal. More specifically, twin-peaked world

distributions of TFP arise if the catch-up term in the Nelson-Phelps equation has a sufficiently low weight. If, on the other hand, technological catch-up is important, the world distribution of TFP is single-peaked. Even in the latter case, however, small idiosyncratic TFP shocks can lead to large long-run differences in TFP levels.

Exhaustible Resources, Technology Choice and Industrialization of Developing Countries studies technology choice in a dynamic model with two technologies for production; one which uses an exhaustible resource and an alternative technology which does not. The main finding is that if the capital stock is large relative to the resource stock, the alternative technology is immediately adopted and the time path of resource extraction is decreasing. If, instead, the capital stock is small, the alternative technology is adopted with a delay and the time path of extraction is inverse-U shaped. I calibrate the model to the case of oil and analyze the effects of industrialization of developing countries on the extraction of oil and technology choice for energy production. Industrialization is modelled as an increase in labor supply. The model predicts that industrialization of developing countries has the following effects: the rate of oil extraction increases immediately, the alternative technology is adopted earlier, and oil as an energy source is abandoned earlier.

To Niclas

Acknowledgments

My Ph.D. studies have been an extraordinary experience filled with new knowledge, new understanding, plentiful challenges and meetings with gifted individuals. The end result, this thesis, could not have been achieved without the help of many other people. Some have helped me directly, by sharing their knowledge and experience with me and, equally important, others have helped me through their love and support. I would like to take the opportunity to express, to all of you, my deepest gratitude.

First and foremost, I want to thank my supervisor John Hassler for his guidance and his enthusiasm. He has been very generous with his time and our discussions have helped me to greatly improve the stringency of my arguments. In addition, I want to thank my coauthor, Per Krusell. It has been very inspiring and rewarding to work with him and his advice has been invaluable. I have also much appreciated our conversations on other essential topics, such as soccer and Karl-Bertil Jonsson. I am also much indebted to Fabrizio Zilibotti for providing encouragement and advice on the first essay of the thesis.

But, to start from the beginning of my Ph.D. studies, I owe special thanks to numerous people at the Stockholm School of Economics, where I spent my first two years of the Ph.D. program. I would like to thank Lars Bergman who encouraged me to apply to the program. I am also indebted to Martin Flodén for his advice. I wish to thank my fellow students from the cohort of 2003: Sebastian Giwa, Erik Lindqvist, Ola Melander, Damián Migueles, Martin Mossberg, Roine Vestman, and Robert Östling. All of you made the first year an enjoyable experience, in spite of all the hard work. I also want to thank the rest of the students for making the time in the office much brighter: Anna Breman, Harald Edquist, Ola Granström, Erik Höglin, Therese Lindahl, Henrik Lundvall, Per Sonnerby, Björn Tyrefors, and Fredrik Wilander, to mention a few.

This thesis has greatly benefited from the friendly and inspiring atmosphere at the Institute for International Economic Studies. I am indebted to Torsten Persson, Nicola Gennaioli, Ethan Kaplan, and David Strömberg for their helpful comments on my work. I owe special thanks to my two office-mates, Olle Folke and Mirco Tonin, and to Martina Björkman, Daria Finocchiaro, Anna Larsson, and Virginia Queijo von Heideken for interesting discussions and many laughs. I have also enjoyed the company of Martin Bech Holte, Dario Caldara, Anders Fredriksson, Ettore Panetti, Maria Perrotta, and Jose Mauricio Prado, to name a few. I am indebted to the administrative staff, in particular Annika Andreasson and Christina Lönnblad, for their help and support. In addition, I would like to thank all participants of Wednesday soccer for excellent games.

During my Ph.D. studies, I have had the privilege to spend a semester at the Massachusetts Institute of Technology, and I wish to thank the faculty and the Ph.D. students at the Department of Economics for an interesting and stimulating time.

Very special thanks go to all my friends outside academia who always cheer me up and remind me that there are many other things in life than economics.

I am deeply grateful to my parents and my brother for their love and support and for their interest in my research topics. Our conversations have given me a greater understanding of the reality I try to capture in my models. In addition, I wish to give the Damsgaard family my warmest thanks.

Finally, I would like to thank my husband Niclas for his love, for his patience in endless discussions over my papers, for always sharing my enthusiasm over progress as well as lessening my worries, and for his everlasting belief in me. I dedicate this thesis to him.

Stockholm, July 2008

Erika Färnstrand Damsgaard

Table of Contents

Chapter 1

Introduction

This thesis consists of three essays which are all related to technology; either to the choice of technology or to spillovers of technology within and across countries. The word technology originates from the Greek word technologia, which means systematic treatment of an art. Today, technology is a broad concept. It can refer to machines, hardware and software as well as systems and methods of organization. Furthermore, it encompasses the practical application of knowledge in a particular area, such as medical technology or information technology.

Among economists, it is generally acknowledged that the main source of long-term growth is new knowledge, generated by research and development. In particular, in the absence of technological progress, the diminishing returns to capital cause growth to cease. Technological progress makes capital and labor more productive and can offset the effect of diminishing returns. This is where new knowledge comes in: generated in a process of research and development, it is the driving force behind technological progress. Among policy-makers, a similar view on long-term growth has emerged. The so-called Lisbon Strategy, the action plan for the European Union launched in 2000, considers research and development as one of the main instruments for boosting growth and employment.

In this context, technology is the mechanism whereby knowledge and ideas are transformed into innovations, i.e. products and processes which generate increases in productivity, output growth and in the end lead to higher standards of living. Therefore, it is of both interest and value to analyze how this mechanism works. The first essay in this thesis investigates how technology choice is affected by institutions that govern incentives to conduct R&D, such as the patent system. The second essay investigates how differences in productivity among countries in the world depend on spillovers of technology within and across countries. In the third essay, I analyze how technology choice is affected by the exhaustibility of natural resources. Below, I give a summary of each essay.

The first essay *Patent Scope and Technology Choice* analyzes how an increase in patent scope affects investments in R&D and innovation, taking into account that patents affect firms' technology choices. It is widely perceived that the scope of intellectual property rights in the US has increased over the last two decades; see, for example, Jaffe (2000) and Gallini (2002). They point at two factors that suggest this to be the case. First, patent holders have been awarded greater power in infringement lawsuits by a broadening of the interpretation of patent claims. Second, patent protection has been extended to cover new areas, notably software, business methods and biotechnology, where a large number of patents with broad scope have been granted. In addition, empirical evidence suggests that patents affect firms' choices of research projects. For example, according to Walsh, Arora and Cohen (2003), firms direct R&D to areas less covered by patents, and Lerner (1995) finds that firms with high litigation costs tend to avoid research areas that are occupied by other firms.

There are several research strategies that firms can pursue in order to find the next generation product in a market. Either they conduct R&D using the patented state-of-the-art technology to make an improvement of that technology, or they choose alternative R&D strategies. The alternative strategies imply using different technologies than the current state-of-the-art. If the state-of-the-art technology is covered by a patent with broad scope, that may induce firms other than the patent holder to use a different technology to avoid the risk of patent infringement. It is probable that pursuing an alternative research strategy is more costly or involves more uncertainty than pursuing a strategy which makes improvements of the technology currently in use.

A broad patent scope on the state-of-the-art technology would therefore reduce the incentives for research by entrant firms and their innovation rates. However, research efforts may be better allocated across different potential projects. If many firms conduct R&D in order to develop the same technology, there may be wasteful duplication of research investment. Firms may, for example, build parallel labs and carry out identical experiments or build identical prototypes, which is a waste of R&D resources from a welfare point of view. If they conduct R&D using different technologies, they are less likely to carry out identical experiments and there is less wasteful duplication. In this essay, I construct a model to analyze this trade-off caused by patent scope.

The model suggests a new explanation for the empirical finding that incumbent firms have high innovation rates relative to entrants. In the standard R&D race models, the incumbent invests less than entrant firms due to the Arrow effect: the incumbent has a lower incentive to innovate since by innovating, he to some extent replaces his current profits. In this model, when the incumbent holds a patent which is broad in scope, it gives him a monopoly on the research which has the highest expected payoff and that increases his incentives to invest. Hence, if the incumbent owns a patent that is broad in scope, he can be more likely to innovate than the entrant.

The model predicts that if the incumbent firm has a high stand-alone incentive to innovate, i.e. the difference in his profits after versus before he innovates is large, and if the patented technology has a small advantage relative to the alternative technology, a broad patent scope gives a higher probability of innovation than a narrow scope. Hence, the negative effects of R&D duplication are under some conditions sufficiently large to warrant a broad patent scope. However, when the model is extended to allow for Stackelberg competition or licensing, the advantage of a broad scope to a large extent disappears. Hence, it is possible that an increase in patent scope increases the probability of innovation in a given industry. Nevertheless, it requires that specific conditions on the form of competition, the technological alternatives, the opportunities for license agreements etc. are met. According to this model, a uniform increase in patent scope, such as awarding patent holders larger powers in infringement lawsuits, cannot be an optimal policy.

The second essay *The World Distribution of Productivity: Country TFP Choice in a Nelson-Phelps Economy,* coauthored with Per Krusell, builds a theory of the shape of the distribution of total-factor productivity (TFP) across countries. This essay is motivated by the empirical fact that there are vast differences in income per capita across countries. Productivity accounting shows that after taking into account differences in observable factors (capital, quality-weighted labor), a large part of the differences in income per capita remains: differences in TFP. Thus, according to the accounting analyses, to understand differences in income, it is not sufficient to understand differences in accumulation of physical and human capital. Therefore, we propose a theory for explaining differences in TFP across countries.

We also take the point of view that a theory of the world distribution of TFP should display two features. One is that the long-run world TFP growth is endogenous. The second feature is that there are technology spillovers across countries. Whereas each country can influence its relative TFP level, its long-run growth of TFP is determined by the rate at which "world TFP", i.e. some average or the frontier TFP grows. One motivation for this view is that there have only been minor increases in the dispersion among countries during the period over which there is reliable income data. Another motivation is that modern economies are arguably highly dependent on world technological developments, as evidenced by, among others, Benhabib and Spiegel (1994), who find adoption of technologies from abroad, or *catch-up*, to be statistically significant for TFP growth.

Given this perspective, we build a theory of the world distribution of TFP that has at its core the specification of technology development proposed by Nelson and Phelps (1966). They formulated a form of technology, or human capital, catch-up. In a country context, the growth rate of technology, or human capital, in a given country can be increased if this country invests, and the further is the distance from the technology frontier, the more productive is such an investment. That is, if a country is further behind, the potential for rapid growth is higher, since the country can benefit from spillovers of technology generated elsewhere. We take this Nelson-Phelps view on development and ask a further question: if countries are subject to this "technology for productivity growth", and each country operates the technology optimally from the viewpoint of maximizing the utility of its citizens, what is the implied equilibrium world distribution of TFP? Since countries benefit from spillovers, there is an obvious force for convergence, but how does this force play out in equilibrium? Thus, we construct a stochastic dynamic general equilibrium model in which countries are ex-ante identical and average long-run growth is endogenous, as is the distribution of TFP across countries.

Our main findings can be described as follows. Although convergence forces are always present in the model, if the catch-up term is weak enough, the long-run distribution of TFP is not single-peaked, but bimodal (twin-peaked). There is one group of countries with high TFP in relative terms, with the remainder of countries operating at a much lower TFP level. However, even when there is a single peak in the TFP distribution, this distribution can show significant dispersion of countries. Thus, the theory embodies strong forces pulling countries apart. These forces are generated by the dynamic gains from TFP investment; each country internalizes the dynamic effects of its own TFP accumulation, through its own investment. Those countries with initially higher TFP levels, or those hit by positive shocks to TFP, have an advantage in further TFP growth.

Aside from providing a theory of relative TFP across countries, with conditions for bimodality, we use the theory to speculate on the evolution of the distribution of world productivity. Acemoglu (2008) plots a sequence of distributions of countries according to GDP per worker for 1960, 1980, and 2000. The plot reveals visible changes in the shape of the distribution. The 1980 and 2000 distributions are "more bimodal" than is the 1960 distribution. In terms of our model, a move toward bimodality can be due to, for example, a decrease in spillovers, a decrease in the cost of trading inputs into production of TFP, a change in the production function for TFP, or transitional dynamics. In the essay, we discuss and interpret each of these channels in turn.

The third essay *Exhaustible Resources, Technology Choice and Industrialization of Developing Countries* studies technology choice in the context of management of exhaustible resources. In recent years, the industrialization of large developing countries, such as China and India, has generated a considerable increase in world demand for exhaustible resources, for example copper, aluminum, iron ore and oil. In 2007, China and India accounted for about 35 percent of global steel consumption and China alone accounted for about one third of world consumption of aluminum. The increase in resource use has resulted in increases in the extraction of resources and through higher resource prices it has had an impact on resource-importing countries in the rest of the world. This development has contributed to the renewed interest in the management of exhaustible resources. At what rate should they be extracted? When will exhaustible resources be substituted for renewable resources and how will the transition take place?

In order to address such questions, this essay constructs a dynamic model of the world economy that exhibits two production technologies; a resource technology which uses an exhaustible resource as input, and an alternative technology which does not. Both technologies produce using capital and labor. In each time period, a social planner decides how much of the resource to extract, which technology or technologies to use, how to allocate capital and labor and how much to save. The model is first solved in a two-period setting and then extended to an infinite time horizon.

The main finding is that technology choice depends on the relative sizes of capital and resource stocks. If the capital stock is large in relation to the resource stock, the alternative technology is immediately adopted. The two technologies coexist until the resource is abandoned and there is a complete switch to the alternative technology. If, instead, the capital stock is small, only the resource technology is used initially and the alternative technology is adopted with a delay. In addition, the resource is abandoned at a later point in time. The intuition for this result is that if the exhaustible resource is used, it is optimal to allocate capital to the resource technology first, such that the resource-capital ratio is constant. Therefore, if the capital stock is small in relation to the amount of resource remaining, all capital will be allocated to the resource technology. Over time, as the resource stock decreases and savings increase the capital stock, it will eventually be optimal to adopt the alternative technology.

Similarly, the time path of resource extraction depends on the relative sizes of capital and resource stocks. If the capital stock is large in relation to the resource stock, resource extraction is decreasing over time. If, instead, the capital stock is small, resource extraction has the shape of an inverse U; it is first increasing and then decreasing. With a low initial capital stock, it is optimal to defer a larger part of the extraction to the future, when the capital stock has increased through savings and the amount of capital available to allocate to the resource technology is higher.

This essay also analyzes the effects of industrialization of developing countries on one of our most important exhaustible resources: oil. I calibrate the model to match the world production of crude oil and analyze the effects of industrialization of developing countries on the extraction of oil and technology choice for energy production. Industrialization is modeled as an increase in the supply of labor. The calibrated model predicts that industrialization of developing countries has the following effects on the world economy: the rate of oil extraction increases immediately, the alternative technology is adopted earlier, and oil as an energy source is abandoned earlier.

Bibliography

[1] Acemoglu, D. (2008), *Introduction to Modern Economic Growth*. Forthcoming at Princeton University Press.

[2] Benhabib, J. and Spiegel, M. (1994), "The Role of Human Capital in Economic Development: Evidence from Aggregate Cross-Country Data", *Journal of Monetary Economics*, Vol. 34, pp. 143-173.

[3] Gallini, N. (2002), "The Economics of Patents: Lessons from Recent U.S. Patent Reform", *Journal of Economic Perspectives*, Vol. 16, No. 2, pp. 131-154.

[4] Jaffe, A. (2000), "The U.S. patent system in transition: policy innovation and the innovation process", *Research Policy*, Vol. 29, pp. 531-557.

[5] Lerner, J. (1995), "Patenting in the Shadow of Competitors", *Journal of Law and Economics*,Vol. 38, No. 2, pp. 463-495.

[6] Nelson, R. and Phelps, E. (1966), "Investment in Humans, Technological Diffusion, and Economic Growth", *The American Economic Review*, Vol. 56, No. 1/2, pp. 69-75.

[7] Walsh, J. Arora, A. and Cohen, W. (2003), "Effects of Research Tool Patents and Licensing on Biomedical Innovation" in Patents in the Knowledge-Based Economy, Wesley M. Cohen and Stephen A. Merrill, *Editors*, Committee on Intellectual Property Rights in the Knowledge-Based Economy, National Research Council.

Chapter 2

Patent Scope and Technology Choice*

1 Introduction

It is widely perceived that the scope of intellectual property rights in the US has increased over the last two decades; see, for example, Jaffe (2000) and Gallini (2002). They point at two factors that suggest this to be the case. First, patent holders have been awarded greater power in infringement lawsuits by a broadening of the interpretation of patent claims. Second, patent protection has been extended to cover new areas, notably software, business methods and biotechnology, where a large number of patents with broad scope have been granted. The purpose of this paper is to analyze how an increase in patent scope affects investments in R&D and innovation.

There are several research strategies that firms can pursue in order to find the next generation product in a market. Either they conduct R&D using the patented state-of-the-art technology to make an improvement of that technology, or they choose alternative R&D strategies. The alternative strategies imply using different technologies than the current state-of-the-art. If the state-of-the-art technology is covered by a patent with broad scope, that may induce firms other than the patent

* I would like to thank Daron Acemoglu, Martin Bech Holte, Bengt Domeij, Ante Farm, John Hassler, Richard Jensen, Stephen Parente, Lars Persson, Fabrizio Zilibotti and seminar participants at IIES Brown Bag Seminar, MIT Brown Bag Seminar, ENTER Conference 2007, Foundation Urrutia Elejalde VII Winter Workshop on Economics and Philosophy 2007, IFN Stockholm Conference 2007, EPIP Conference 2007, REER 2007, and ZEW Conference on the Economics of Innovation and Patenting 2008 for valuable comments and suggestions. I would also like to thank Christina Lönnblad for editorial assistance. All remaining errors are my own.

holder to conduct R&D on an alternative technology to avoid the risk of patent infringement.[1] Lerner (1995) finds that firms with high litigation costs tend to avoid research areas that are occupied by other firms, particularly when these firms have low litigation costs. Walsh, Arora and Cohen (2003) analyze the effect of patents on R&D in the pharmaceutical industry. They find that firms tend to direct R&D investment to research areas less covered by patents.

It is probable that pursuing an alternative research strategy is more costly or involves more uncertainty than pursuing a strategy which makes improvements of the technology currently in use. A broad patent scope on the state-of-the-art technology would therefore reduce the incentives for research by entrant firms, and their innovation rates. However, research efforts may be better allocated across different potential projects. If many firms conduct R&D in order to develop the same technology, there may be wasteful duplication of research investment. Firms may, for example, build parallel labs and carry out identical experiments or build identical prototypes, which is a waste of R&D resources from a welfare point of view. If they conduct R&D using different technologies, they are less likely to carry out identical experiments and there is less wasteful duplication. Direct evidence of duplication of R&D is given by simultaneous innovation, which is common in science. Two examples discussed by Chatterjee and Evans (2004) are the parallel inventions of the first electronic mini-calculator by Casio and Texas Instruments in 1972, and the parallel discovery of the process for synthesis of leukotrienes by two competing research teams in 1979. Duplication of research effort which does not directly result in innovations is certainly more common.

In sum, an increase in patent scope can have several effects on investments in R&D and the rate of innovation in the economy. It reduces entrant firms' incentives for research and their innovation rates. At the same time, an increase in patent scope decreases duplication of research effort, and directs research towards potentially fruitful new technologies and methods. The question is whether an improvement in allocation across projects can offset a decrease in individual firms' innovation rates and, if so, under what conditions.

In this paper, I construct a model to analyze this trade-off caused by patent

[1] According to patent law, an invention which builds on a patented invention infringes that patent, even if it is patentable in its own right.

scope. For simplicity, I use a duopoly model with an incumbent firm and an entrant firm. The incumbent owns a patent connected to the state-of-the-art technology, and produces the corresponding product. There are two possible research strategies to follow: the first is to build on the state-of-the-art technology, and the second to use an alternative technology, which is less promising. I consider two alternative scenarios for patent scope: one in which the patent scope on the state-of-the-art technology is narrow, and one in which the patent scope is broad. In the case of a narrow scope, both firms can choose to conduct research on the state-of-the-art technology. In the case of a broad scope, the entrant has to choose the alternative, less promising technology in order to avoid infringing on the patent. I describe the possible equilibria that can arise under narrow and broad scope, respectively, and compare the resulting R&D investments and probabilities of innovation.

The model suggests a new explanation for the empirical finding that incumbent firms have high innovation rates relative to entrants. In the standard R&D race models, the incumbent invests less than entrant firms due to the Arrow effect: the incumbent has a lower incentive to innovate since by innovating, he to some extent replaces his current profits. In this model, when the incumbent firm holds a patent which is broad in scope, it gives him a monopoly on research that has the highest expected payoff. This effect can increase the incumbent's incentives for R&D sufficiently to outweigh the Arrow effect. Hence, the incumbent can be more likely to innovate than the entrant when he has an advantage originating from policy, namely the scope of the patent he owns.

The model predicts that if the incumbent firm has a high stand-alone incentive to innovate, i.e. the difference in his profits after versus before he innovates is large, and if the patented technology has a small advantage relative to the alternative technology, a broad patent scope gives a higher probability of innovation than a narrow scope. Hence, the negative effects of R&D duplication are under some conditions sufficiently large to warrant a broad patent scope. Conversely, when the incumbent's stand-alone incentive to innovate is low, or the patented technology has a large advantage, a narrow scope gives a higher probability of innovation. If the incumbent is able to commit to an investment level or if license agreements can be made, the first result is partly reversed; in instances where the highest innovation probability was previously given by a broad scope, it is now obtained under a narrow scope. Hence, the benefit of a broad patent scope largely relies on the assumptions that the firms act simultaneously and that there are no possibilities for license agreements.

The paper is organized as follows. The related literature is presented in Section 2 and an introduction to the determination of patent scope is given in Section 3. Section 4 describes the model, and characterizes the possible equilibria. Section 5 entails the investments and the probabilities of innovation resulting from the narrow and broad patent scope, respectively. Section 6 describes the conditions under which each patent regime gives the highest innovation probability and the highest social surplus. Section 7 extends the model to allow for Stackelberg competition and licensing. Section 8 concludes.

2 Related Literature

There is a large theoretical literature on the economic effects of intellectual property rights. An increasingly spreading view is that the current system of intellectual property rights in the US offers innovators too much protection of their innovations. Heller and Eisenberg (1998) argue that there is a "tragedy of the anticommons" in biomedical research as there are numerous patents to each separate building block for a new product. Acquiring the rights to use all of them is costly and potentially difficult, as the owners of the rights may have heterogeneous interests. Therefore, patenting can constitute an obstacle to future research. Similarly, Shapiro (2001) argues that in several industries, such as semiconductors and software, the current patent system is creating a patent thicket, an overlapping set of patents, which requires innovators of new technology to obtain licenses from multiple patent holders. The high transaction costs involved imply that stronger patent rights may stifle innovation. Bessen and Maskin (2002) show that when innovations are sequential, stronger intellectual property rights protection may reduce innovation even in the case when there is only one patent holder. On the other hand, Green and Scotchmer (1995) also present a model of sequential innovation and find that a broad patent scope can be necessary to give the first innovator sufficient incentives to invest.

In the law and economics literature, Kitch (1977) argues that pioneering technologies should be granted patents with broad scope, since it will allow the innovator to coordinate further development of the technology by granting licenses and thereby, wasteful duplication of effort is reduced. His view is challenged by Merges and Nelson (1990). Their argument is that uncertainty and high transaction costs of licensing reduce the effectiveness of coordination, and that broad patent scope can instead block technology development. Technical advance is likely to be faster when there is competition, as the patent holder has higher incentives to develop his technology. Domeij (2000) discusses the trade-off between total investment in R&D and duplication of investments induced by patent scope in the context of the pharmaceutical industry. When the second generation product is a new indication, i.e. the same compound is used to treat other types of illnesses, Domeij argues that the patent holder has a high incentive to search for new innovations, since they are intended for new markets. In addition, the patent holder has a technological advantage over competitors in finding this type of innovations. Consequently, he concludes that a broad patent scope is preferable.

In the literature on firms' choices of research strategies, Dasgupta and Maskin (1987) find that competition encourages firms to choose research projects that are too similar from a welfare point of view. Chatterjee and Evans (2004) show that if the projects differ in other dimensions than the probability of leading to an innovation, firms may either choose projects that are too similar, or projects that are too different relative to what is socially optimal. Previous literature on duplication of effort in research and development includes Tandon (1983), Jones and Williams (2000), and Zeira (2003). These works model identical firms and do not take into account the different incentives facing incumbent and entrant firms. Cabral and Polak (2004) present a duopoly model with an incumbent and an entrant. They investigate how an increase in consumer valuation of the incumbent firm's good, interpreted as an increase in its dominance, affects the amount of duplication of R&D by the two firms and the rate of innovation. Their conclusion is that increased dominance has a positive effect on innovation when intellectual property rights are strong. However, neither of the models has a mechanism by which entrant firms' technology choices affect the duplication of R&D.

This paper is also related to the literature concerned with why incumbent firms have high innovation rates relative to entrant firms. As shown by Reinganum (1983), an incumbent firm invests less in R&D than an entrant when the innovation process is stochastic, and is less likely to innovate. However, empirical evidence points to the opposite. For example, Blundell et al. (1999) find that within industries, firms with high market share innovate more. Several explanations for this observation have been proposed, most of them relying on a technological advantage for the incumbent. One example is Segerstrom and Zolnierek (1999) where the incumbent has lower R&D costs than entrant firms. Another is Etro (2004), where the explanation is a first mover advantage for the incumbent in combination with free entry.

3 Determination of patent scope

The scope of a patent is central to this analysis. Therefore, I will start with a brief introduction to the determination of patent scope in patent law and practice, as described in Merges and Nelson (1990). A patent application consists of a specification of the innovation and a set of claims. The specification is written as an engineering article and describes the problem the innovator faced, and how it was solved. The claims define what the inventor considers to be the scope of the innovation, the "technological territory" where he can sue other parties for infringement. The general rule is that a patent's claims should extend beyond the precise disclosure of the innovation in the specification. Otherwise, imitators could make minor changes to that example without infringing and the patent would be of little value. The inventor naturally wants to make the claims as broad as possible, and the patent examiner must decide what scope is appropriate, which claims should be admitted and which should not.

In infringement cases, the court first examines whether there is "literal infringement", namely the product literally falls within the boundaries of the patent claims. If not, the court also examines whether the product infringes under the doctrine of equivalents. The doctrine of equivalents says that a product is infringing if it does the work in substantially the same way and accomplishes substantially the same result as the patented product. Consequently, patent scope is determined in two instances, by two separate authorities. Ex ante, if the patent holder has not sued any other party for infringement, the patent scope is defined by the claims as determined by the Patent Office. Ex post, in an infringement case, the patent scope is determined by the court, in its decision on whether the patent has been infringed.

4 The Model

The economy has two firms, an incumbent and an entrant. Both firms make investments in R&D in order to find the next innovation, which has private value V when patented. Both firms have quadratic investment cost functions. The incumbent firm holds a patent connected to the current state-of-the-art technology, and earns a profit from producing the corresponding product. The profit is expressed as a share of the value of the next innovation, αV, where $\alpha \in [0, 1]$. The entrant earns no current profits. Innovation is drastic; new innovations replace previous ones.

There are two possible research strategies for a firm to pursue. Strategy C is to build on the current state-of-the-art technology, technology C, and make an improved product. Strategy A is to use an alternative technology, technology A, for which there is no risk of patent infringement. In this context, an alternative technology should be more broadly interpreted as using another material, algorithm, chemical compound etc., depending on the industry and the nature of the product. Irrespective of which technology is used in R&D, the private value of an innovation is V. There is no strong reason to believe that using different technologies to develop a certain product will generate innovations of exactly the same value. However, this simplification enables me to distinguish the effects of different patent regimes on total innovation from effects of a higher value of an innovation.

Each technology has an exogenous probability γ_k, $k \in \{C, A\}$, of leading to the next innovation. The alternative technology has a weakly lower probability of leading to the next innovation than the state-of-the-art: $\gamma_A \leq \gamma_C$. The difference between γ_C and γ_A reflects the relative advantage of the state-of-the-art technology. I assume that either technology C or A leads to a new innovation, but not both. This is a simplification of technology development, but it is made for tractability. I will discuss the implications of the assumption further below. In addition, I normalize the sum of γ_A and γ_C to 1, since it reduces the number of model parameters. Hence, $\gamma_A = 1 - \gamma_C$. This does not affect the main results, since what is important in the model is the ratio $\frac{\gamma_C}{\gamma_A}$.

In this paper, the R&D process is modeled as a one-shot game. This modeling choice is motivated by the fact that firms' R&D projects for development of new products are often close to irreversible. This is especially true in biotechnology and pharmaceutical industries. As a consequence of this structure, the model has a positive probability that both firms innovate if they choose the same technology on which to conduct R&D. It is necessary to specify the payoffs to both firms if this event occurs. Let the game be interpreted as a time period of five years, a period over which it is reasonable to assume that the R&D strategy cannot easily be changed. If both firms innovate during this period, a patent will be awarded to the firm which innovated first. Suppose that innovations arrive with a hazard rate that is constant over the period. Then, conditional on both firms having innovated at the end of the period, the firms have the same probability of innovating at each point in time. Therefore, I assume that each firm has probability $\frac{1}{2}$ of obtaining the patent if both innovate.

Further, I assume that the incumbent always chooses to invest in technology C, irrespective of the entrant's technology choice. A justification for this assumption is that using a particular technology requires a fixed cost or an investment in human capital.[2] In the baseline model, firms act simultaneously. In Section 7, the model is extended to Stackelberg competition.

The patent regimes are modeled as follows. The scope of the patent on the state-of-the-art technology can be either narrow or broad. In the case of a narrow patent scope, the entrant can choose between the two technologies when investing in R&D, and he selects the technology which gives the highest expected payoff. In the case of a broad patent scope, the incumbent can use his patent to block the entrant's innovation, if it is based on technology C. Therefore, the entrant automatically chooses technology A in order to avoid infringement. This assumption will be relaxed in Section 7, where the model is extended to allow for license agreements between the firms.

The possibility of duplication of R&D resources can be illustrated in terms of two urns, A and C, filled with marbles. Each urn corresponds to a technology. Suppose that a firm's investment in R&D can be described as drawing a number of marbles from one of the urns and then replacing them. Drawing one marble is equivalent to conducting one experiment. Each urn has its own set of marbles and the number of marbles is n_k, $k \in \{C, A\}$. Only one marble corresponds to a successful experiment, i.e. an innovation, and this marble is denoted 1.

[2] The assumption rules out an equilibrium in which the incumbent would choose to abandon his patented technology and invest in an alternative technology that is ex ante less attractive, only in order to escape competition from the entrant.

With probability γ_C, marble number 1 is in urn C. Firm j purchases t_j, $j \in \{I, E\}$ marbles from urn k and the probability that it innovates, conditional on having chosen the right urn, is $\frac{t_j}{n_k}$. Firm j can increase this probability by buying more marbles at the per marble price. The draws of different firms are independent events.

Suppose first that the incumbent and the entrant both choose urn C. The incumbent draws t_I marbles and replaces them in the urn, which gives him an innovation probability $\frac{t_I}{n_C}$. Then, the entrant draws t_E marbles, resulting in an innovation probability $\frac{t_E}{n_C}$. It is possible that both firms draw the same marble, that is, conduct the same experiment. This is a duplication of R&D resources from the point of view of society. No individual firm draws the same marble twice and there are no duplicate experiments at the firm level. However, a social planner is interested in the probability of any of the firms drawing marble number 1. For two events, A and B, the probability of at least one event occurring as:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

If the two events are independent, $P(A \cap B) = P(A)P(B)$. In our example, the probability of at least one innovation is:

$$P(no\ 1\ at\ least\ once) = \gamma_C \left(\frac{t_I}{n_C} + \frac{t_E}{n_C} - \frac{t_I}{n_C}\frac{t_E}{n_C} \right).$$

The product $\frac{t_I}{n_C}\frac{t_E}{n_C}$ represents a waste of resources due to duplication.

Now, suppose instead that the incumbent draws marbles from urn C and the entrant draws marbles from urn A. The incumbent and the entrant have probabilities $\frac{t_I}{n_C}$ and $\frac{t_E}{n_A}$, respectively, of drawing marble number 1, conditional on choosing the right urn. The probability that both firms draw the same marble is zero. Hence, the probability of at least one innovation is:

$$P(no\ 1\ at\ least\ once) = \gamma_C \frac{t_I}{n_C} + (1 - \gamma_C)\frac{t_E}{n_A}.$$

There is no waste of resources due to duplication. Next, I turn to the characterization of the equilibrium investments in R&D.

4.1 Equilibrium investments

In this model, the incumbent always invests in his patented technology when he conducts R&D. He invests the amount of resources, p_I, which maximizes his expected payoff Π_I, where subscript I denotes incumbent. The entrant, on the other hand, chooses both which technology to invest in, denoted k, and the level of investment, p_E, which maximizes his expected payoff Π_E, where subscript E denotes entrant. Each firm's investment translates directly into its probability of innovating. The timing of the game is as follows: First, the entrant chooses which technology to invest in. Second, given the entrant's technology choice, both firms simultaneously decide how much to invest. An equilibrium is a triplet $\{k^*, p_I^*, p_E^*\}$, $k \in \{C, A\}$ and $p_I, p_E \in [0,1]$ such that $k^* = \arg\max_k \Pi_E(k, p_I^*(k), p_E^*(k))$, $p_E^*(k^*) = \arg\max_{p_E} \Pi_E(k^*, p_I^*(k^*), p_E)$ and $p_I^*(k^*) = \arg\max_{p_I} \Pi_I(k^*, p_I, p_E^*(k^*))$. I divide the equilibria into two types, given the entrant's choice of technology:

- If the entrant chooses C, the equilibrium is of type C
- If the entrant chooses A, the equilibrium is of type A

First, the investments in equilibrium of type C are characterized and after that, the investments in equilibrium of type A. In order to interpret the firms' investments as probabilities of innovation, each investment must be bounded above by 1. I focus on the case when the optimal investment levels by both firms are interior solutions. In the baseline model, this is achieved by setting V equal to 1. The effects of varying V will be analyzed in Section 6.

4.2 Equilibrium of type C

The expected payoff to the incumbent when both firms choose technology C is

$$\begin{aligned}\Pi_I(C, p_I, p_E) &= \alpha V + \gamma_C p_I(1-p_E)(V-\alpha V) + \gamma_C p_E(1-p_I)(0-\alpha V) \\ &\quad + \gamma_C p_E p_I \left(\frac{1}{2}(V-\alpha V) + \frac{1}{2}(0-\alpha V)\right) - \frac{(p_I)^2}{2}. \qquad (2.1)\end{aligned}$$

With probability $\gamma_C p_I(1-p_E)$, the incumbent innovates whereas the entrant does not. The gain is $V(1-\alpha)$, the value of the innovation net of current profit, since the new product replaces the old one. Following Katz and Shapiro (1987), I refer to

$V(1-\alpha)$ as the incumbent's stand-alone incentive to innovate, i.e. the difference in profit after versus before he innovates if he believes that his rival will not innovate. With probability $\gamma_C p_E(1-p_I)$ the entrant innovates but not the incumbent, and the latter loses his current profits. With probability $\gamma_C p_E p_I$ both firms innovate, in which case the incumbent has probability $\frac{1}{2}$ of obtaining the patent. The variable cost of R&D is $\frac{(p_I)^2}{2}$. The first-order condition yields

$$p_I = \gamma_C V \left(1 - \alpha - p_E \left(\frac{1}{2} - \alpha\right)\right).$$

The incumbent's investment p_I is increasing in γ_C and in $V(1-\alpha)$. It can be increasing or decreasing in p_E, depending on the value of α. There are two opposing forces at work: a higher investment by the entrant reduces the probability that the incumbent wins the patent, given his own investment, which decreases his incentives to invest in order to win. At the same time, a higher investment by the entrant increases the probability that the entrant wins. This increases the incumbent's returns to investing in order not to lose current profit and to increase the probability that both innovate, which increases his expected payoff by $\frac{V}{2}$. This effect increases the incumbent's incentive to invest. When $\alpha > \frac{1}{2}$, current profits are high relative to the value of innovation and the expected payoff from winning is low. The latter effect dominates. When, $\alpha < \frac{1}{2}$, current profits are low relative to the value of innovation, and the first effect dominates. The cutoff point is at $\alpha = \frac{1}{2}$, which follows from the assumption that if both firms innovate, each firm has probability $\frac{1}{2}$ of obtaining the patent.

The expected payoff to the entrant when both firms choose technology C is

$$\Pi_E\left(C, p_I, p_E\right) = \gamma_C p_E(1-p_I)V + \gamma_C p_E p_I \frac{1}{2} V - \frac{(p_E)^2}{2}. \tag{2.2}$$

With probability $\gamma_C p_E(1-p_I)$, the entrant wins V. With probability $\gamma_C p_E p_I$ both firms innovate, in which case the entrant gets V with probability $\frac{1}{2}$. The first-order condition yields

$$p_E = \gamma_C V \left(1 - \frac{p_I}{2}\right). \tag{2.3}$$

The entrant's investment is decreasing in the incumbent's investment, for all parameter values. Solving for Nash equilibrium investment levels, given $V = 1$, yields the following investment by the incumbent and the entrant, respectively

$$p_I^C(\alpha, \gamma_C) = \frac{2\gamma_C\left(2(1-\alpha) + 2\alpha\gamma_C - \gamma_C\right)}{\left(4 + 2\alpha\gamma_C^2 - \gamma_C^2\right)} \tag{2.4}$$

$$p_E^C(\alpha, \gamma_C) = \frac{2\gamma_C\left(2 + \alpha\gamma_C - \gamma_C\right)}{\left(4 + 2\alpha\gamma_C^2 - \gamma_C^2\right)}, \tag{2.5}$$

where superscript C indicates that the equilibrium is of type C. The incumbent's equilibrium investment $p_I^C(\alpha, \gamma_C)$ is increasing in γ_C and decreasing in α. The entrant's equilibrium investment $p_E^C(\alpha, \gamma_C)$ is increasing in γ_C and α. An increase in γ_C implies a higher probability that technology C leads to the next innovation, which increases both firms' investments. An increase in α decreases the incumbent's stand-alone incentive to innovate, $V(1-\alpha)$, which reduces his investment. The entrant responds to this reduction by increasing his investment.

As long as $\alpha > 0$, $p_I^C(\alpha, \gamma_C) < p_E^C(\alpha, \gamma_C)$. The fact that the incumbent stands to lose current profit from innovating, while the entrant does not, implies that in equilibrium, the incumbent invests less. This is the Arrow effect. When the two firms invest in the same technologies, innovation is characterized by leapfrogging, i.e. the incumbent is less likely than the entrant to be the next innovator.

4.3 Equilibrium of type A

The expected payoff to the incumbent when the entrant chooses technology A is

$$\Pi_I\left(A, p_I, p_E\right) = \alpha V + \gamma_C p_I(V - \alpha V) + (1-\gamma_C)p_E(0 - \alpha V) - \frac{\left(p_I\right)^2}{2}.$$

The assumption that one of the technologies leads to innovation, but not both, implies that the incumbent's optimal investment is independent of that of the entrant. Taking the first-order condition, given $V = 1$, yields

$$p_I^A(\alpha, \gamma_C) = \gamma_C\left(1 - \alpha\right)$$

where superscript A indicates that the equilibrium is of type A. The expected payoff to the entrant when he chooses technology A is

$$\Pi_E\left(A, p_I, p_E\right) = (1-\gamma_C)p_E V - \frac{(p_E)^2}{2}.$$

Taking the first-order condition, given $V = 1$, yields

$$p_E^A(\gamma_C) = (1-\gamma_C).$$

As above, the entrant's optimal investment is independent of the investment by the incumbent.

In this equilibrium, the incumbent invests in a technology that is more likely to lead to the next innovation, which increases his incentives to invest relative to those of the entrant. If this effect is sufficiently strong, it can dominate the Arrow effect. If the following condition holds

$$\gamma_C > \frac{1}{2-\alpha} \tag{2.6}$$

the incumbent is more likely to innovate than the entrant. The threshold value for γ_C defined by (2.6) is increasing in α and takes values in the interval $\left(\frac{1}{2}, 1\right)$. A lower stand-alone incentive for the incumbent to innovate implies that for the incumbent to be more likely to innovate, a higher probability of success for technology C is required.

4.4 The entrant's choice of technology

Let us return to the entrant's decision of in which technology to invest. The entrant chooses the technology that gives the highest expected payoff, given the equilibrium investments described above. The condition for when choosing C has a higher expected payoff than choosing A is given below.

Proposition 1 *If $\alpha > \bar{\alpha}$, the Nash equilibrium is of type C, where*

$$\bar{\alpha} = \frac{4 - 8\gamma_C + \gamma_C^2 + \gamma_C^3}{2\gamma_C^3}.$$

Proof. See Appendix A1. ■

The higher is α, the lower will the incumbent's investment be, which increases the entrant's expected payoff from choosing C relative to A. The threshold $\bar{\alpha}$ is decreasing in γ_C, since a larger probability of success for technology C increases the entrant's relative expected payoff from choosing C. [3]

5 Patent scope

In this model, the scope of a patent can either be narrow or broad. Patent scope is defined such that if the scope of the patent connected to technology C is narrow, the entrant can choose between technology C and A and hence, the possible types of equilibria are both C and A. If the patent scope is broad, the entrant has to choose technology A in order to avoid patent infringement, and the equilibrium is always of type A. Consequently, the patent scope determines which strategies are available to the entrant.

I assume that the patent scope does not affect V, the private value of the innovation, or α, the incumbent's current profit relative to V. It may be argued that a broad scope can increase the current profits accruing to the patent holder as it discourages the development of substitutes during the patented product's life. This effect would reduce the incumbent's investment in R&D under a broad scope relative to a narrow scope. The assumption that α is independent of scope gives an upper bound to the incumbent's investment under a broad scope. It may also be argued that a firm's expectation of patent scope affects the expected value of innovating. In a dynamic model, V would correspond to the present discounted value of future profits, and if future patents are expected to be broad in scope, that may translate into a higher V, given expectations of γ_C and α. Hence, expectations of patent scope can affect firms' investments independently of their technology choices but to assess the magnitude of this effect, a dynamic model is required. In this paper, I abstract from the potential effects of patent scope on innovation through current profit and expectations of future scope, and analyze the effect through technology choice alone. Nevertheless, as a robustness check I also allow V to take a higher, exogenously given value under a broad relative to a narrow scope, to assess the impact on the model's predictions. The result is reported in Section 6.

[3] I assume that if indifferent, the entrant chooses technology A.

In the case of a narrow patent scope, the entrant will choose C if $\alpha > \bar{\alpha}$. If not, patent scope is irrelevant for the entrant's technology choice, as he chooses technology A under a narrow as well as under a broad patent scope. Therefore, the comparison of investment and innovation probabilities under differing patent scope is meaningful only under the condition $\alpha > \bar{\alpha}$.

5.1 Narrow patent scope

I start with a characterization of the investments by the two firms and the aggregate innovation probability under a narrow patent scope. Suppose that $\alpha > \bar{\alpha}$ so that the equilibrium is C. In this type of equilibrium, the entrant is more likely to innovate than the incumbent and there will be leapfrogging, as in the standard stochastic racing and endogenous growth models.

The aggregate innovation probability is defined as the probability of at least one firm innovating. When both the entrant and the incumbent invest in the same technology, there is a duplication of R&D investment and, in analogy with the example in Section 4, the innovation probability is:

$$i^N = \gamma_C \left[p_I^C(\alpha, \gamma_C) + p_E^C(\alpha, \gamma_C) - p_I^C(\alpha, \gamma_C) p_E^C(\alpha, \gamma_C) \right], \tag{2.7}$$

where superscript N denotes narrow patent scope. The amount of duplication for a given total investment is highest when the two firms' investments are equal, and it decreases as investments become more asymmetric. It reflects the fact that no firm duplicates its own research, but the higher is the potential overlap in experiments with the other firm, the higher is the probability of duplication. The innovation probability i^N is increasing in γ_C. Inspection of $\frac{\partial i^N}{\partial \alpha}$ shows that i^N is increasing in α if

$$\alpha > \frac{(\gamma_C - 2)^2}{2\gamma_C^2},$$

and decreasing otherwise. An increase in α decreases the incumbent's investment, and increases that of the entrant. The net effect is a decrease in total investment, but also a decrease in duplication as the investments become more unequal. When α is sufficiently high, the reduction in total investment is offset by the decrease in duplication.

5.2 Broad patent scope

Now, I turn to a characterization of investments and the innovation probability under a broad patent scope. A broad patent scope implies that firms are in an equilibrium of type A. The fact that the incumbent has a monopoly on the more promising technology, given by the broad scope of the patent, provides him with an additional incentive to invest. If the patented technology has a sufficiently large advantage relative to the alternative, γ_C is sufficiently high, the incumbent is more likely to innovate than the entrant. Under a narrow scope, in contrast, the entrant is always more likely to innovate, irrespective of the value of γ_C. As described above, Etro (2004) explains the empirical pattern of innovation by incumbents with a first mover advantage for the incumbent, and Segerstrom and Zolnierek (1999), among others, with a technological advantage. This paper suggests an additional source of advantage for the incumbent resulting from policy, namely the scope of his patent.

The innovation probability under a broad patent scope is

$$i^B = \gamma_C p_I^A(\alpha, \gamma_C) + (1-\gamma_C) p_E^A(\gamma_C), \tag{2.8}$$

where superscript B denotes broad patent scope. Note that there is no duplication of R&D. It follows that i^B is decreasing in α and increasing in γ_C if (2.6) holds.

6 Effects of patent scope on innovation

The previous section characterized the economy's innovation probability under the two patent regimes; a narrow and a broad scope, respectively. The effects of differing patent scope on innovation can be analyzed by comparing the ratio of the resulting innovation probabilities

$$\frac{i^N}{i^B} = \frac{\gamma_C \left[p_I^C(\alpha,\gamma_C) + p_E^C(\alpha,\gamma_C) - p_I^C(\alpha,\gamma_C)\, p_E^C(\alpha,\gamma_C)\right]}{\gamma_C p_I^A(\alpha,\gamma_C) + (1-\gamma_C) p_E^A(\gamma_C)}. \tag{2.9}$$

First, I describe how $\frac{i^N}{i^B}$ varies with the two key parameters in the model: α and γ_C.

Proposition 2 *$\frac{i^N}{i^B}$ is increasing in α.*

Proof. See Appendix A1. ■

If the incumbent has a low stand-alone incentive to innovate, α is high, the benefit of introducing competition in R&D on technology C is large. The entrant invests more in case he gets access to this technology. In addition, the small investment by the incumbent relative to that of the entrant implies a low amount of duplication.

Proposition 3 *For all* $\alpha \in (0,1)$ $\arg\max_{\gamma_C}\left(\frac{i^N}{i^B}\right) < 1$.

Proof. See Appendix A1. ■

The ratio $\frac{i^N}{i^B}$ takes its highest value for $\gamma_C < 1$. The numerical solution shows that $\frac{i^N}{i^B}$ has an inverted U-shape over γ_C for all values of $\alpha \in (0,1)$. One might have thought that the largest gain from a narrow scope would be obtained when the patented technology leads to the next innovation with probability 1, that is when there are no expected gains from conducting research on an alternative technology. The intuition for this result is that for γ_C close to 1, an increase in γ_C increases total investment under both a narrow and a broad scope, but under a narrow scope there is a high degree of duplication. The duplication effect implies that $\frac{i^N}{i^B}$ decreases.

Now, I return to the assumption that both technologies cannot simultaneously lead to an innovation. This assumption is made for tractability, rather than as a description of reality. Relaxing the assumption will have the following implication for the results. Given a positive probability that both firms find an innovation when they invest in different technologies, there is now strategic interaction between the two firms in equilibrium of type A, which reduces the level of total investment in that equilibrium. Suppose that both firms find an innovation. Each firm can obtain a patent and if they can collude, each firm will earn $\frac{V}{2}$ from selling its innovation. If not, the expected gains from innovation are lower, which further reduces the level of investment in equilibrium of type A. Consequently, the assumption does not affect the main results of the model, but introduces a level effect on the investments under a broad scope.

6.1 Does a broad scope give a higher probability of innovation?

To assess the effects on innovation of an increase in patent scope, it is instructive to return to the trade-off between total investment in R&D and the allocation of investment. A narrow patent scope allows both firms to do research on the most promising technology, but gives rise to duplication of R&D. This effect decreases the numerator of $\frac{i^N}{i^B}$. A broad patent scope forces the entrant to do research that is ex ante less promising and he has a lower probability of innovation, which decreases the denominator of $\frac{i^N}{i^B}$. To answer the question: does a broader scope give a higher probability of innovation?, it remains to determine which effect dominates and under what conditions. That is, when is $i^B > i^N$ and vice versa? I solve for the innovation probabilities for all values of $\gamma_C \in [0.5, 1]$ and $\alpha \in [0, 1]$ and the result is shown in Figure 2.1. In the figure, the area labeled 1 is the area in which the entrant chooses equilibrium A even under a narrow scope and the patent scope has no effect on the innovation probabilities. The area labeled 2 is the one in which the broad scope gives the highest probability of innovation, whereas area 3 is the one in which a narrow scope gives the highest probability of innovation. Figure 2.1 shows that a broad scope gives a higher innovation probability for low values of γ_C and α, that is when the patented technology has a small advantage relative to the alternative and the incumbent's stand-alone incentive to innovate is high. When technology C has a small advantage, the entrant does not reduce his investment to any considerable extent, if forced to conduct R&D on technology A. Consequently, a broad patent scope gives a higher probability of innovation. When the incumbent has a high stand-alone incentive to innovate, the amount of duplication under a narrow scope is high and a broad patent scope gives a higher probability of innovation. As seen in Figure 2.1, a narrow patent scope gives a higher probability of innovation for a lion's share of the parameter space.

6.2 Social surplus

The previous section shows under what conditions a broad and a narrow patent scope, respectively, give the highest probability of innovation. However, maximizing the probability of innovation is desirable only insofar as it is also socially optimal. In addition to the duplication effect, a social planner must take two other effects of R&D into account when choosing patent scope. The first effect is the social value of innovation, which is typically considered to be larger than the private value. One reason is that creation of new knowledge generates spillovers across sectors in the economy and across time. The second effect is the business stealing effect; entrant firms do not take into account the fact that as they innovate, the incumbent's profit is lost. In order to analyze which patent scope is socially optimal in this model, I define the social surplus under a narrow and broad scope as s^N and s^B, respectively. I assume that the private value of innovation is proportional to the social value. In addition, the social value of the new innovation is S and the social value of the current innovation is αS. The increase in social value from innovation is then $S(1-\alpha)$, when accounting for the business stealing effect, and it comes at a cost equal to the sum of the two firms' investment costs. The ratio of social surpluses is

$$\frac{s^N}{s^B} = \frac{i^N S(1-\alpha) - \frac{\left(p_I^C(\alpha,\gamma_C)\right)^2 + \left(p_E^C(\alpha,\gamma_C)\right)^2}{2}}{i^B S(1-\alpha) - \frac{\left(p_I^A(\alpha,\gamma_C)\right)^2 + \left(p_E^A(\gamma_C)\right)^2}{2}}.$$

Suppose that the social value is five times larger than the private value of innovation; $S = 5V$. I solve numerically for the social surpluses to investigate when $\frac{s^N}{s^B} > 1$. The result is shown in Figure 2.2. In the comparison of social surpluses in Figure 2.2, it is notable that for most values of α, the patent scope which maximizes the probability of innovation is also the scope that is socially optimal. However, when α is close to 1, a broad scope, which gives the lowest probability of innovation in this parameter range, gives the highest surplus. The reason is that the innovation generates such a small increase in social value that it is optimal to restrict the investments in R&D. The level of α above which restricting investments is optimal depends on the ratio of social to private value of innovation, which in this example was set to 5. If the ratio is sufficiently large, restricting investment will never be optimal. The tentative conclusion is that the socially optimal patent scope is that which maximizes the probability of innovation, except when the increase in social value from the innovation is small.

6.3 Effects of varying the private value of innovation

In the baseline model, I have set $V = 1$ to ensure that equilibrium investments are bounded above by 1. This precludes any analysis of the effects of varying the private value of innovation. Now, I allow for corner solutions where p_I and p_E equal 1, and analyze the effects of an increase in V. It is still assumed that both technologies give rise to innovations of equal value. The result is that an increase in the value of innovation has two effects. First, compared to Figure 2.1, it increases the area of parameter space for which patent scope is inconsequential. The intuition for this result is that an increase in V increases the incumbent's investment, which decreases the entrant's payoff in equilibrium of type C but not A. Second, it increases the area of parameter space for which a broad scope gives a higher innovation probability than a narrow scope. The reason is that an increase in V increases the investment by both firms, but under a narrow scope, there is also an increase in the amount of duplication.

As a robustness check, I also allow V to take a higher, exogenously given value under a broad relative to a narrow patent scope. If firms expect the value of innovation to be higher under a broad scope, this increases incentives to invest under a broad relative to a narrow scope, and a decrease in $\frac{i^N}{i^B}$ may be expected. Let $V^B = \theta V^N, \theta > 1$. First, the model is solved for $\theta = 1.5$; firms expect that a broad patent scope increases the value of innovation by 50 percent. The result is an increase in the area of parameter space for which a broad scope gives a higher innovation probability than a narrow scope, as compared to Figure 2.1. However, it is still the case that a broad patent scope gives the highest innovation probability for *less* than half the total area of parameter space spanned by α and γ_C. The model is also solved for $\theta = 2$; firms expect the broad patent scope to double the value of innovation. Nevertheless, a broad patent scope gives the highest innovation probability only for *less* than two thirds of the total area of the parameter space.

7 Extensions of the model

Until now, it has been assumed that the firms simultaneously decide how much to invest. In addition, it has been assumed that the entrant cannot enter a license agreement with the incumbent in case of infringement on the incumbent's patent. However, many industries are characterized by precommitment in R&D investment or license agreements among firms. Therefore, it is important to investigate if, and how, the effects of an increase in patent scope depend on these assumptions. In this section, each of the two assumptions will be relaxed in turn.

7.1 Stackelberg competition

Suppose that the incumbent can commit to an investment in R&D. For example, he builds a new research lab or employs researchers. The incumbent then acts as a Stackelberg leader. The entrant observes the incumbent's investment and then decides which technology to invest in and how much to invest. In the equilibrium of type C, the firms' optimal investments are dependent on each other. If the incumbent is a Stackelberg leader, he can affect the entrant's optimal investment level. In addition, the incumbent can affect the entrant's technology choice. If the incumbent's investment is sufficiently large, the entrant will get a higher expected payoff from choosing technology A than from choosing C. Hence, by sufficient overinvestment, the incumbent can keep the entrant out of technology C. When the equilibrium is of type A, the firms' investments are independent of each other. Hence, unlike in equilibrium of type C, the incumbent is not able to affect the entrant's optimal investment level in this equilibrium by moving first. The entrant optimally invests $p_E^A(\gamma_C)$ under both Stackelberg competition and simultaneous moves.

7.1.1 Investments in equilibrium of type C

When the incumbent invests first, he takes into account the entrant's optimal response to his investment. Since the entrant's investment is no longer taken as given, there is an additional effect of the incumbent's investment on his own expected payoff. As shown in (2.3), the entrant's investment is decreasing in the investment by the incumbent. This implies that by investing more, the incumbent does not only increase his probability of winning, but also indirectly decreases the entrant's probability of winning as the entrant is induced to invest less.

Let the optimal investments by the two firms in equilibrium of type C be $p_{I,S}$ and $p_{E,S}$ where subscript S denotes Stackelberg competition. To find the optimal investment by the incumbent, I insert (2.3) into (2.1). Taking the first-order condition yields

$$p_{I,S}(\alpha, \gamma_C) = \frac{\gamma_C \left(1 - \alpha - \frac{1}{2}\gamma_C + \frac{3}{2}\alpha\gamma_C\right)}{\left(1 - \frac{1}{2}\gamma_C^2 + \alpha\gamma_C^2\right)}. \tag{2.10}$$

The optimal investment by the entrant is (as given by (2.10) and (2.3))

$$p_{E,S}(\alpha, \gamma_C) = \frac{\gamma_C \left(2\alpha\gamma_C - 2\gamma_C - \gamma_C^2 + \alpha\gamma_C^2 + 4\right)}{2\left(2\alpha\gamma_C^2 - \gamma_C^2 + 2\right)}.$$

Since the entrant's investment is decreasing in the incumbent's investment, the following holds.

Proposition 4 *For all* $\alpha \in [0,1]$ *and all* $\gamma_C \in \left[\frac{1}{2}, 1\right]$, $p_{I,S}(\alpha, \gamma_C) > p_I^C(\alpha, \gamma_C)$.

Proof. See Appendix A1. ■

In an equilibrium of type C, the incumbent always invests more if he is a Stackelberg leader, than if the two firms move simultaneously.

7.1.2 The entrant's choice of technology

The level of investment by the incumbent which induces the entrant to choose technology A is denoted $\bar{p}_I$ and can be expressed as

$$\bar{p}_I(\gamma_C) = \frac{2(2\gamma_C - 1)}{\gamma_C}. \tag{2.11}$$

The investment level $\bar{p}_I(\gamma_C)$ is increasing in γ_C. The higher is the relative advantage of technology C, the larger is the investment required to keep the entrant out of it. Note that if $\gamma_C > \frac{2}{3}$, not even the maximal investment by the incumbent, $\bar{p}_I(\gamma_C) = 1$, can prevent the entrant from choosing technology C under a narrow patent scope and the equilibrium is always C.

7.1.3 Equilibria under narrow patent scope

Suppose that the patent scope is narrow. In order to establish which equilibrium arises under Stackelberg competition, it is necessary to determine which investment level by the incumbent gives him the highest expected payoff $\Pi_I(k, p_I, p_E)$, given the entrant's optimal response to that investment level, both as regards the technology choice and investment level.

I define a threshold $\bar{\alpha}_S \in (0,1)$, where S denotes Stackelberg competition, such that the incumbent's payoffs in the two types of equilibria are equal:

$$\Pi_I(A, p_I(\bar{\alpha}_S, \gamma_C), p_E^A(\gamma_C)) = \Pi_I(C, p_{I,S}(\bar{\alpha}_S, \gamma_C), p_{E,S}(\bar{\alpha}_S, \gamma_C)).$$

Using $\bar{\alpha}_S$, it is possible to show the following:

Proposition 5 *If $\alpha < \bar{\alpha}_S$ the equilibrium is of type A, and if $\alpha \geq \bar{\alpha}_S$ the equilibrium is of type C.*

Proof. See Appendix A1. ■

If $\alpha < \bar{\alpha}_S$, the incumbent will choose to strategically overinvest and thereby, he induces the entrant to choose to conduct R&D on technology A. If $\alpha \geq \bar{\alpha}_S$ the incumbent finds it optimal not to strategically overinvest, and the entrant chooses technology C. Figures 2.3 and 2.4 depict the incumbent's expected payoff Π_I as a function of his investment in equilibrium candidates C and A. The vertical line indicates the threshold level of investment $\bar{p}_I(\gamma_C)$, at which the entrant chooses technology A. Figure 2.3 displays the case when $\alpha < \bar{\alpha}_S$. Even though investment is not at its optimal level in A, the expected payoff in A is larger than in C and the incumbent prefers A. Consequently, he will strategically overinvest such that the entrant chooses technology A. Figure 2.4 displays the case when $\alpha > \bar{\alpha}_S$. As α increases, the incumbent's incentives for innovation decrease and he prefers to invest less. However, it is only in equilibrium of type C that he can reduce his investment, since he must invest at least $\bar{p}_I(\gamma_C)$ in equilibrium of type A. Hence, the payoff of choosing A relative to C decreases. For α above the threshold $\bar{\alpha}_S$, the incumbent has a higher expected payoff in equilibrium of type C and will induce the entrant to choose C.

7.1.4 Effects of patent scope

If the patent scope is broad, the Stackelberg competition has no effect on equilibrium investments, since the equilibrium is of type A. The innovation probability is identical to that under a broad scope with simultaneous moves, given by (2.8).

If the patent scope is narrow, the probability of innovation depends on which type of equilibrium firms are in. Let $i^{N,S}$ denote the innovation probability under a narrow scope. If the equilibrium is of type A, there is no duplication of R&D, and hence the only difference between the patent regimes is that under a narrow scope, the incumbent strategically overinvests, and $\bar{p}_I(\gamma_C) > p_I^A(\alpha, \gamma_C)$.[4] Hence, total investment in R&D is higher under a narrow scope and it follows that $i^{N,S} > i^B$. If

[4] If $\bar{p}_I(\gamma_C) \leq p_I^A(\alpha, \gamma_C)$, the entrant chooses technology A even under a narrow patent scope and therefore patent scope has no effect on innovation probabilities.

the equilibrium is of type C, the innovation probability is

$$i^{N,S} = \gamma_C \left[p_{I,S}(\alpha, \gamma_C) + p_{E,S}(\alpha, \gamma_C) - p_{I,S}(\alpha, \gamma_C) p_{E,S}(\alpha, \gamma_C) \right].$$

In this equilibrium, there is duplication of R&D, as under simultaneous moves. Which patent scope gives the highest innovation probability depends on α and γ_C, as seen in Figure 2.5.

Figure 2.5 shows that a broad scope gives a higher probability of innovation for a small subset of parameter space, denoted area 2. This holds for values of α close to zero, and values of γ_C close to 0.7. The vertical line gives the threshold $\gamma_C = \frac{2}{3}$. In area 2, the incumbent is not able to overinvest and keep the entrant out since $\gamma_C > \frac{2}{3}$, and the equilibrium is C. However, the incumbent can still affect the entrant's optimal investment level, and the first mover advantage implies that the incumbent invests more, and the entrant less, relative to the case of simultaneous moves. Now, the negative effects of duplication are sufficiently large for a broad scope to give a higher probability of innovation. To the left of area 2, the equilibrium under a narrow scope is of type A, as the incumbent chooses to overinvest sufficiently to keep the entrant out of technology C. There is no duplication and a narrow scope gives a higher probability of innovation. To the right of area 2, a higher value of γ_C increases total investments and decreases duplication under a narrow scope sufficiently to render it a higher innovation probability than a broad scope.

If we compare Figures 2.1 and 2.5, it is clear that the subset of parameter space for which a broad patent scope gives a higher innovation probability is now substantially smaller. The conclusion is that the effect of patent scope on innovation probability depends on whether the incumbent can commit to investing or not. If commitment is possible, the potential benefit of a broad scope is substantially smaller.

7.2 Licensing

Until now, any license agreement between the two firms has been precluded. It has been assumed that the incumbent always chooses to block the entrant's innovation, if it infringes on his patent. However, if the two firms can write a license agreement, the incumbent may choose to license its technology to the entrant, in return for a license fee. Suppose that the patent on technology C is broad in scope, but that the entrant nevertheless chooses to conduct R&D on technology C. If he innovates and the incumbent does not, the incumbent has two options: he can block the entrant's innovation, and earn his current monopoly profit, αV, or he can license his technology to the entrant, lose current profit but earn a license revenue in form of a fixed fee, F. I assume that the agreement is written ex post, after the entrant has innovated. If the incumbent agrees to license, it implies that the entrant has two strategies available under a broad patent scope. Either he chooses technology A, which gives him V if he innovates, or he chooses technology C which gives him V less the license fee F if he innovates.

The license fee will be determined by bargaining between the licensor and the licensee. The share of the surplus from the license agreement that goes to each firm depends on its outside option and its relative bargaining power. The incumbent's outside option is to continue selling his patented product, with profit αV. The entrant's outside option is his expected payoff from choosing to conduct R&D on technology A. First, suppose that the incumbent has all the bargaining power. Then, he will demand a license fee such that the entrant receives only his outside option, and the entrant always chooses to conduct R&D on technology A. This implies that if the incumbent has all bargaining power, allowing for license agreements does not have any effect on equilibrium investments nor on innovation probabilities. This maximum fee gives the lower bound on the effects of license agreements on investments, which is zero. If, on the other hand, the entrant has all bargaining power, the incumbent will receive a fee equal to his outside option, $F = \alpha V$. The lower is the license fee, the more likely it is that the entrant will choose technology C. Hence, the minimum license fee gives the upper bound on the effects of license agreements. If the bargaining powers lie in between these two extremes, the effect of licensing on investments and innovation falls between zero and the upper bound. To find the upper bound, I determine the effect of licensing on equilibrium outcomes for the minimum license fee $F = \alpha V$. [5]

[5] As in the baseline model, I assume that if indifferent, the entrant chooses technology A.

7.2.1 Equilibrium of type C

The expected payoff to the incumbent when both firms choose technology C and the entrant obtains a license in case he innovates is

$$\begin{aligned}\Pi_{I,L}(C,p_I,p_E) &= \alpha V+\gamma_C p_I(1-p_E)(V-\alpha V)+\gamma_C p_E(1-p_I)\,(0)\\ &\quad+\gamma_C p_E p_I\left(\frac{1}{2}\,(V-\alpha V)+\frac{1}{2}\,(0)\right)-\frac{(p_I)^2}{2},\end{aligned}$$

where subscript L denotes licensing. The difference between this expected payoff and (2.1) is that in case the entrant wins, the incumbent licenses the technology, gets license fee αV and loses current profit αV. The net gain is zero. In case both innovate and the entrant gets the patent, the net gain is zero.

The expected payoff to the entrant when both firms choose technology C and the entrant obtains a license in case he innovates is

$$\Pi_{E,L}(C,p_I,p_E)=\gamma_C p_E(1-p_I)(V-\alpha V)+\gamma_C p_E p_I\frac{1}{2}(V-\alpha V)-\frac{(p_E)^2}{2}.$$

The difference between this expected payoff and (2.2) is that the entrant has to pay a license fee in case he wins and the net gain is $V-\alpha V$. Solving for the Nash equilibrium yields:

$$p_{I,L}\left(\alpha,\gamma_C\right)=p_{E,L}\left(\alpha,\gamma_C\right)=\frac{2\gamma_C(1-\alpha)}{2+\gamma_C(1-\alpha)}.$$

The optimal investments for the entrant and the incumbent are identical. The reason is that the entrant indirectly takes into account the incumbent's profit loss through the license fee. In addition, the incumbent's expected payoff from not innovating when the entrant does is zero since the license revenue compensates him for the loss of current profit. Comparing these investments to their counterparts in equilibrium C in the baseline model, (2.4) and (2.5), I find that for all $\alpha>0$,

$$p_{I,L}\left(\alpha,\gamma_C\right) < p_I^C\left(\alpha,\gamma_C\right), \tag{2.12}$$

$$p_{E,L}\left(\alpha,\gamma_C\right) < p_E^C\left(\alpha,\gamma_C\right). \tag{2.13}$$

The entrant invests less under licensing because the net reward for innovation is lower, and the incumbent invests less because he has less to lose from not innovating.

7.2.2 The entrant's choice of technology

Under licensing, the entrant chooses between doing R&D on technology C which has a higher probability of success, but where the payoff is reduced to $V - \alpha V$ or technology A, which has a lower probability of success but yields a payoff of V. It is possible to derive a condition for when the entrant chooses technology C over A.

Proposition 6 *Let the license fee be αV. The entrant chooses technology C even under a broad patent scope if $\alpha < \bar{\alpha}_L$, where*

$$\bar{\alpha}_L = \frac{3\gamma_C - 2 + \gamma_C^2}{\gamma_C + \gamma_C^2}.$$

Proof. See part A1 of the Appendix. ■

The entrant chooses technology C for a sufficiently low α, that is when the license fee is sufficiently low. The threshold $\bar{\alpha}_L$, where L denotes licensing, is increasing in γ_C. A higher advantage for technology C relative to A increases the payoff to the entrant from choosing C.

7.2.3 Effects of patent scope

When licensing is allowed, a broad patent scope does not necessarily reduce duplication. When $\alpha < \bar{\alpha}_L$, the entrant chooses technology C even though a new innovation would infringe on the patent. Hence, the two firms invest in the same technology even under a broad patent scope. The probability of innovation is:

$$i^{B,L} = \gamma_C \left[p_{I,L}(\alpha, \gamma_C) + p_{E,L}(\alpha, \gamma_C) - p_{I,L}(\alpha, \gamma_C)\, p_{E,L}(\alpha, \gamma_C) \right],$$

where superscript L denotes licensing. Under a narrow scope, no license is required and the innovation probability is identical to that under no licensing, as given by (2.7). The subset of parameter space where the entrant chooses to conduct R&D on technology C and obtain a license is shown in Figure 2.6. In the figure, areas 1, 2 and 3 are the subsets of parameter space where under a broad scope, the entrant chooses technology A even when he has the option to get a license. The innovation probabilities are unaffected by the licensing option. In area 4, the entrant finds it profitable to use the licensing option and chooses C. For this subset of parameter space, the innovation probability is higher under a narrow scope. The reason is that there is now duplication of R&D under both narrow and broad scope and, in addition, the incumbent and entrant both invest less under a broad scope, as shown by (2.12) and (2.13).[6] Comparing Figures 2.1 and 2.6, it can be concluded that the area of the parameter space where a broad scope gives a higher probability of innovation is substantially smaller with than without licensing. If licensing is possible and the entrant has some bargaining power, the benefit of a broad patent scope is substantially smaller.

8 Concluding comments

The model developed in this paper is motivated by the perceived increase in patent scope in the US in the last two decades. The model predicts the level of investment in R&D and the innovation probability resulting from a narrow and broad patent scope, respectively. It suggests a new explanation for the empirical fact that incumbent firms have high innovation rates relative to entrant firms. An incumbent firm can be more likely to innovate even in the absence of any technological or cost advantage, if the firm has an advantage originating from policy, namely a broad scope of the patent he owns.

The main finding is that if the incumbent has a high stand-alone incentive to innovate and the patented technology has a small advantage relative to the alternative technology, a broad patent scope gives a higher probability of innovation than a narrow scope. Consequently, the negative effects of duplication of R&D investments are under some conditions sufficiently large to warrant a broad patent scope. Conversely, when the incumbent's stand-alone incentive is low or the patented technology has a large advantage, a narrow patent scope gives a higher probability of innovation. When the incumbent can commit to an investment level, or when license agreements can be made, the first result is partly reversed; in instances where the highest innovation probability was previously given by a broad scope, it is now obtained under a narrow scope. Consequently, the benefit of a broad patent scope largely relies on the assumptions that the firms act simultaneously and that there

[6] Figure 2.6 depicts the maximal effect of licensing agreements, which is when the entrant has all the bargaining power. If the incumbent has some bargaining power, this increases the license fee, which shifts area 4 to the right. If the incumbent has all the bargaining power, area 4 disappears completely and licensing has no effect on innovation probabilities.

is no possibility for license agreements.

The model shows that the effects of an increase in patent scope depend on innovation and industry characteristics. It is possible that an increase in patent scope increases the probability of innovation in a given industry. However, it requires that specific conditions on the form of competition, the technological alternatives, the incumbent's profit and the opportunities for license agreements are met. If not, the result is a reduction of the probability of innovation. According to this model, a uniform increase in patent scope, such as awarding patent holders larger powers in infringement lawsuits, cannot be an optimal policy.

The conclusion raises a new question: is the optimal policy implementable? To set the optimal scope ex ante, the Patent Office must make predictions of, for example, the technology's advantage relative to alternatives and the patent holder's incentive for further improvement of the innovation he seeks to patent. This might seem an inherently difficult task for the patent examiner. However, the patent scope is also determined ex post, if the patent holder sues another party for infringement. At this point in time, the industry characteristics are observed rather than predicted. The court, deciding whether a product infringes on the patent or not, can at least in principle obtain information on alternative technologies and the incumbent's incentives for innovation. If the court finds that a narrow scope would have generated a higher rate of innovation, it should decide that the product was not infringing on the patent. If entrant firms anticipate such a decision by the court, they will make the desirable technology choice.

A direction for future research is to increase the realism of the model by extending it to a dynamic framework, where the effects of expectations and the dynamics of technology development can be analyzed.

Bibliography

[1] Bessen, J. and Maskin, E. (2002), “Sequential Innovation, Patents and Imitation”, Institute for Advanced Study, School of Social Science Economics Working Papers No. 25.

[2] Blundell, R. Griffith, R. and Van Reenen, J. (1999), “Market Share, Market Value and Innovation in a Panel of British Manufacturing Firms” *Review of Economic Studies*, Vol. 66, No. 3, pp. 529-554.

[3] Cabral, L. and Polak, B. (2004), “Does Microsoft Stifle Innovation? Dominant Firms, Imitation and R&D Incentives”, CEPR Discussion Paper No. 4577.

[4] Chatterjee, K. and Evans, R. (2004), “Rivals’ search for buried treasure: competition and duplication in R&D”, *RAND Journal of Economics*, Vol. 35, No. 1, pp. 160-183.

[5] Dasgupta, P. and Maskin, E. (1987), “The Simple Economics of Research Portfolios”, *Economic Journal*, Vol. 97, No. 387, pp. 581-595.

[6] Domeij, B. (2000), *Pharmaceutical Patents in Europe*, Stockholm: Norstedt.

[7] Etro, F. (2004), “Innovation by Leaders”, *Economic Journal*, Vol. 114, No. 495, pp. 281-303.

[8] Gallini, N. (2002), “The Economics of Patents: Lessons from Recent U.S. Patent Reform”, *Journal of Economic Perspectives*, Vol. 16, No. 2, pp. 131-154.

[9] Green, J. and Scotchmer, S. (1995), “On the Division of Profit in Sequential Innovation”, *RAND Journal of Economics*, Vol. 26, No. 1, pp. 20-33.

[10] Heller, M. and Eisenberg, R. (1998), “Can Patents Deter Innovation? The Anticommons in Biomedical Research”, *Science*, Vol. 280, pp. 698-701.

[11] Jaffe, A. (2000), "The U.S. patent system in transition: policy innovation and the innovation process", *Research Policy*, Vol. 29, pp. 531-557.

[12] Jones, C. and Williams, J. (2000), "Too Much of a Good Thing? The Economics of Investment in R&D", *Journal of Economic Growth*, Vol. 5, No. 1, pp. 65-85.

[13] Katz, M. and Shapiro, C. (1987), "R & D Rivalry with Licensing or Imitation", *American Economic Review*, Vol. 77, No. 3, pp. 402-420.

[14] Kitch, E. (1977), "The Nature and Function of the Patent System" *Journal of Law and Economics*, Vol. 20, No. 2, pp. 265-290.

[15] Lerner, J. (1995), "Patenting in the Shadow of Competitors", *Journal of Law and Economics*,Vol. 38, No. 2, pp. 463-495.

[16] Merges, R. and Nelson, R. (1990), "On the Complex Economics of Patent Scope", *Columbia Law Review*, Vol. 90, No. 4, pp. 839-916.

[17] Reinganum, J. (1983), "Uncertain Innovation and the Persistence of Monopoly", *American Economic Review*, Vol. 73, No. 4, pp. 741-748.

[18] Segerstrom, P. and Zolnierek, J. (1999), "The R&D Incentives of Industry Leaders", *International Economic Review*, Vol. 40, No. 3, pp. 745-766.

[19] Shapiro, C. (2001), "Navigating the Patent Thicket: Cross Licenses, Patent Pools, and Standard-Setting " in Innovation Policy and the Economy, Cambridge, Mass. MIT Press.

[20] Tandon, P. (1983), "Rivalry and the Excessive Allocation of Resources to Research", *Bell Journal of Economics*, Vol. 14, No. 1, pp. 152-165.

[21] Walsh, J. Arora, A. and Cohen, W. (2003), "Effects of Research Tool Patents and Licensing on Biomedical Innovation" in Patents in the Knowledge-Based Economy, Wesley M. Cohen and Stephen A. Merrill, *Editors*, Committee on Intellectual Property Rights in the Knowledge-Based Economy, National Research Council.

[22] Zeira, J. (2003), "Innovations, Patent Races and Endogenous Growth", CEPR Discussion Paper No. 3974.

Appendix

A1 Proofs

Proof of Proposition 1

To compute the expected payoff to the entrant in equilibrium of type C, I insert the equilibrium investments $p_I^C(\alpha,\gamma_C) = \frac{2\gamma_C(2(1-\alpha)+2\alpha\gamma_C-\gamma_C)}{\left(4+2\alpha\gamma_C^2-\gamma_C^2\right)}$, and $p_E^C(\alpha,\gamma_C) = \frac{2\gamma_C(2+\alpha\gamma_C-\gamma_C)}{\left(4+2\alpha\gamma_C^2-\gamma_C^2\right)}$ into

$$\Pi_E\left(C,p_I,p_E\right) = \gamma_C p_E(1-p_I)V + \gamma_C p_E p_I \frac{1}{2}V - \frac{(p_E)^2}{2}.$$

Given $V=1$, the expression can be simplified to

$$\Pi_E\left(C,p_I^C(\alpha,\gamma_C),p_E^C(\alpha,\gamma_C)\right) = \frac{2\gamma_C^2\left(2+\alpha\gamma_C-\gamma_C\right)^2}{\left(4+2\alpha\gamma_C^2-\gamma_C^2\right)^2}.$$

To compute the expected payoff to the entrant in equilibrium of type A, I insert $p_E^A\left(\gamma_C\right) = (1-\gamma_C)$ into

$$\Pi_E\left(A,p_I,p_E\right) = (1-\gamma_C)p_E V - \frac{(p_E)^2}{2}.$$

Given $V=1$, the expression can be simplified to

$$\Pi_E\left(A,p_I,p_E^A\left(\gamma_C\right)\right) = \frac{1}{2}\left(1-\gamma_C\right)^2.$$

The entrant chooses technology C if

$$\frac{2\gamma_C^2\left(2+\alpha\gamma_C-\gamma_C\right)^2}{\left(4+2\alpha\gamma_C^2-\gamma_C^2\right)^2} > \frac{1}{2}\left(1-\gamma_C\right)^2$$

which can be simplified to

$$\alpha > \frac{4-8\gamma_C+\gamma_C^2+\gamma_C^3}{2\gamma_C^3}.$$

Proof of Proposition 2

The ratio of innovation probabilities is

$$\frac{i^N}{i^B} = \frac{2\gamma_C^2\left(20\alpha\gamma_C - 16\gamma_C - 8\alpha + 4\gamma_C^2 - 6\alpha\gamma_C^2 - \alpha\gamma_C^3 + 2\alpha^2\gamma_C^3 + 16\right)}{\left(\gamma_C^2\left(1-\alpha\right) + \left(1-\gamma_C\right)^2\right)\left(4 + 2\alpha\gamma_C^2 - \gamma_C^2\right)^2}.$$

The derivative of $\frac{i^N}{i^B}$ with respect to α is

$$\frac{\partial\left(\frac{i^N}{i^B}\right)}{\partial\alpha} = 2\gamma_C^2\frac{\begin{pmatrix} 144\gamma_C - 240\gamma_C^2 + 296\gamma_C^3 - 250\gamma_C^4 + 117\gamma_C^5 - 26\gamma_C^6 + 2\gamma_C^7 \\ +16\alpha\gamma_C^2 - 56\alpha\gamma_C^3 + 188\alpha\gamma_C^4 - 170\alpha\gamma_C^5 + 52\alpha\gamma_C^6 - 4\alpha\gamma_C^7 \\ -32\alpha^2\gamma_C^4 + 72\alpha^2\gamma_C^5 - 24\alpha^2\gamma_C^6 - 2\alpha^2\gamma_C^7 + 4\alpha^3\gamma_C^7 - 32 \end{pmatrix}}{\left(2\alpha\gamma_C^2 - \gamma_C^2 + 4\right)^3\left(2\gamma_C - 2\gamma_C^2 + \alpha\gamma_C^2 - 1\right)^2}.$$

The denominator of the above expression is positive since $4 > \gamma_C^2$. Let

$$\begin{aligned} F(\gamma_C, \alpha) \; = \; & 144\gamma_C - 240\gamma_C^2 + 296\gamma_C^3 - 250\gamma_C^4 + 117\gamma_C^5 - 26\gamma_C^6 + 2\gamma_C^7 + 16\alpha\gamma_C^2 \\ & -56\alpha\gamma_C^3 + 188\alpha\gamma_C^4 - 170\alpha\gamma_C^5 + 52\alpha\gamma_C^6 - 4\alpha\gamma_C^7 - 32\alpha^2\gamma_C^4 + 72\alpha^2\gamma_C^5 \\ & -24\alpha^2\gamma_C^6 - 2\alpha^2\gamma_C^7 + 4\alpha^3\gamma_C^7 - 32. \end{aligned}$$

$\frac{\partial\left(\frac{i^N}{i^B}\right)}{\partial\alpha}$ is positive if $F(\gamma_C, \alpha) > 0$. Applying constrained optimization, the problem can be formulated as $\min\limits_{\gamma_C,\alpha} F(\gamma_C, \alpha)$ subject to $0 \leqslant \alpha \leqslant 1$ and $0.5 \leqslant \gamma_C \leqslant 1$, where $F(\gamma_C, \alpha)$ is continuously differentiable. The global minimum of $F(\gamma_C, \alpha)$ is $F(0.5, 0) = 4.6$. Hence, for $\gamma_C \in [0.5, 1]$ and $\alpha \in [0, 1]$, we have that $\frac{\partial\left(\frac{i^N}{i^B}\right)}{\partial\alpha} > 0$.

Proof of Proposition 3

The derivative of $\frac{i^N}{i^B}$ with respect to γ_C is

$$\frac{\partial\left(\frac{i^N}{i^B}\right)}{\partial\gamma_C} = 2\gamma_C\frac{\begin{pmatrix} 304\alpha\gamma_C - 320\gamma_C - 64\alpha + 352\gamma_C^2 - 336\gamma_C^3 + 256\gamma_C^4 - 104\gamma_C^5 \\ +16\gamma_C^6 - 496\alpha\gamma_C^2 + 640\alpha\gamma_C^3 - 688\alpha\gamma_C^4 + 365\alpha\gamma_C^5 - 64\alpha\gamma_C^6 - 2\alpha\gamma_C^7 \\ +32\alpha^2\gamma_C^2 - 176\alpha^2\gamma_C^3 + 432\alpha^2\gamma_C^4 - 364\alpha^2\gamma_C^5 - 64\alpha^3\gamma_C^4 \\ +76\alpha^2\gamma_C^6 + 100\alpha^3\gamma_C^5 + 9\alpha^2\gamma_C^7 - 24\alpha^3\gamma_C^6 - 12\alpha^3\gamma_C^7 + 4\alpha^4\gamma_C^7 + 128 \end{pmatrix}}{\left(2\alpha\gamma_C^2 - \gamma_C^2 + 4\right)^3\left(2\gamma_C - 2\gamma_C^2 + \alpha\gamma_C^2 - 1\right)^2}.$$

The denominator of this expression is positive since $4 > \gamma_C^2$. Let

$$\begin{aligned} G(\gamma_C, \alpha) &= 304\alpha\gamma_C - 320\gamma_C - 64\alpha + 352\gamma_C^2 - 336\gamma_C^3 + 256\gamma_C^4 - 104\gamma_C^5 + 16\gamma_C^6 \\ &\quad -496\alpha\gamma_C^2 + 640\alpha\gamma_C^3 - 688\alpha\gamma_C^4 + 365\alpha\gamma_C^5 - 64\alpha\gamma_C^6 - 2\alpha\gamma_C^7 \\ &\quad +32\alpha^2\gamma_C^2 - 176\alpha^2\gamma_C^3 + 432\alpha^2\gamma_C^4 - 364\alpha^2\gamma_C^5 - 64\alpha^3\gamma_C^4 + 76\alpha^2\gamma_C^6 \\ &\quad +100\alpha^3\gamma_C^5 + 9\alpha^2\gamma_C^7 - 24\alpha^3\gamma_C^6 - 12\alpha^3\gamma_C^7 + 4\alpha^4\gamma_C^7 + 128. \end{aligned}$$

At $\gamma_C = 0$, the expression reduces to $G(\gamma_C, \alpha) = 128 - 64\alpha$, which is positive for all $\alpha \in [0, 1]$. At $\gamma_C = 1$ the expression reduces to $G(\gamma_C, \alpha) = 9\alpha^2 - 5\alpha + 4\alpha^4 - 8$, which is negative for all $\alpha \in [0, 1)$. $G(\gamma_C, \alpha)$ is continuous and the intermediate value theorem can be applied. Hence, for $\alpha \in (0, 1)$, there exists at least one maximum of the ratio $\frac{i^N}{i^B}$ in the interval $\gamma_C \in (0, 1)$.

Proof of Proposition 4

The incumbent's investments are: $p_{I,S}(\alpha, \gamma_C) = \frac{\gamma_C\left(1-\alpha-\frac{1}{2}\gamma_C+\frac{3}{2}\alpha\gamma_C\right)}{1-\frac{1}{2}\gamma_C^2+\alpha\gamma_C^2}$, and $p_I^C(\alpha, \gamma_C) = \frac{2\gamma_C(2(1-\alpha)+2\alpha\gamma_C-\gamma_C)}{4+2\alpha\gamma_C^2-\gamma_C^2}$.

$$\begin{aligned} p_{I,S}(\alpha, \gamma_C) &> p_I^C(\alpha, \gamma_C) \Leftrightarrow \\ 0 &> \frac{1}{2}\gamma_C\left(6\alpha\gamma_C - 2\gamma_C - 4\alpha + \gamma_C^2 - 3\alpha\gamma_C^2 - 4\alpha^2\gamma_C + 2\alpha^2\gamma_C^2\right) \end{aligned}$$

Let

$$J(\gamma_C, \alpha) = 6\alpha\gamma_C - 2\gamma_C - 4\alpha + \gamma_C^2 - 3\alpha\gamma_C^2 - 4\alpha^2\gamma_C + 2\alpha^2\gamma_C^2.$$

Applying constrained optimization, the problem can be formulated as $\max\limits_{\gamma_C,\alpha} J(\gamma_C, \alpha)$ subject to $0 \leqslant \alpha \leqslant 1$ and $\frac{1}{2} \leqslant \gamma_C \leqslant 1$, where $J(\gamma_C, \alpha)$ is continuously differentiable. The optimization yields a global maximum of $J(\gamma_C, \alpha)$ at $J(0.5, 0) = -0.75$. It implies that $p_{I,S}(\alpha, \gamma_C) > p_I^C(\alpha, \gamma_C)$.

Proof of Proposition 5

In equilibrium candidate of type C, the incumbents investment is

$p_{I,S}^C(\alpha, \gamma_C) = \min(p_{I,S}(\alpha, \gamma_C), \bar{p}_I(\gamma_C))$. Let $\hat{\alpha}_1 = \frac{3\gamma_C^3-8\gamma_C+4}{\gamma_C^2(5\gamma_C-2)}$. If $\alpha > \hat{\alpha}_1$, then $p_{I,S}^C(\alpha, \gamma_C) = p_{I,S}(\alpha, \gamma_C)$ and, if $\alpha \leq \hat{\alpha}_1$, $p_{I,S}^C(\alpha, \gamma_C) = \bar{p}_I(\gamma_C)$. In equilibrium candidate of type A the incumbents investment is

$p_{I,S}^A(\alpha, \gamma_C) = \max(p_I^A(\alpha, \gamma_C), \bar{p}_I(\gamma_C))$. Let $\hat{\alpha}_2 = \frac{\gamma_C^2-4\gamma_C+2}{\gamma_C^2}$. If $\alpha > \hat{\alpha}_2$, then $p_{I,S}^A(\alpha, \gamma_C) = \bar{p}_I(\gamma_C)$ and, if $\alpha \leq \hat{\alpha}_2$, then $p_{I,S}^A(\alpha, \gamma_C) = p_I^A(\alpha, \gamma_C)$.

There are four cases to consider: Case 1: $\alpha \leq \hat{\alpha}_1$ and $\alpha \leq \hat{\alpha}_2$. Case 2: $\alpha > \hat{\alpha}_1$ and $\alpha \leq \hat{\alpha}_2$. Case 3: $\alpha \leq \hat{\alpha}_1$ and $\alpha > \hat{\alpha}_2$. Case 4: $\alpha > \hat{\alpha}_1$ and $\alpha > \hat{\alpha}_2$. I start with case 3, and then proceed to cases 1,2 and 4.

Case 3. Compare $\Pi_I(C, \bar{p}_I(\gamma_C), p_E)$ and $\Pi_I\left(A, \bar{p}_I(\gamma_C), p_E^A(\gamma_C)\right)$. In equilibrium of type C, $p_E = \gamma_C\left(1 - \frac{\bar{p}_I(\gamma_C)}{2}\right)$. The payoff functions are

$$\begin{aligned}
\Pi_I\left(A, \bar{p}_I(\gamma_C), p_E^A(\gamma_C)\right) &= \frac{-10\gamma_C^2 + 8\gamma_C + 4\gamma_C^3 + 2\alpha\gamma_C^2 - 2\alpha\gamma_C^3 - \alpha\gamma_C^4 - 2}{\gamma_C^2}, \\
\Pi_I(C, \bar{p}_I(\gamma_C), p_E) &= \frac{-9\gamma_C^2 + 8\gamma_C + \gamma_C^3 + 2\gamma_C^4 + \alpha\gamma_C^2 + \alpha\gamma_C^3 - 3\alpha\gamma_C^4 - 2}{\gamma_C^2}.
\end{aligned}$$

$$\begin{aligned}
\Pi_I\left(A, \bar{p}_I(\gamma_C), p_E^A(\gamma_C)\right) &> \Pi_I(C, \bar{p}_I(\gamma_C), p_E) \Leftrightarrow \\
\frac{\begin{array}{c}-10\gamma_C^2 + 8\gamma_C + 4\gamma_C^3 + 2\alpha\gamma_C^2 \\ -2\alpha\gamma_C^3 - \alpha\gamma_C^4 - 2\end{array}}{\gamma_C^2} &> \frac{\begin{array}{c}-9\gamma_C^2 + 8\gamma_C + \gamma_C^3 + 2\gamma_C^4 + \alpha\gamma_C^2 \\ +\alpha\gamma_C^3 - 3\alpha\gamma_C^4 - 2\end{array}}{\gamma_C^2}
\end{aligned}$$

which can be simplified to $\gamma_C^2\left(2\gamma_C - 1\right)\left(1 - \gamma_C\right)\left(1 - \alpha\right) > 0$. For $\gamma_C > \frac{1}{2}$ and $\alpha < 1 : \gamma_C^2\left(2\gamma_C - 1\right)\left(1 - \gamma_C\right)\left(1 - \alpha\right) > 0$. The incumbent prefers A.

Case 1. Compare $\Pi_I(C, \bar{p}_I(\gamma_C), p_E)$ and $\Pi_I\left(A, p_I^A(\alpha, \gamma_C), p_E^A(\gamma_C)\right)$. From Case 3, it is clear that $\Pi_I\left(A, \bar{p}_I(\gamma_C), p_E^A(\gamma_C)\right) > \Pi_I(C, \bar{p}_I(\gamma_C), p_E)$.

In addition, $\Pi_I\left(A, p_I^A(\alpha, \gamma_C), p_E^A(\gamma_C)\right) \geq \Pi_I(A, \bar{p}_I(\gamma_C), p_E^A(\gamma_C))$ since $p_I^A(\alpha, \gamma_C) = \arg\max_{p_I} \Pi_I(A, p_I, p_E)$. Hence, $\Pi_I\left(A, p_I^A(\alpha, \gamma_C), p_E^A(\gamma_C)\right) > \Pi_I(C, \bar{p}_I(\gamma_C), p_E)$ and the incumbent prefers A.

Case 2. Compare $\Pi_I\left(C, p_{I,S}(\alpha, \gamma_C), p_{E,S}(\alpha, \gamma_C)\right)$ and $\Pi_I\left(A, p_I^A(\alpha, \gamma_C), p_E^A(\gamma_C)\right)$. The payoff functions are

$$\begin{aligned}
\Pi_I\left(C, p_{I,S}(\alpha, \gamma_C), p_{E,S}(\alpha, \gamma_C)\right) &= \frac{\begin{array}{c}8\alpha + 4\gamma_C^2 - 4\gamma_C^3 + \gamma_C^4 - 20\alpha\gamma_C^2 + 16\alpha\gamma_C^3 \\ -2\alpha\gamma_C^4 + 12\alpha^2\gamma_C^2 - 12\alpha^2\gamma_C^3 + \alpha^2\gamma_C^4\end{array}}{4\left(2\alpha\gamma_C^2 - \gamma_C^2 + 2\right)}, \\
\Pi_I\left(A, p_I^A(\alpha, \gamma_C), p_E^A(\gamma_C)\right) &= \frac{1}{2}\gamma_C\left(4\alpha + \gamma_C - 4\alpha\gamma_C + \alpha^2\gamma_C\right).
\end{aligned}$$

$$\begin{aligned}
\Pi_I\left(A, p_I^A(\alpha,\gamma_C), p_E^A(\gamma_C)\right) &> \Pi_I\left(C, p_{I,S}(\alpha,\gamma_C), p_{E,S}(\alpha,\gamma_C)\right) \Leftrightarrow \\
0 &> \\
&\frac{\begin{array}{c} 8\alpha + 4\gamma_C^2 - 4\gamma_C^3 + \gamma_C^4 - 20\alpha\gamma_C^2 + 16\alpha\gamma_C^3 \\ -2\alpha\gamma_C^4 + 12\alpha^2\gamma_C^2 - 12\alpha^2\gamma_C^3 + \alpha^2\gamma_C^4 \end{array}}{4\left(2\alpha\gamma_C^2 - \gamma_C^2 + 2\right)} \\
&-\frac{1}{2}\gamma_C\left(4\alpha + \gamma_C - 4\alpha\gamma_C + \alpha^2\gamma_C\right).
\end{aligned}$$

Let

$$\begin{aligned}
K(\gamma_C, \alpha) &= \frac{\begin{array}{c} 8\alpha + 4\gamma_C^2 - 4\gamma_C^3 + \gamma_C^4 - 20\alpha\gamma_C^2 + 16\alpha\gamma_C^3 - 2\alpha\gamma_C^4 + 12\alpha^2\gamma_C^2 \\ -12\alpha^2\gamma_C^3 + \alpha^2\gamma_C^4 \end{array}}{4\left(2\alpha\gamma_C^2 - \gamma_C^2 + 2\right)} \\
&\quad -\frac{1}{2}\gamma_C\left(4\alpha + \gamma_C - 4\alpha\gamma_C + \alpha^2\gamma_C\right)
\end{aligned}$$

I want to find the maximum value of $K(\gamma_C, \alpha)$. Applying constrained optimization, the problem can be formulated as $\max_{\gamma_C,\alpha} K(\gamma_C, \alpha)$ subject to $\alpha \in [0.000, 0.280]$ and $\gamma_C \in [0.550, 0.590]$, where $K(\gamma_C, \alpha)$ is continuously differentiable. The optimization yields a global maximum of $K(\gamma_C, \alpha)$ at $K(0.550, 0.280) = -0.03$. It implies that $\Pi_I\left(A, p_I^A(\alpha,\gamma_C), p_E^A(\gamma_C)\right) > \Pi_I\left(C, p_{I,S}(\alpha,\gamma_C), p_{E,S}(\alpha,\gamma_C)\right)$ and the incumbent prefers A.

Case 4. Compare $\Pi_I\left(C, p_{I,S}(\alpha,\gamma_C), p_{E,S}(\alpha,\gamma_C)\right)$ and $\Pi_I\left(A, \bar{p}_I(\gamma_C), p_E^A(\gamma_C)\right)$.

$$\Pi_I\left(C, p_{I,S}(\alpha,\gamma_C), p_{E,S}(\alpha,\gamma_C)\right) > \Pi_I\left(A, \bar{p}_I(\gamma_C), p_E^A(\gamma_C)\right) \Leftrightarrow$$

$$\begin{aligned}
&\frac{8\alpha + 4\gamma_C^2 - 4\gamma_C^3 + \gamma_C^4 - 20\alpha\gamma_C^2 + 16\alpha\gamma_C^3 - 2\alpha\gamma_C^4 + 12\alpha^2\gamma_C^2 - 12\alpha^2\gamma_C^3 + \alpha^2\gamma_C^4}{4\left(2\alpha\gamma_C^2 - \gamma_C^2 + 2\right)} \\
> \; &\frac{-10\gamma_C^2 + 8\gamma_C + 4\gamma_C^3 + 2\alpha\gamma_C^2 - 2\alpha\gamma_C^3 - \alpha\gamma_C^4 - 2}{\gamma_C^2}
\end{aligned}$$

Let

$$\begin{aligned}
\bar{\alpha}_S &= \frac{-4\gamma_C^2 + 24\gamma_C^3 - 38\gamma_C^4 + 12\gamma_C^5 + 3\gamma_C^6}{-4\gamma_C^4 + 4\gamma_C^5 + 9\gamma_C^6} + \\
&\quad \frac{2\sqrt{20\gamma_C^4 - 128\gamma_C^5 + 320\gamma_C^6 - 408\gamma_C^7 + 301\gamma_C^8 - 144\gamma_C^9 + 49\gamma_C^{10} - 10\gamma_C^{11}}}{-4\gamma_C^4 + 4\gamma_C^5 + 9\gamma_C^6}.
\end{aligned}$$

This gives

$$\begin{aligned}
\alpha &\geq \bar{\alpha}_S : \Pi_I\left(C, p_{I,S}(\alpha,\gamma_C), p_{E,S}(\alpha,\gamma_C)\right) \geq \Pi_I\left(A, \bar{p}_I(\gamma_C), p_E^A(\gamma_C)\right) \\
\alpha &< \bar{\alpha}_S : \Pi_I\left(C, p_{I,S}(\alpha,\gamma_C), p_{E,S}(\alpha,\gamma_C)\right) < \Pi_I\left(A, \bar{p}_I(\gamma_C), p_E^A(\gamma_C)\right).
\end{aligned}$$

I assume that if indifferent, the incumbent chooses C. Show that $\alpha_1 < \bar{\alpha}_S$ for $\gamma_C \in [0.5332, 0.6667]$.

$$\begin{aligned}
\alpha_1 &< \bar{\alpha}_S \Leftrightarrow \\
\frac{3\gamma_C^3 - 8\gamma_C + 4}{\gamma_C^2\left(5\gamma_C - 2\right)} &< \frac{-4\gamma_C^2 + 24\gamma_C^3 - 38\gamma_C^4 + 12\gamma_C^5 + 3\gamma_C^6}{-4\gamma_C^4 + 4\gamma_C^5 + 9\gamma_C^6} \\
&+ \frac{2\sqrt{\begin{array}{c} 20\gamma_C^4 - 128\gamma_C^5 + 320\gamma_C^6 - 408\gamma_C^7 + 301\gamma_C^8 \\ -144\gamma_C^9 + 49\gamma_C^{10} - 10\gamma_C^{11} \end{array}}}{-4\gamma_C^4 + 4\gamma_C^5 + 9\gamma_C^6}
\end{aligned}$$

The above expression can be simplified to

$$0 < 8\gamma_C^4\left(1-\gamma_C\right)\left(2\gamma_C - 1\right)\left(4\gamma_C + 9\gamma_C^2 - 4\right)\left(4\gamma_C - \gamma_C^2 + \gamma_C^3 - 2\right)\left(2\gamma_C^2 - 6\gamma_C + \gamma_C^3 + 4\right)$$

which holds for $\gamma_C \in [0.5332, 0.6667]$.

Show that $\alpha_2 < \bar{\alpha}_S$ for $\gamma_C \in [0.5332, 0.6667]$.

$$\begin{aligned}
\alpha_2 &< \bar{\alpha}_S \Leftrightarrow \\
\frac{\gamma_C^2 - 4\gamma_C + 2}{\gamma_C^2} &< \frac{-4\gamma_C^2 + 24\gamma_C^3 - 38\gamma_C^4 + 12\gamma_C^5 + 3\gamma_C^6}{-4\gamma_C^4 + 4\gamma_C^5 + 9\gamma_C^6} \\
&+ \frac{\left(2\sqrt{\begin{array}{c} 20\gamma_C^4 - 128\gamma_C^5 + 320\gamma_C^6 - 408\gamma_C^7 + 301\gamma_C^8 \\ -144\gamma_C^9 + 49\gamma_C^{10} - 10\gamma_C^{11} \end{array}}\right)}{-4\gamma_C^4 + 4\gamma_C^5 + 9\gamma_C^6}
\end{aligned}$$

The above expression can be simplified to

$$0 < 392\gamma_C^2 - 128\gamma_C - 496\gamma_C^3 - 11\gamma_C^4 + 648\gamma_C^5 - 543\gamma_C^6 + 122\gamma_C^7 - 9\gamma_C^8 + 16.$$

Let

$$f_{\alpha_2}(\gamma_C) = 392\gamma_C^2 - 128\gamma_C - 496\gamma_C^3 - 11\gamma_C^4 + 648\gamma_C^5 - 543\gamma_C^6 + 122\gamma_C^7 - 9\gamma_C^8 + 16.$$

$\frac{\partial f_{\alpha_2}(\gamma_C)}{\partial \gamma_C} > 0$ for $\gamma_C \in [0.5332, 0.6667]$ and $f_{\alpha_2}(0.5332) = 4.39 \times 10^{-2}$. It follows that for $\alpha < \bar{\alpha}_S$, the equilibrium is of type A and for $\alpha \geq \bar{\alpha}_S$ the equilibrium is of type C.

Proof of Proposition 6

The expected payoff to the entrant in equilibrium A is $\Pi_E\left(A, p_I, p_E^A(\gamma_C)\right) = \frac{1}{2}(1-\gamma_C)^2$. The expected payoff to the entrant from choosing technology C when the patent scope is broad and the incumbent demands a license fee αV is obtained by inserting $p_{I,L} = p_{E,L} = \frac{2\gamma_C(1-\alpha)}{2+\gamma_C(1-\alpha)}$ into

$$\Pi_{E,L}(C, p_I, p_E) = \gamma_C p_E(1-p_I)(V-\alpha V) + \gamma_C p_E p_I \frac{1}{2}(V-\alpha V) - \frac{(p_E)^2}{2},$$

which can be simplified to

$$\Pi_{E,L}(C, p_{E,L}, p_{I,L}) = \frac{2\gamma_C^2(1-\alpha)^2}{(2+\gamma_C(1-\alpha))^2}.$$

The entrant chooses technology C if

$$\frac{2\gamma_C^2(1-\alpha)^2}{(2+\gamma_C(1-\alpha))^2} > \frac{1}{2}(1-\gamma_C)^2,$$

which can be simplified to

$$\alpha < \frac{3\gamma_C - 2 + \gamma_C^2}{\gamma_C + \gamma_C^2}.$$

A2 Figures

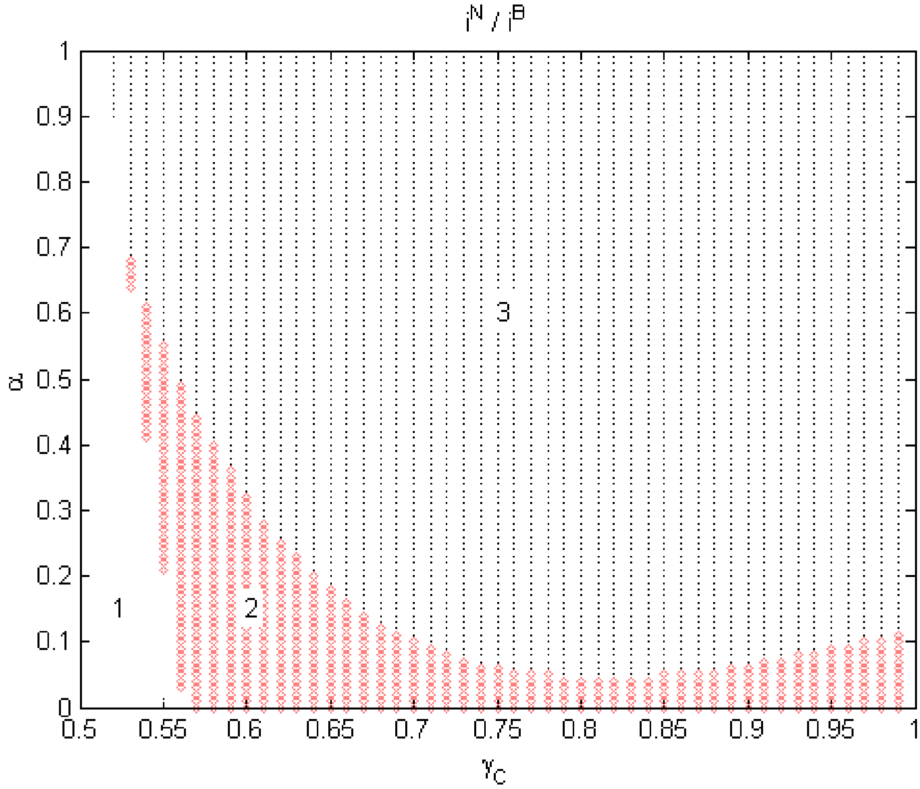

Figure 2.1: The ratio of innovation probabilities. Area 1: patent scope is inconsequential. Area 2: $i^N < i^B$. Area 3: $i^N > i^B$.

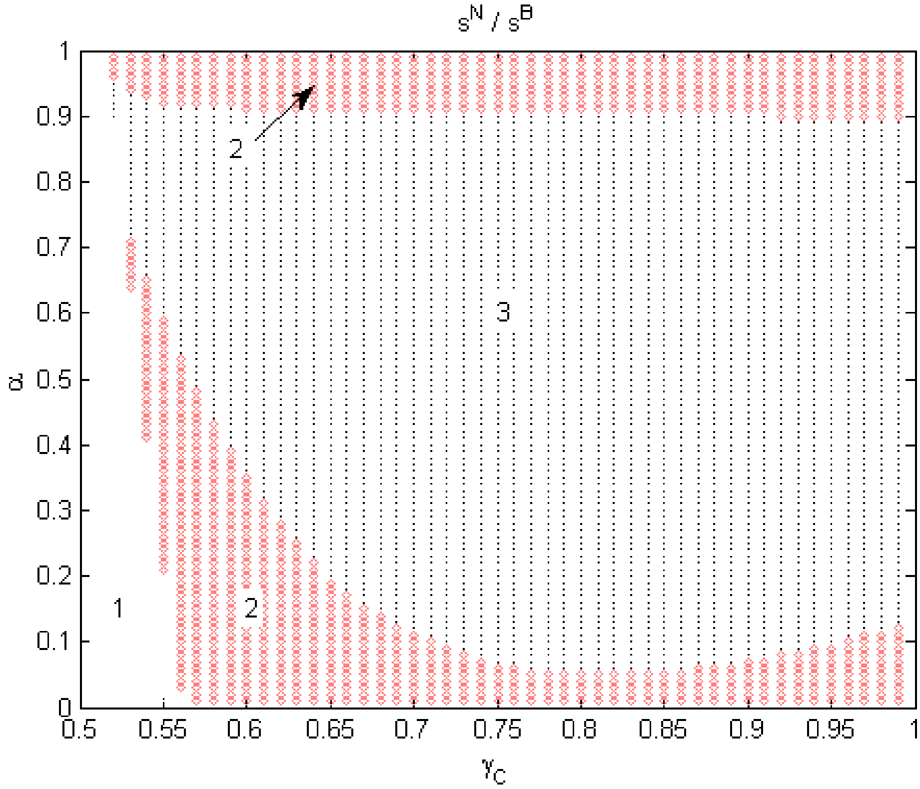

Figure 2.2: Social surplus. Area 1: patent scope is inconsequential. Area 2: $s^N < s^B$. Area 3: $s^N > s^B$.

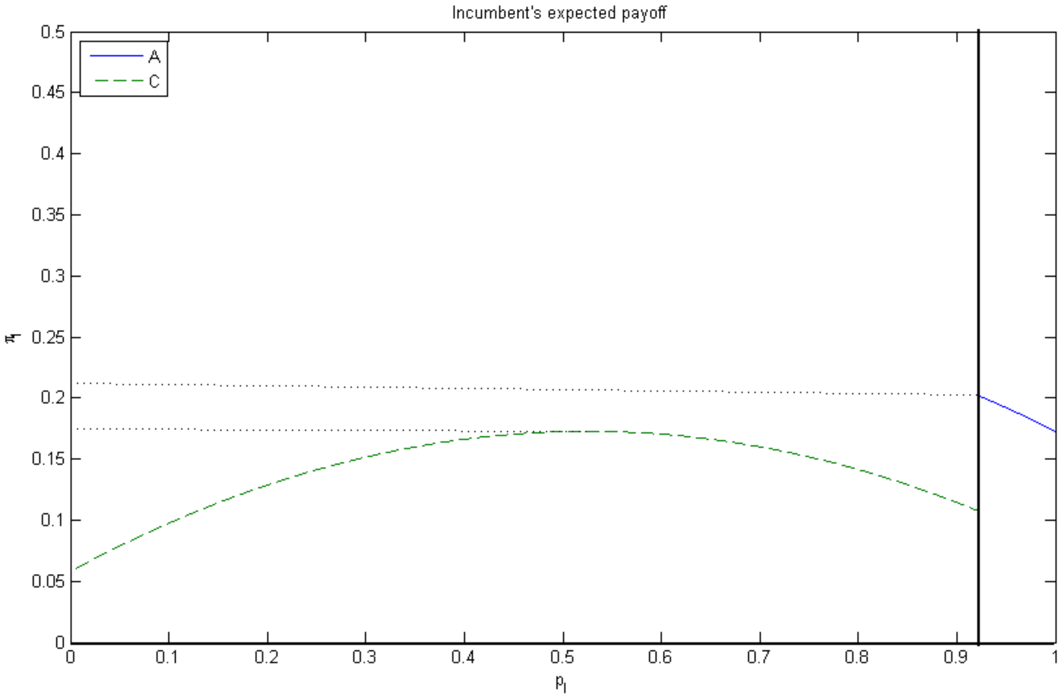

Figure 2.3: Incumbent's expected payoff in equilibrium candidates C and A when $\alpha < \bar{\alpha}_S$.

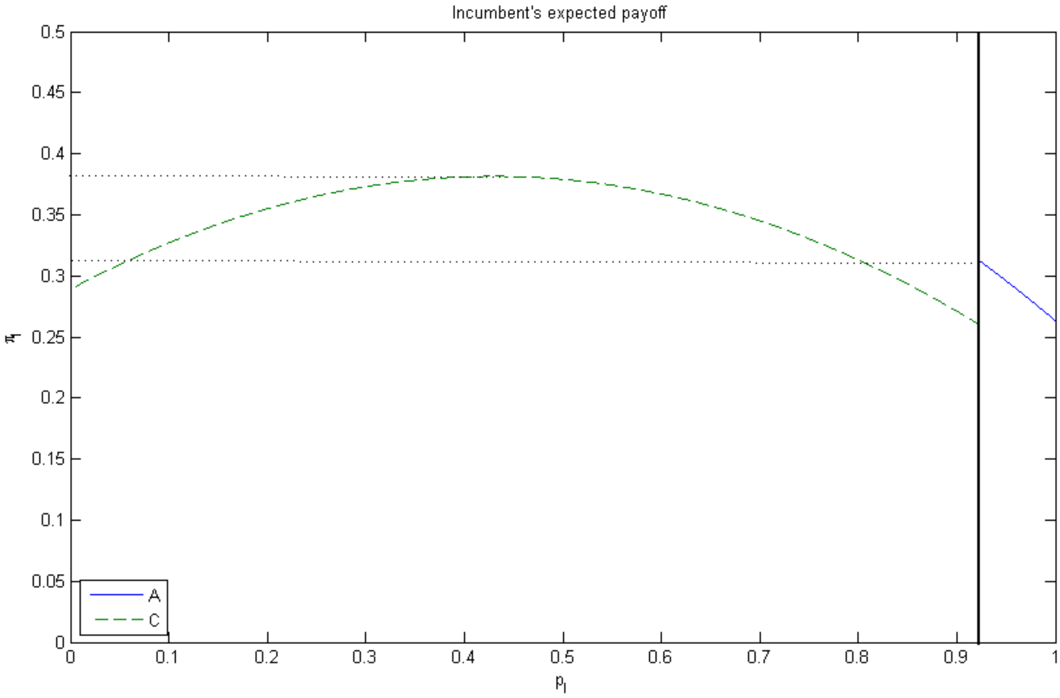

Figure 2.4: Incumbent's expected payoff in equilibrium candidates C and A when $\alpha > \bar{\alpha}_S$.

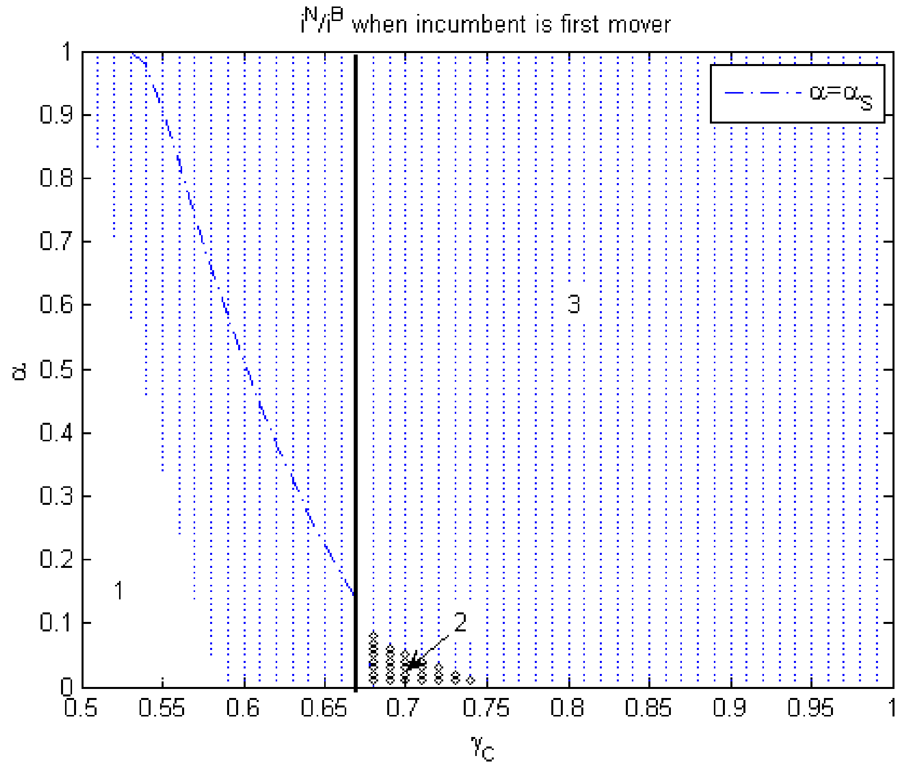

Figure 2.5: The ratio of innovation probabilities under Stackelberg competition. Area 1: patent scope is inconsequential. Area 2: $i^{N,S} < i^B$. Area 3: $i^{N,S} > i^B$.

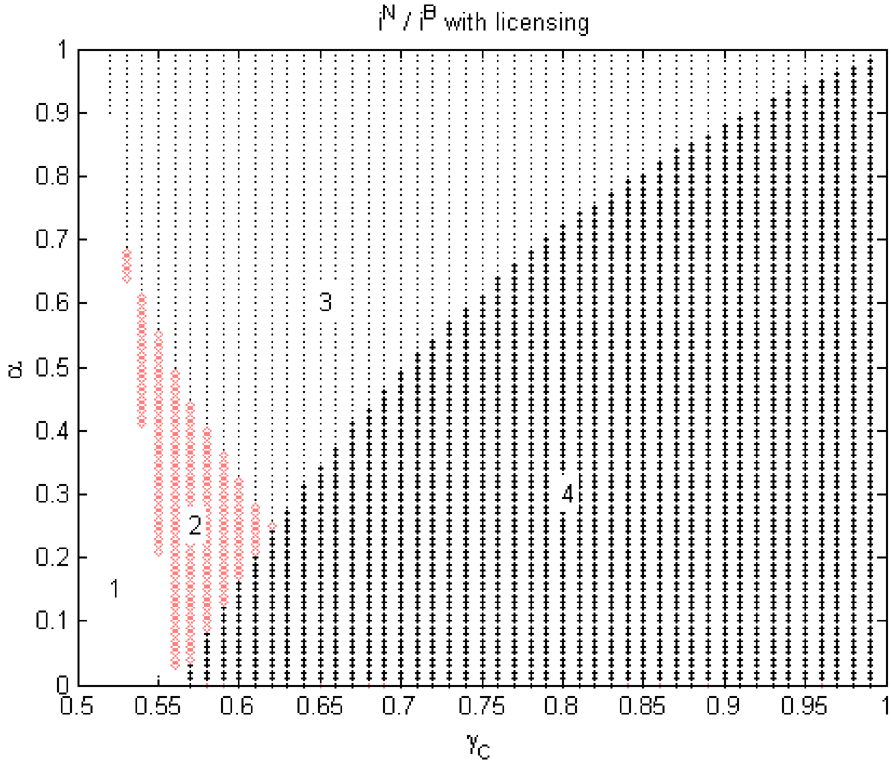

Figure 2.6: The ratio of innovation probabilities with licensing. Area 1: patent scope is inconsequential. Area 2: $i^N < i^B$. Area 3: $i^N > i^B$. Area 4: $\alpha < \bar{\alpha}_L$ and $i^N > i^{B,L}$.

Chapter 3

The World Distribution of Productivity: Country TFP Choice in a Nelson-Phelps Economy [*]

1 Introduction

What explains the relative levels of riches across countries? Building on the approached used in Klenow and Rodríguez-Clare (1997) and Hall and Jones (1999), Jones (2008) offers a recent account attributing the bulk of observed income-per-capita differences among countries to "TFP", or Total-Factor-Productivity, i.e., essentially to residually measured productivity differences. TFP differences are residual in that they capture the productivity differences that remain after differences in physical and human capital (per capita, or per worker) have been measured and taken into account, using the assumption of an aggregate Cobb-Douglas production function. Thus, under the assumptions maintained in these accounting analyses, to understand differences in income across countries it is not sufficient to understand differences in rates of accumulation of physical or human capital, at least if the available measures of physical and human capital are accurate. Therefore, much of the

[*] This is a joint essay with Per Krusell. We would like to thank John Hassler and seminar participants at IIES Brown Bag Seminar, University of Munich, Queen's University, NHH-UiO Workshop on economic dynamics 2007, SAET conference 2007, the Swedish Riksbank, and SED Annual Meeting 2008 for valuable comments and suggestions. We would also like to thank Christina Lönnblad for editorial assistance. All remaining errors are our own.

growth theory that analyzes such accumulation may "miss the point", if the purpose is to explain relative income differences across countries quantitatively. In conclusion, it seems reasonable to concur with the way Prescott (1998) puts it: "Needed: A Theory of Total Factor Productivity". The present paper proposes one direction in which theories with this aim might be built.

We also take the point of view that a theory of the world distribution of TFP should, ideally, display two features. One is that long-run world TFP growth is endogenous, i.e., "nontrivially" determined within the model. Thus, in this sense we very much take the view of the endogenous growth literature (Romer (1986), Lucas (1988), Rebelo (1991), and others). Whether the long-run world growth *rate* (or perhaps just the level increments) is endogenous is perhaps not central; for simplicity here though, we assume that it is. The second feature that we view as essential is that whereas each given country can influence its relative TFP level, it can do little to influence its long-run rate of TFP growth. More precisely, we view countries as small and their long-run growth rates as determined by the rate at which "world TFP" (i.e., some average, or perhaps the frontier) grows. This view is partly based on there only having been rather minor increases in the dispersion among countries during the period over which there is reliable country income data (essentially Heston, Summers, and Aten (2006)). In addition, modern economies are arguably (i) highly dependent on world technological developments, as evidenced by, among others, Benhabib and Spiegel (1994), who find productivity *catch-up* to be statistically significant for growth of TFP; and (ii) open in the sense that even if conventional trade is not entirely free, technology spillovers flow rather freely across countries in the world. In sum, we view the world average growth rate as endogenous but, conditional on this rate, any individual country's growth rate as exogenous.

Given this perspective, we construct a theory of the world TFP distribution that has at its core the specification of technology advancement proposed in Nelson and Phelps (1966). They hypothesized a form of technology, or human-capital, catch-up: in a country context, the growth rate of technology, or human capital, in a given country can be increased if this country invests, and the further is the distance from the technology frontier, the more productive is such an investment. That is, if a country is further behind, the potential for rapid growth is higher, since the country can "free-ride" on technologies/human capital accumulated elsewhere. A Nelson-

Phelps view of development is indeed taken explicitly in Jones's (1998) textbook on economic growth: he embeds the Nelson-Phelps formulation into a developing-country growth model, and uses it as a framework for productivity accounting and growth dynamics. In this paper, we thus take the Nelson-Phelps-Jones perspective on development but then go on to ask a further question: if countries are subject to this "technology for productivity growth", and each country operates the technology optimally from the viewpoint of maximizing the utility of its citizens, what is the implied *equilibrium world distribution of country TFP*? Since the catch-up term by definition means that countries benefit from spillovers, there is an obvious force for convergence, but how does this force play out in equilibrium?

Our main findings can be described as follows. Although convergence forces are always present in the model, if the catch-up term is weak enough, the stable long-run world distribution of TFP is not single-peaked but bimodal. There is one group of countries with high TFP in relative terms, with the remainder of the countries operating at a much lower TFP level. All countries grow at the same rate, but the high-TFP countries invest more in technology than do low-TFP countries. The catch-up term thus allows the low-TFP countries to grow at the same rate and not fall further behind in relative terms. The bimodality is first derived using a model without idiosyncratic (country-specific) TFP shocks, and with this setting bimodality means that the long-run TFP distributions have two (and exactly two) groups of countries, within each of which all countries have the same TFP level. In the model with shocks, in contrast, there is a "smooth" long-run distribution, and this distribution can have two visible peaks or just one. However, even when there is a single peak in the TFP distribution, and where the corresponding long-run, no-shocks outcome has a single group of countries with identical TFP, the distribution can have significant dispersion around the mode. Thus, the theory embodies strong forces pulling countries apart.

Formally, in our model, individual countries can invest in a technology-enhancing input, e, in order to increase their TFP, T. We think of this investment input as a traded one (such as $e$ducated workers), which can thus be allocated across countries, and there will be an equilibrium world price for this input. The accumulation of TFP, as mentioned, is of the Nelson-Phelps (1966) form: TFP growth in a given country depends positively on investment in TFP and, in addition, on the country's distance to the world technology frontier, which generates catch-up. The formal specification is that $\frac{T_{i,t+1}}{T_{i,t}} = \left(\frac{\bar{T}_t}{T_{i,t}}\right)^{\gamma} H(e_{i,t})$ where i is a country index and t stands for time; $\bar{T}$ is world average, or frontier, TFP, and the function H, which is increasing, describes how e units translate into TFP growth units. In the model, all countries are symmetric; they have identical technologies and only potentially differ by their initial conditions. The distribution of countries over TFP levels is determined by two counteracting forces. First, the technological catch-up generates convergence, and second, the internalization of country-specific dynamic gains from TFP investment generates divergence. Thus, in terms of this specification, we find that the distribution of TFP can be bimodal, provided that the weight on the catch-up factor in the Nelson-Phelps equation, γ, is sufficiently close to zero.

The intuition why catch-up is a force for convergence is straightforward. Why, though, is there a force for divergence if γ is small? Consider the extreme case, i.e., $\gamma = 0$. Here, we can write the TFP technology as $T_{i,t+1} = T_{i,t}H(e_{i,t})$ and it is clear that, as a dynamic technology for accumulation, the system has increasing returns to scale, as long as H is increasing: if $T_{i,t}$ and $e_{i,t}$ both double, $T_{i,t+1}$ more than doubles, or doubles exactly if H is flat. Thus, when individual country planners operate technologies of this sort, there are dynamic increasing returns to scale, and internalizing these gains is important and leads to a force for divergence: those countries with initially (or, by idiosyncratic stochastic events) high T_i levels have an advantage in further growth.

How can our findings be used to interpret the available country data? Aside from providing a theory of relative TFPs in the world (and a potential for bimodality), we use the theory to speculate on the findings in Acemoglu (2008). He plots a sequence of distributions of countries according to GDP per worker for 1960, 1980, and 2000, and it reveals some visible changes in the shape of the distribution and, in particular, in the "degree of bimodality". Thus, the 1980 and 2000 distributions are "more bimodal" than is the 1960 distribution. In terms of the model, a move toward bimodality can be due to, for example, (i) a decrease in effective spillovers (a lower γ); (ii) a decrease in the cost of trading e; (iii) increases in the extent to which e affects TFP (a change in the shape of H); or (iv) transitional dynamics. We discuss and interpret these channels in Section 8 of the paper.

Our present setting is not designed as a full-fledged quantitative assessment of the kind of theory we propose; among other things, we abstract from capital accumulation (thus, TFP and labor productivity are synonymous). Perhaps more importantly, a serious quantitative model of world TFP would require more geographic detail than our setting with (ex-ante) identical countries, since we envisage the critical distinction here—between which TFP investments are internalized and which are not—to coincide with country borders. Thus, for example, γ should in principle be determined based on how large countries are, and it would arguably also be more appropriate with a model which has heterogeneity in γ across countries. Such an extension, and a full quantitative investigation, would be an interesting one to pursue. Thus, our present analysis is to be interpreted as a first pass at how "rational TFP accumulation decisions on the country level", in combination with technology spillovers as proposed by Nelson-Phelps, deliver a model of relative riches in the world.

The paper is organized as follows. The related literature is presented in Section 2. Section 3 describes the model. The balanced growth equilibrium of the model is defined in Section 4. Section 5 describes the symmetric balanced growth equilibrium, and analyses its stability. Asymmetric balanced growth equilibria and their stability properties are characterized in Section 6. In Section 7, the model is extended to allow for country-specific shocks to TFP. Section 8 entails an analysis of what model parameters might change the distribution of relative TFP over time, and Section 9 concludes.

2 Related literature

In an influential article, Nelson and Phelps (1966) argued that in an economy with technological change, the more educated the workforce is, the faster new technologies of production will be introduced. The argument was formalized in a model where advancement of technology depends positively on investment in education and on the gap between the best-practice, or frontier, technology and the technology currently used. Parente and Prescott (1994) incorporate this mechanism into a model of an economy featuring firms, households, and a government. They propose that differences in barriers to technology adoption can account for the observed income disparities across countries. In the model, a firm can invest in the adoption of new and more productive technologies, and the amount of investment needed depends on the barriers to technology adoption in the country where it operates. Parente and Prescott calibrate the model and argue that the differences in barriers required to account for the observed cross-country income differences are not implausibly large. Similarly, Jones (1998) embeds the model proposed by Nelson and Phelps into the Romer model. He assumes that a country's human capital or skill accumulation depends on investment in education as well as technology spillovers from more advanced countries. Even though Parente and Prescott (1994) and Jones (1998) introduce the Nelson-Phelps framework into growth models, both explore the implications for income levels and growth rates in a partial equilibrium setting. In this paper, we extend the analysis to general equilibrium.

Large differences in per capita income across countries as well as a "twin-peaked" distribution of world income have been documented by, for example, Quah (1993), Quah (1997), and Kremer, Onatski, and Stock (2001). Several economists have constructed models aimed at explaining these empirical findings. An example is Chari, Kehoe and McGrattan (1997) which uses a neoclassical growth model to determine how much of the variation in incomes across countries that can be explained by distortions to capital accumulation. The distortions are modeled as a stochastic process for the price of capital, and the variation in incomes generated by the model is about 4/5 of the observed variation. Similarly, Acemoglu and Ventura (2002) explain the world income distribution by accumulation of capital in combination with international trade and specialization. The determinants of income differences across countries are technology levels and policies affecting incentives to invest. However, technology, rather than physical or human capital, appears to be the main determinant of the differences in incomes. For example, Klenow and Rodríguez-Clare (1997) find that 50 percent or more of cross-country variation in GDP per worker is explained by productivity differences. Similarly, Jones (2008) finds that approximately one-third of differences in income across countries can be explained by differences in capital per person, while differences in TFP explain the remaining two-thirds. Therefore, along with Parente and Prescott (1994), a number of models have been developed to explain income differences by modeling differences in productivity rather than in accumulation of physical capital.

Several models constructed to explain the differences in income across countries have used the Schumpeterian growth model as a point of departure. For example, Howitt (2000) analyzes a multi-country version of the Aghion-Howitt endogenous growth model with perfect technology transfer across countries. Under the assumption that countries differ in their R&D productivities, R&D subsidy rates, or investment rates, the model can generate "club convergence", whereby countries which invest in R&D will converge to parallel growth paths, and countries which do not will stagnate. Howitt and Mayer-Foulkes (2005) use a similar model with technology spillovers across countries. However, the extent to which a country benefits from spillovers depends on its level of human capital, in accordance with the argument by Nelson and Phelps (1966). They show that countries sort into three convergence groups characterized by R&D, implementation and stagnation, respectively. The cross-country differences in income are explained by the countries' levels of "competitiveness" and educational attainment.

Aghion, Howitt and Mayer-Foulkes (2005) introduce credit market imperfections in the multi-country Schumpeterian model. The model exhibits technology spillovers across countries, and an investment in R&D is a prerequisite for the receiving country to benefit from spillovers. It is assumed that R&D requires access to external finance and this access is restricted by the level of financial development. The model predicts that countries above a certain threshold of financial development will converge to a high growth rate, whereas all other countries will converge to strictly lower growth rates.

The present paper also models investments in productivity, but aside from taking an explicit Nelson-Phelps view on technology diffusion, it differs from the models described above in two additional ways. First, rather than assuming that countries have different characteristics, it treats all countries symmetrically. In addition, it models internalization of the dynamic gains from productivity investments at the country level.

3 Model

The world consists of a continuum of countries indexed i. Each country produces output and invests in TFP accumulation in order to increase future output. The investment in TFP can be R&D, technology adoption, improving institutions etc.

Each country is endowed with both low-skilled and high-skilled labor, where low-skilled labor is used in production of output and high-skilled labor in the accumulation of TFP. The number of high-skilled workers employed in country i at time t is denoted by $e_{i,t}$. It is assumed that low-skilled labor is immobile while high-skilled labor flows freely across countries. The total amount of high-skilled labor in the world is fixed and it is equal to e_W. All countries are of the same size. Country i's endowment of low-skilled labor is normalized to one, and its endowment of high-skilled labor is equal to the world total (and average), e_W.

Country i's income $Y_{i,t}$ is the output produced net of TFP investment costs:

$$Y_{i,t} = T_{i,t} \cdot 1 - w_t e_{i,t} + w_t e_W. \tag{3.1}$$

In (3.1), $T_{i,t}$ is country $i's$ level of TFP. Since the amount of low-skilled workers is normalized to 1, it is also equal to country $i's$ output. w_t is the world wage rate for high-skilled workers, and the investment cost constitutes of wage payments to foreign high-skilled workers. This formulation for income is admittedly simple, but it is a starting point for the analysis.

3.1 TFP accumulation

All countries have the same technology for TFP accumulation but possibly different starting levels of TFP. In each country, the TFP investment is chosen by a country planner.

In the model by Nelson and Phelps (1966), the advancement of technology depends positively on investment in education and on the gap between the best-practice, or frontier, technology and the technology currently used. This paper follows their formulation, but views the investment as a general investment in TFP, and allows the importance of technological catch-up to vary by a parameter, γ. Consequently, TFP in country i is accumulated according to

$$\frac{T_{i,t+1}}{T_{i,t}} = \left(\frac{\bar{T}_t}{T_{i,t}}\right)^{\gamma} H(e_{i,t}) \tag{3.2}$$

where the distance to the world TFP frontier is captured by the term $\frac{\bar{T}_t}{T_{i,t}}$, and $\bar{T}_t$ is the world average, or frontier, TFP level. Common access to the frontier TFP generates a faster catch-up the further behind a country is. The parameter $\gamma \in [0, 1]$ measures the strength of the catch-up effect. The investment in TFP is captured by the TFP production function, $H(e_{i,t})$.

It is assumed that $H(e_{i,t})$ is strictly increasing and strictly concave in $e_{i,t}$. In addition, it must satisfy the following conditions: $H(0) = 1$, $H'(0) = \infty$, and $H(e_{i,t})$ is bounded above by b, where b is not too large. The following functional form for $H(e_{i,t})$

$$H(e_{i,t}) = (b-1)\left(1+\frac{1}{e_{i,t}}\right)^{-\kappa} + 1$$

satisfies the conditions while being relatively simple, and this specification of $H(e_{i,t})$ will be used throughout the analysis.

The world TFP average, $\bar{T}_t$, has the form

$$\bar{T}_t \equiv \left(\int T_{i,t}^{\psi} di\right)^{1/\psi}.$$

The parameter ψ determines the extent to which $\bar{T}_t$ depends on the leading, or frontier, TFP level in the world. For $\psi = 1$, $\bar{T}_t$ is the arithmetic average of the TFP levels of all countries, and for $\psi = \infty$, it is equal to the highest TFP level. The world TFP level grows according to

$$\bar{T}_{t+1} = \bar{T}_t(1+g_t)$$

where g_t is endogenously determined.

A rewriting of the TFP accumulation function in (3.2) gives

$$T_{i,t+1} = T_{i,t}^{1-\gamma}\bar{T}_t^{\gamma} H(e_{i,t}). \tag{3.3}$$

From this expression, it is clear that investments in TFP have dynamic effects, some of which are specific to the country, as captured by the term $T_{i,t}$ and others which are international, as captured by $\bar{T}_t$. The parameter γ governs the share of these dynamic effects that is country-specific. For example, if $\gamma = 1$, dynamic gains of TFP investment arise only through technological catch-up. If $\gamma = 0$, dynamic gains of TFP investment are completely internalized within the country. The parameter γ shows to be crucial for the results of the model, and will be discussed at length below. For comparison, Jones (1998) uses the same formulation as (3.2) but views $T_{i,t}$ as human capital and $e_{i,t}$ as education expenses.

3.2 Consumers

Each country i has a dynastic household which maximizes utility given the utility function

$$U(C_i) = \sum_{t=0}^{\infty} \beta^t \log(C_{i,t}) \tag{3.4}$$

where β is the discount factor. The consumer is endowed with 1 unit of low-skilled labor and e_W units of high-skilled labor.

3.3 Country planner problem

The aim of this paper is to explain the distribution of TFP across countries. Therefore, the model focuses on inter-country relationships and is solved in general equilibrium, while intra-country relationships are given a cursory representation. We do not consider the aggregation of individual firms' TFP into country TFP in a given country, and output is specified only at the country level, as given by the expression in (3.1). Hence, it is assumed that technology flows freely within countries, and all dynamic effects are internalized within a country. Therefore, we will characterize the country planner's solution of the individual country's optimization problem. The country planner chooses a sequence of consumption allocations and investments in TFP so as to maximize consumer utility, taking the sequence of world prices and average TFP, $\{p_t, w_t, \bar{T}_t\}_{t=0}^{\infty}$ as given. The problem can be stated as follows.

$$\begin{aligned}
& \max_{\{T_{i,t}, C_{i,t}, e_{i,t}\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t \log(C_{i,t}) \\
& s.t. \\
\sum_{t=0}^{\infty} p_t C_{i,t} &= \sum_{t=0}^{\infty} p_t \left(T_{i,t} \cdot 1 - w_t e_{i,t} + w_t e_W\right) \\
T_{i,t+1} &= T_{i,t}^{1-\gamma} \bar{T}_t^{\gamma} H(e_{i,t}),
\end{aligned} \tag{3.5}$$

where p_t is the time-0 price of the time t good and $p_0 = 1$. The country planner maximizes utility of consumption given two constraints. The first is the resource constraint and the second governs the accumulation of TFP. Since each country is assumed to be small, its TFP choice has no effect on the average TFP level and therefore, the country planner takes $\bar{T}_t$ as given.

3.4 World equilibrium

A **world equilibrium** consists of sequences of allocations $\{T_{i,t}, e_{i,t}, C_{i,t}\}_{t=0}^{\infty}$ for all i and prices $\{w_t, p_t\}_{t=0}^{\infty}$, such that

1. $\{T_{i,t}, C_{i,t}, e_{i,t}\}_{t=0}^{\infty}$ solves the problem in (3.5) for all i;
2. $\bar{T}_t = \left(\int T_{i,t}^{\psi} di\right)^{1/\psi}$ for all t and
3. $e_W = \int e_{i,t} di$ for all t.

Condition 2 states that average TFP in the world is consistent with individual countries' TFP choices. Condition 3 ensures that there is market clearing in the market for high-skilled labor.

3.5 Initial characterization of a country's investment decision

In the model, it is assumed that there are perfect world capital markets. This implies that the interest rates p_t/p_{t+1} are exogenous from the point of view of an individual country. Given this assumption, each planner's utility maximization problem can be separated into two independent problems: an income-maximization problem and an intertemporal consumption allocation problem. We state each of them in turn.

In the income-maximization problem, the country planner chooses a sequence of future investments in TFP so as to maximize output net of investment costs

$$\max_{\{T_{i,t+1}\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} p_t \left(T_{i,t} - w_t H^{-1} \left(\frac{T_{i,t+1}}{T_{i,t}^{1-\gamma} \bar{T}_t^{\gamma}} \right) \right), \tag{3.6}$$

taking the sequence of world prices and average TFP, $\{p_t, w_t, \bar{T}_t\}_{t=0}^{\infty}$, as given. The expression in (3.6) is obtained by inserting the expression for $e_{i,t}$ from (3.3) into (3.1). The term $w_t e_W$ can be dropped since it is constant from the individual country's point of view.

Next, we turn to the intertemporal consumption allocation problem. Given sequences of prices and country income, $\{p_t, Y_{i,t}\}_{t=0}^{\infty}$, the household in country i chooses its intertemporal consumption allocation so as to maximize

$$\begin{aligned}
&\max_{\{C_{i,t}\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t \log(C_{i,t}) \\
&s.t. \\
&\sum_{t=0}^{\infty} p_t C_{i,t} = \sum_{t=0}^{\infty} p_t Y_{i,t}.
\end{aligned}$$

Optimization yields the following relationship between consumption growth and prices

$$\frac{C_{i,t+1}}{C_{i,t}} = \beta \frac{p_t}{p_{t+1}}. \tag{3.7}$$

In a world equilibrium, country and world consumption growth between t and $t+1$ equals $\beta(p_t/p_{t+1})$. Using the separation of the optimization problem, the world equilibrium can be redefined as follows.

A **world equilibrium** thus consists of sequences of allocations $\{T_{i,t}, e_{i,t}, C_{i,t}\}_{t=0}^{\infty}$ for all i and prices $\{w_t, p_t\}_{t=0}^{\infty}$, such that

1. $\{T_{i,t}\}_{t=0}^{\infty}$ solves the problem in (3.6) for all i, and $e_{i,t} = H^{-1}\left(\frac{T_{i,t+1}}{T_{i,t}^{1-\gamma}\bar{T}_t^{\gamma}}\right)$;

2. $\bar{T}_t = \left(\int T_{i,t}^{\psi} di\right)^{1/\psi}$ for all t;

3. $e_W = \int e_{i,t} di$ for all t;

4. $\{C_{i,t}\}_{t=0}^{\infty}$ is given by
$$\sum_{t=0}^{\infty} p_t C_{i,t} = \sum_{t=0}^{\infty} p_t Y_{i,t}$$
and $\frac{C_{i,t+1}}{C_{i,t}} = \beta \frac{p_t}{p_{t+1}}$ for all i, t.

In Section 5, the world equilibrium will also be defined recursively, for the special case of a distribution consisting of two groups of countries.

4 Balanced growth equilibrium

This analysis will focus on the long-run world distributions of TFP. Therefore, we restrict our attention to balanced growth equilibria. A balanced growth equilibrium is a world equilibrium, as defined in Section 3.5, where all variables grow at constant rates. There is a common world growth rate g of $\bar{T}_t$, $T_{i,t}$, $C_{i,t}$ and w_t. The common growth rate allows us to define a TFP-adjusted wage for high-skilled workers, $\hat{w}$:

$$\hat{w} \equiv \frac{w_0}{\bar{T}_0} = \frac{w_t}{\bar{T}_t}$$

for all t.

Similarly, the relationship in (3.7) can be rewritten as

$$\frac{p_t}{p_{t+1}} = \frac{1+g}{\beta}. \tag{3.8}$$

Before defining the balanced growth equilibrium, we restate the optimization problem in terms of relative TFP levels, $z_{i,t}$:

$$z_{i,t} \equiv \frac{T_{i,t}}{\bar{T}_t}.$$

This implies that (3.3) can be expressed as

$$z_{i,t+1} = z_{i,t}^{1-\gamma} \frac{H(e_{i,t})}{1+g}. \tag{3.9}$$

Using (3.8) and the variables thus defined, the country planner's income maximization problem in a balanced growth equilibrium can be stated as follows. The country planner chooses a sequence of future relative TFP levels so as to maximize output net of investment costs

$$\max_{\{z_{i,t+1}\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t \left(z_{i,t} - \hat{w} H^{-1} \left(\frac{z_{i,t+1}}{z_{i,t}^{1-\gamma}} (1+g) \right) \right),$$

taking $\hat{w}$ and g as given.

The solution to the optimization problem above results in the following Euler equation (where subscript i is omitted)

$$\frac{\hat{w}}{H'(e_t)}\frac{1}{z_t^{1-\gamma}} = \frac{\beta}{1+g} + \frac{\hat{w}\beta\,(1-\gamma)}{H'(e_{t+1})}\frac{z_{t+2}}{z_{t+1}^{2-\gamma}}. \tag{3.10}$$

As can be seen from (3.10), there are three effects of an increase in the relative TFP level on income. First, there is an increase in investment costs at time t, as captured by the term on the left-hand side. Second, there is an increase in output at time $t+1$, corresponding to the first term on the right-hand side. Third, there is a decrease in investment costs at time $t+1$, as given by the second term on the right-hand side. The last effect depends directly on γ. For low values of γ, a large part of the dynamic gains to TFP are country-specific, and investment in TFP today generates large decreases in future investment costs. As γ increases, the country-specific gains decrease and catch-up with the frontier becomes relatively more important.

4.1 Definition of a balanced growth equilibrium

Formally, a balanced growth equilibrium, a BGE, is a world equilibrium, as defined in Section 3.5, such that $p_t = \beta^t$, $e_{i,t} = e_i$ $\forall i,t$ and $T_{i,t} = T_i(1+g)^t$ $\forall i,t$ for $g > 0$.

What is the distribution of countries over relative TFP levels in a balanced growth equilibrium? Let this distribution be described by $\Gamma(z)$, a probability measure on (S, β_s) where $S \in [z_{\min}, z_{\max}]$ and β_s is the associated Borel σ-algebra. This measure will be further discussed below. Using $\Gamma(z)$, we can redefine the balanced growth equilibrium in recursive notation.

A **balanced growth equilibrium** consists of a stationary probability measure $\Gamma(z)$, variables $\hat{w}$ and g, and functions $v(z)$ and $E(z)$ such that

1. $\forall z$, $v(z)$ solves
$$v(z) = \max_e z - \hat{w}e + \beta v\left(z^{1-\gamma}\frac{H(e)}{1+g}\right);$$
2. $\forall z$, $E(z)$ is
$$E(z) = \arg\max_e v(z);$$
3. $\int_S (\frac{z^{1-\gamma}}{1+g}H(E(z)))^{1/\psi} d\Gamma(z) = 1;$
4. $\int_S E(z)d\Gamma(z) = e_W$; and

5. $\Gamma(z)$ satisfies
$$\Gamma(B) = \int_{z \in S: \frac{z^{1-\gamma}}{1+g} H(E(z)) \in B} d\Gamma(z) \quad \forall B \in \beta_S.$$

The first and second conditions give the value function and policy function, respectively. The third condition states that the integral over all relative TFP levels must equal 1. The fourth condition is the market-clearing condition for high-skilled workers. The last condition ensures that the probability measure $\Gamma(z)$ is stationary. $\Gamma(z)$ is stationary if, for each set $B \in \beta_S$, the distribution of countries over relative TFP levels is time-invariant.

The astute reader has noticed that the consumption allocation is absent from the definition of the balanced growth equilibrium above. The reason is twofold. First, the relationship between consumption growth and prices, as specified in (3.8), is already embedded in the definition of $v(z)$. Second, given the assumption of perfect capital markets, the distribution of consumption across countries is independent of the relative TFP levels, z. In a balanced growth equilibrium, the level of consumption for a country is given by the consumer's intertemporal consumption allocation problem. Although consumption growth is identical in all countries, the level of consumption in a given country depends on initial conditions.

An obvious candidate for a balanced growth equilibrium is one where all countries behave identically: a symmetric one. In a symmetric BGE, or SBGE, the measure $\Gamma(z)$ is degenerate with its entire mass at $z = 1$. For any other shape of $\Gamma(z)$ it must be the case that the function $E(z)$ implies multiple stationary points.[1] We will discuss the different possible outcomes in turn, starting with the SBGE.

5 Symmetric balanced growth equilibria

A candidate for an SBGE has the following two characteristics. First, all countries choose a level of employment of high-skilled labor equal to the world average, e_W. Second, the world distribution of country relative TFP levels, $\Gamma(z)$, is degenerate at $z_i = 1$ for all i; thus it is trivially single-peaked, or unimodal.

In an SBGE, the world growth rate g is determined by (3.9) which, evaluated in

[1] A stationary point is a stationary solution to (3.9) or, in recursive terms, $z = z^{1-\gamma} \frac{H(E(z))}{1+g}$.

$z_{t+1} = z_t = 1$, gives

$$g = H(e_W) - 1.$$

Is this allocation optimal for each country? The TFP-adjusted wage rate for high-skilled workers, $\hat{w}$, will be such that the first-order condition from the country planner's optimization problem is satisfied. Therefore, the Euler equation, (3.10), gives the wage rate as

$$\hat{w} = \frac{\beta H'(e_W)}{(1+g)(1-\beta(1-\gamma))}.$$

As a result of the symmetry across countries, both g and $\hat{w}$ are determined by e_W, the average number of high-skilled workers in the world. In addition, $\hat{w}$ is decreasing in γ whereas g is independent of γ.

Through the determination of the wage rate, the necessary condition for optimality is thus satisfied. However, the first-order condition is not automatically sufficient, as the objective function is not necessarily concave. In fact, the concavity of the objective function is determined by the parameter γ. We resort to numerical solutions for $v(z)$ and $E(z)$ which show that below a threshold level of γ, the second-order conditions for optimality are not satisfied at $z = 1$ and an SBGE does not exist. The following sections contain this analysis. In Section 5.3, we then consider stability analysis and transitional dynamics.

5.1 A numerical example

To illustrate the characteristics of the balanced growth equilibrium, we provide a numerical example. The parameter values have been set as follows. The parameter b, which governs the upper bound of the TFP production function, is set to 4.5. κ, which governs the concavity of the function, is set to 0.4. The consumers' discount factor, β, equals 0.9. The total number of high-skilled workers in the world, e_W, is set to 0.1. Finally, the parameter governing the weight of frontier countries' TFP in average TFP, ψ, is set to 1 which implies that $\bar{T}$ is an arithmetic average of all T_i. Note that, at this stage, the parameter values are not chosen to match real-world data.

5.2 Existence in an SBGE: a country's policy function

Using the numerical example described above, we calculate an individual country planner's optimal policy function, $z_{t+1} = f(z_t)$. Figures 3.1-3.4 show the properties of the policy function for different values of γ, under the assumption that all other countries are in a symmetric balanced growth equilibrium. The stationary points depicted are those for which sufficient conditions for optimality are satisfied. Figure 3.1 depicts the case when γ is high, i.e., when technological catch-up is important for TFP growth. The policy function has a unique stationary point which is the symmetric equilibrium $z = 1$. As γ decreases, the policy function starts to bend downward to the left of $z = 1$ and upward to the right of $z = 1$. This case is shown in Figure 3.2. Figure 3.3 illustrates that as γ decreases further, two new stationary points emerge, one on each side of $z = 1$. The stationary point $z = 1$ itself now becomes unstable, a result that will be discussed in the stability analysis below. Finally, as depicted in Figure 3.4, for sufficiently low values of γ, the symmetric stationary point ceases to exist, while the two asymmetric points remain, albeit further apart. A more detailed account of how the existence of the SBGE depends on γ in our numerical example is given in Table 3.1.

5.3 Stability and transitional dynamics

The previous section showed that for high values of γ, an SBGE exists. The one-group model does not exhibit any transitional dynamics around this SBGE. Thus, if the initial distribution of relative TFP levels, $z_{i,0}$, is identical for all countries i, then all countries will choose the same investment in TFP in the initial period, e_W, and the resulting growth rate will be identical to that of the SBGE, g, from the beginning of time. If all countries start in a symmetric equilibrium, the consumption levels of all countries are identical in all periods. If a country starts with an initial level z_0 higher (lower) than 1, it will have a higher (lower) level of consumption in every period than a country with $z_0 = 1$.[2] Next, we want to ascertain whether the symmetric equilibrium is stable.

There are several kinds of perturbations of countries, or groups of countries, with

[2] If a nontrivial distribution of initial asset positions is allowed, that will also influence the level of consumption.

respect to which the equilibrium could be stable (or unstable). First, the stability with respect to a perturbation of one single country can be determined. Thus, one country is given a relative TFP level z that is slightly different from $z = 1$, while the rest are at $z = 1$. What path will that country then follow? If, and only if, it converges back to $z = 1$, then the SBGE is stable with respect to perturbations of a single country. In addition, stability can be determined with respect to perturbations of groups of countries, where the groups could be of any size. Now, suppose countries are divided into n groups, of arbitrary size; within a group, all countries have the same level of z. Suppose that all groups are given initial relative TFP levels z that are slightly different from $z = 1$. If all groups converge back to $z = 1$, then the SBGE is stable with respect to perturbations of groups of countries. In this analysis, we will determine the stability of the equilibrium with respect to two kinds of perturbations; to a perturbation of one country, which we denote *measure-zero stability*, and to a 2-group perturbation, denoted *2-group stability*.

In order to perform the stability analysis, we define a 2-group recursive world equilibrium where transitions are possible, i.e., where the equilibrium does not have to exhibit balanced growth. This definition will also be the point of departure for a characterization of asymmetric balanced growth equilibria, which are discussed in the next section.

Let the two groups have relative TFP levels z_1 and z_2. Within a group, all countries are identical. As in the numerical example, ψ is set to 1, which implies that $\bar{T}_t$ is an arithmetic average of all $T_{i,t}$. Let φ be the share of countries belonging to group 1, which we denote the low-TFP group. The group with relative TFP level z_2 is denoted the high-TFP group. The sum of relative TFP levels must equal 1, which implies that $z_2 = \frac{1-\varphi z_1}{1-\varphi}$. Consequently, the programming problem can be defined using only one aggregate state variable; z_1. A recursive 2-group world equilibrium can be defined as follows.

A recursive **2-group world equilibrium** consists of $v(z, z_1)$, $E(z, z_1)$, $w(z_1)$, $f(z_1)$, and $g(z_1)$ such that (subscripts i omitted for convenience)

1. $\forall(z, z_1)$, $v(z, z_1)$ solves

$$v(z, z_1) = \max_{e} z - w(z_1)e + \beta v\left(z^{1-\gamma}\frac{H(e)}{f(z_1)}, g(z_1)\right);$$

2. $\forall (z, z_1)$, $E(z, z_1)$ is
$$E(z, z_1) = \arg\max_e v(z, z_1);$$

3. $\forall z_1$, $g(z_1) = z_1^{1-\gamma} \frac{H(E(z_1,z_1))}{f(z_1)}$;

4. $\forall z_1$, $f(z_1) = \varphi z_1^{1-\gamma} H(E(z_1, z_1)) + (1-\varphi) \left(\frac{1-\varphi z_1}{1-\varphi}\right)^{1-\gamma} H\left(E\left(\frac{1-\varphi z_1}{1-\varphi}, z_1\right)\right)$; and

5. $\forall z_1$,
$$e_W = \varphi E(z_1, z_1) + (1-\varphi) E\left(\frac{1-\varphi z_1}{1-\varphi}, z_1\right). \tag{3.11}$$

The first and second conditions give the value function and the policy function, respectively, for a country which faces an aggregate relative TFP level z_1 and chooses its individual relative TFP level z. The third condition states the law of motion for the aggregate state variable z_1. The function $f(z_1)$ determines the gross aggregate growth rate $1+g_t$. The last condition is the market-clearing condition for high-skilled workers.

5.3.1 Measure-zero stability

To determine whether the symmetric equilibrium is stable with respect to measure-zero perturbations, we analyze the behavior of an individual country whose relative TFP level differs slightly from the symmetric steady-state value, $z = 1$. Using the notation introduced in the definition of the 2-group world equilibrium above, (3.9) can be restated as

$$z' = z^{1-\gamma} \frac{H(E(z, z_1))}{f(z_1)}. \tag{3.12}$$

Measure-zero stability can then be established based on the derivative of (3.12):

$$\frac{\partial z'}{\partial z} = (1-\gamma) z^{-\gamma} \frac{H(E(z, z_1))}{f(z_1)} + z^{1-\gamma} \frac{H'(E(z, z_1)) E_1(z, z_1)}{f(z_1)}, \tag{3.13}$$

where $E_1(z, z_1)$ is the derivative of the policy function $E(z, z_1)$ with respect to its first argument. It is possible to deduce $E_1(z, z_1)$ from the recursive version of the Euler equation, (3.10). By taking the derivative with respect to z, a second-order equation in $E_1(z, z_1)$ is obtained. The two solutions for $E_1(z, z_1)$ result in a pair of expressions for $\frac{\partial z'}{\partial z}$. Saddle-path stability corresponds to one expression larger than

1 in absolute value and one expression less than 1 in absolute value. Evaluating (3.13) at $z_1 = z = 1$, $f(z_1) = H(e_W)$, and $E(z, z_1) = e_W$ yields

$$\frac{\partial z'}{\partial z} = 1 - \gamma + \frac{H'(e_W)}{H(e_W)} E_1(1,1). \tag{3.14}$$

Here, we see that the stability properties of the SBGE will depend heavily on γ.

Equivalently, linearization of (3.10) around $z = 1$ yields the characteristic equation

$$x^2 + Kx + \frac{1}{\beta} = 0 \tag{3.15}$$

where

$$G(x) = \frac{x}{H'(H^{-1}(x(1+g)))}$$

and

$$K = \frac{(1-(1-\gamma)\beta)\, G(1)}{(1-\gamma)\,\beta G'(z^\gamma)} - \frac{1}{(1-\gamma)\,\beta} - (1-\gamma)\,. \tag{3.16}$$

The roots to the equation in (3.15) determine the stability of the system with respect to measure-zero perturbations around the value $z = 1$. The product of the roots is equal to $1/\beta$, which implies that at least one of the roots is larger than 1. As seen from (3.16) the values of the roots depend on γ. Using our numerical example, we show that for high values of γ, one root is less than 1 in absolute value and one root larger than 1. Hence, the SBGE is saddle-path stable. This is the case depicted in Figures 3.1 and 3.2. For intermediate values of γ, both roots are larger than 1 in absolute value and, consequently, the SBGE is unstable. Figure 3.3 corresponds to this case. Any country with a relative TFP level smaller than $z = 1$ will converge to the stationary point to the left of $z = 1$ and any country with a relative TFP level larger than $z = 1$ will converge to the stationary point to the right of $z = 1$. For low values of γ, the roots are complex, which contradicts optimality and there is no SBGE, as shown in Figure 3.4. Hence, for values of γ below some threshold, a country whose initial level of TFP is slightly different from $z = 1$ will not converge to the SBGE. The measure-zero stability of the numerically calculated examples is listed in Table 3.1.

The stability analysis established that for some values of γ, a single country perturbed from the SBGE will not converge to it. The next step is to determine to what relative TFP level the country will converge. In terms of Figures 3.1-3.4, these levels correspond to the stationary points other than $z = 1$ appearing in some cases. For this purpose, we examine the stationary version of (3.10)

$$z^{1-\gamma} = \frac{H'(e_W)}{H'(H^{-1}(z^{\gamma}(1+g)))}. \tag{3.17}$$

First, note that this equation is satisfied for $z = 1$, the symmetric BGE. Moreover, both the left-hand side and the right-hand side of the equation are increasing in z, indicating that the equation can have multiple solutions. In our numerical example, we found that when γ is low, there are at least two additional solutions to (3.17). This case is depicted in Figures 3.3 and 3.4.

Why do the asymmetric stationary points arise? The model exhibits two countervailing forces; the catch-up effect, which generates convergence, and the dynamic increasing returns, which generate divergence. When technological catch-up is less important, i.e., γ is low, the divergence effect dominates. Countries with a higher relative TFP level will find it optimal to invest more in TFP than countries with a lower relative TFP level.

5.3.2 2-group stability

In this section, we examine whether the SBGE is stable with respect to 2-group perturbations. As in the above definition, all countries are divided into two groups, denoted by their relative TFP levels: z_1 and z_2. The SBGE is stable if a group of countries, when given initial values of z slightly different from 1, converges back to the SBGE. Suppose that group 1 is given a TFP level slightly different from $z_1 = 1$. Whether the countries in group 1 will converge back to $z_1 = 1$ is determined by $g(z_1)$. The derivative of $g(z_1)$ is

$$\begin{aligned} g'(z_1) &= (1-\gamma)\, z_1^{-\gamma} \frac{H(E(z_1,z_1))}{f(z_1)} + z_1^{1-\gamma} \frac{H'(E(z_1,z_1))\,(E_1(z_1,z_1)+E_2(z_1,z_1))}{f(z_1)} \\ &\quad - z_1^{1-\gamma} H(E(z_1,z_1)) \frac{f'(z_1)}{f(z_1)^2} \end{aligned} \tag{3.18}$$

which, evaluated at $z_1 = 1$, $f(z_1) = H(e_W)$, and $E(z_1, z_1) = e_W$ equals

$$g'(1) = 1 - \gamma + \frac{H'(e_W)}{H(e_W)}\,(E_1(1,1) + E_2(1,1)) - \frac{f'(1)}{H(e_W)}. \tag{3.19}$$

In Appendix A1, we show that both $f'(1)$ and $E_2(1,1)$ are equal to zero. Hence, (3.19) is identical to (3.14) and, consequently, the 2-group stability analysis yields conclusions identical to those of the measure-zero stability analysis; for values of γ below a certain threshold, a group of countries whose initial levels of TFP are slightly different from $z = 1$ will not converge to the SBGE. The 2-group stability of the numerically calculated examples is shown in Table 3.1. The table shows computed SBGE for values of γ ranging from 0.216 to 0.226. The growth rate g is equal to 1.34 for all symmetric BGE, since it is independent of γ. This rate is very high, but it is a result of the choice of parameter values for the function $H(e_{i,t})$. The TFP-adjusted wage rate $\hat{w}$ is decreasing in γ. Symmetry across countries implies that the amount of high-skilled labor employed in TFP accumulation is 0.1, which is equal to the total amount of high-skilled workers in the world. The last two columns in Table 3.1 indicate whether the equilibrium is stable or not. For $\gamma = 0.224$ or higher, the equilibrium is stable. However, as γ decreases, it becomes unstable and finally, for $\gamma = 0.214$ or lower, the equilibrium ceases to exist.

6 Asymmetric balanced growth equilibria

An asymmetric balanced growth equilibrium, an ABGE, is one where all variables grow at constant rates and there is more than one level of relative TFP chosen by the country planners. We will focus on a particular type of ABGE, namely 2-group BGE. These are "natural" outcomes (as opposed to outcomes with more than two groups), as will be argued below.

6.1 2-group balanced growth equilibria

The 2-group BGE is a specific case of the 2-group world equilibrium defined in Section 5, where the growth rate is constant: $T_{i,t} = T_i(1+g)^t \; \forall i, t$ for $g > 0$. Since there are no transitional dynamics by definition in a BGE, we can omit the aggregate state variable z_1 from the optimization problem.

A recursive **2-group balanced growth equilibrium** consists of $v(z)$, $E(z)$, $w(z_1)$, $f(z_1)$, and $g(z_1)$ such that

1. $\forall z$, $v(z)$ solves
$$v(z) = \max_e z - w(z_1)e + \beta v\left(z^{1-\gamma}\frac{H(e)}{f(z_1)}\right);$$

2. $\forall z$, $E(z)$ is
$$E(z) = \arg\max_e v(z);$$

3. $\forall z_1$, $g(z_1) = z_1$ and $g(z_2) = z_2$, where $g(z_1) = z_1^{1-\gamma}\frac{H(E(z_1))}{f(z_1)}$;

4. $\forall z_1$, $f(z_1) = \varphi z_1^{1-\gamma} H(E(z_1)) + (1-\varphi)\left(\frac{1-\varphi z_1}{1-\varphi}\right)^{1-\gamma} H\left(E\left(\frac{1-\varphi z_1}{1-\varphi}\right)\right)$;

   and

5. $\forall z_1$,
$$e_W = \varphi E(z_1) + (1-\varphi)E\left(\frac{1-\varphi z_1}{1-\varphi}\right). \tag{3.20}$$

The first and second conditions give the value function and the policy function, respectively. The third condition ensures that the 2-group distribution is stationary. The function $f(z_1)$ determines the constant gross aggregate growth rate $1+g$. In the ABGE, the world growth rate is determined by the division of high-skilled labor into the low- and high-TFP groups, and the relative size of the two groups. In addition, it depends on γ, whereas the growth rate in the SBGE is independent of γ. The last condition is the market-clearing condition for high-skilled workers.

Let e_1 and e_2 be the amount of high-skilled labor employed in the low-TFP group and the high-TFP group, respectively. The unknown parameters e_1, e_2, and φ are determined by combining the Euler equations in steady state, (3.17), for both groups;
$$H(e_1)^{\frac{1-\gamma}{\gamma}} H'(e_1) = H(e_2)^{\frac{1-\gamma}{\gamma}} H'(e_2) \tag{3.21}$$

with the market-clearing condition for high-skilled labor (condition 5 in (3.20)). This system of two equations is underdetermined; it has one more unknown than the number of equations. Consequently, if one solution exists, there is an infinite number of solutions, indeed a whole continuum. Each solution has a corresponding distinct world growth rate and TFP-adjusted wage rate for high-skilled labor.

Equation (3.21) allows us to identify a key necessary condition for a 2-group ABGE to exist: what is required is non-monotonicity of $H(e)^{\frac{1-\gamma}{\gamma}}H'(e)$. Whether or not there is non-monotonicity depends on the primitives γ and $H(e)$. Since H is increasing and strictly concave, there are opposing forces, as a straightforward derivative of this expression reveals. We do not attempt to provide general conditions on H and γ that satisfy the necessary non-monotonicity here; for the specific functional form for H we consider in this paper, however, non-monotonicity is satisfied in a range $(0, \bar{\gamma})$, with $\bar{\gamma} < 1$.

Also, note that for ABGE with more than 2 groups, when there is non-monotonicity, for a generic e, there is an odd number of solutions to (3.21). Here, stability arguments (like those discussed in Section 5.3) make us focus on 2 (as opposed to 3) groups. We have not fully analyzed the case where (3.21) would admit more than 3 (say, 5) solutions, but such a possibility does at least not seem feasible with the functional form for H used here.

In an ABGE, the ratio of TFP between the low-and high-TFP group, $\frac{T_1}{T_2}$, is given by

$$\frac{T_1}{T_2} = \left(\frac{H(e_1)}{H(e_2)}\right)^{\frac{1}{\gamma}}$$

and the TFP-adjusted wage rate for high-skilled workers is given by

$$\hat{w} = \frac{\beta z_i^{1-\gamma} H'(e_i)}{(1+g)\,(1-\beta(1-\gamma))},$$

where $i \in \{1, 2\}$.

The assumption of free movement of high-skilled labor across countries ensures that the wage rate paid to high-skilled workers is identical in the two groups. Therefore, the TFP-adjusted wage rate can be obtained from the Euler equation for either group. In the asymmetric equilibrium, the wage rate depends on the relative TFP levels as well as on the division of high-skilled labor across the two groups.

As in the symmetric one, the asymmetric equilibrium has a consumption growth that is identical for both groups and thus for all countries, but the level of consumption in a given country depends on its total income and on initial conditions. All countries in the low-TFP group will have a lower total income and hence, a lower level of consumption than countries in the high-TFP group. If all countries start in the asymmetric equilibrium, countries within the same group will have identical consumption levels.

6.2 Numerical example

The numerical example presented in Section 5.1 can be used to characterize the 2-group asymmetric equilibria as well as the symmetric ones. Table 3.2 shows the computed symmetric and asymmetric balanced growth equilibria for different values of γ. For the symmetric BGE, it shows whether the equilibrium is stable along with its growth rate g. For the asymmetric BGE, it shows the resulting values for the "gap", i.e., the ratio of TFP between the low-and high-TFP group, as well as $\hat{w}$, g, and φ. The indeterminacy of the system of equations in the 2-group BGE implies that the numerical solutions entail ranges of values for the parameters. The table shows that for values of γ of 0.226 or higher, only the symmetric BGE exists and it is stable. For γ equal to 0.220, the symmetric BGE is unstable and there exists a continuum of asymmetric BGE.[3] As γ decreases to 0.214 or less, the symmetric BGE ceases to exist, while the asymmetric BGE remain. Within the group of ABGE, the table shows that the growth rate g and the TFP-adjusted wage rate $\hat{w}$ decrease in γ.

From Table 3.2, we can conclude that if technological catch-up is important for TFP growth, i.e., γ is high, the distribution of TFP is symmetric. If, instead, technological catch-up is less important, i.e., γ is low, the distribution of TFP is asymmetric: twin-peaked. As mentioned above, the intuition for the existence of asymmetric BGE is that for sufficiently low values of γ, the dynamic increasing returns effect (which creates divergence) starts to dominate the catch-up effect (which creates convergence). This implies that countries with a higher relative TFP level will find it optimal to invest more in TFP than countries with a lower relative TFP level. The former become technological leaders while the latter become technological laggards which benefit from technology diffusion from the leaders. If the countries have different initial TFP levels, then countries with lower initial relative TFP will invest less, such that they eventually reach z_1, which constitutes the low-TFP group of the ABGE. Similarly, countries with higher initial relative TFP will invest more,

[3] For a small range of values for γ, symmetric and asymmetric BGE coexist. Within that range, as γ decreases, the symmetric BGE goes from being stable to unstable.

such that they eventually reach z_2, which constitutes the high-TFP group.[4]

6.3 Stability properties of the ABGE

To ascertain whether the asymmetric balanced growth equilibria are stable, we perform the same type of stability analyses as for the symmetric balanced growth equilibria: measure-zero stability and 2-group stability.

6.3.1 Measure-zero stability

An ABGE characterized by the triplet e_1, e_2, and φ is stable with respect to measure-zero perturbations if a single country, which is given an initial relative TFP level slightly different from z_1, converges back to z_1, and a single country perturbed away from z_2 converges back to z_2. Whether the country converges back or not is determined by $\frac{\partial z'}{\partial z}$, as given by (3.13), evaluated at $z = z_1$ and $z = z_2$, respectively. As in the case of the measure-zero stability analysis of the SBGE, the derivative of the recursive version of (3.10) with respect to z gives a second-order equation in $E_1(z, z_1)$, which results in a pair of expressions for $\frac{\partial z'}{\partial z}$. This pair is then evaluated at both z_1 and z_2. We compute measure-zero stability for the ABGE in the numerical example characterized above and Table 3.3 displays the results.

6.3.2 2-group stability

An ABGE characterized by the triplet e_1, e_2, and φ is stable with respect to 2-group perturbations if the following holds: when the low-TFP group is given an initial relative TFP level slightly different from z_1, it converges back to z_1. (Since a perturbation of one group affects the remaining group, it is sufficient to analyze perturbations of one group only.) Suppose that the low-TFP group is perturbed away from z_1. Whether it will converge back to z_1 is determined by the derivative of $g(z_1)$, as given by (3.18).

[4] If all countries start at the same initial TFP level, our conjecture is that they will split up into two groups, one which starts to invest less, such that it eventually reaches z_1, and one which starts to invest more, such that it eventually reaches z_2. In the initial period, the sequence of wages and relative TFP levels, $\{w_t, z_t\}_{t=0}^{\infty}$, must be such that the countries are indifferent between joining the low- and the high-TFP groups, and that they choose to split up into groups of a relative size which is consistent with the ABGE.

Unlike the in SBGE, the derivatives $f'(z_1)$ and $E_2(z_1, z_1)$ are not equal to zero, and must therefore be solved for in order to evaluate $g'(z_1)$. To that end, we compute the derivative with respect to z_1 of the recursive version of (3.10), and evaluate it at $z = z_1$ and $z = z_2$, respectively. Combining the resulting two equations with the expressions for $f'(z_1)$, $g'(z_1)$, and the derivative with respect to z_1 of condition 5 in (3.11), we obtain a system of five equations. There are five unknowns; $f'(z_1)$, $g'(z_1)$, $E_2(z_1, z_1)$, $E_2(z_2, z_1)$, and $w'(z_1)$. The system yields a second-order equation in $E_2(z_1, z_1)$ and therefore has two solutions. We solve the system of equations and for each solution obtain an expression for $g'(z_1)$; see Appendix A1 for details. If one of the expressions is larger than 1 in absolute value, and one is less than 1 in absolute value, there is saddle-path stability. We compute the values of $g'(z_1)$ for the ABGE in the numerical example characterized above, and the results are reported in Table 3.3. As shown in Table 3.3, the ABGE in the numerical example are stable, both with respect to measure-zero and to 2-group perturbations.

7 TFP shocks

In this section, the model is extended to allow for country-specific shocks to TFP. The motivation for this extension is twofold. First, it is to create a more smooth and realistic world TFP distribution where individual countries can move between groups and potentially experience both growth miracles and growth disasters. Second, it is an attempt to eliminate the indeterminacy of the asymmetric steady states obtained in the baseline model.

7.1 Model

The following assumptions are added to the model described in Section 3. Each country is subject to a TFP shock $\varepsilon_{i,t}$, $\varepsilon \in \{\varepsilon_L, \varepsilon_H\}$, where $\varepsilon_L < \varepsilon_H$. The probability of a good shock is denoted $P(\varepsilon_H) = \pi$. The shock ε is *iid* across countries and across time. The country planner sets $e_{i,t}$ before observing the shock $\varepsilon_{i,t}$. With TFP shocks, the accumulation of TFP has the following form

$$T_{i,t+1} = (1 + \varepsilon_{i,t})\, T_{i,t}^{1-\gamma} \bar{T}_t^{\gamma} H(e_{i,t})$$

which is (3.3) with the addition of the shock $\varepsilon_{i,t}$.

It is also assumed that the world has perfect consumption insurance and frictionless borrowing and lending. This ensures that a separation of the optimization problem into an income-maximization problem and an intertemporal consumption allocation problem is still valid. The intertemporal consumption allocation problem is the same as that in the model without technology shocks, and the resulting allocation will be the same, given total income. However, the income-maximization problem is different, as will be described below.

7.2 A country's investment decision

As in the model without shocks, we restrict our attention to balanced growth equilibria. In the income-maximization problem, the country planner chooses an investment in TFP so as to maximize output net of investment costs. The recursive formulation of the optimization problem is

$$v(z) = \max_{e} z - \hat{w}e + \beta(\pi v(z'_H) + (1-\pi)v(z'_L))$$

subject to, for $j = L, H$

$$z'_j = z^{1-\gamma}\frac{1+\varepsilon_j}{1+g}H(e).$$

The country planner employs an amount e of high-skilled workers in a given time period. With probability π the country is hit by a positive shock, and the resulting next period relative TFP level is z'_H and the corresponding value function is $v(z'_H)$. With probability $1-\pi$ the country is hit by a negative shock, and the resulting next period relative TFP level is z'_L, associated with the value function $v(z'_L)$.

7.3 Definition of a balanced growth equilibrium

A balanced growth equilibrium is a world equilibrium, where all variables grow at rate g on average. Individual countries can have a TFP growth that is faster or slower than this average rate g. The distribution of z is constant. As in the model without shocks, let $\Gamma(z)$ be a probability measure on (S, β_s) where $S \in [z_{\min}, z_{\max}]$ and β_s is the associated Borel σ-algebra. ψ is set to 1, which implies that $\bar{T}$ is an arithmetic average of all T_i.

A **balanced growth equilibrium** consists of a stationary probability measure $\Gamma(z)$, variables $\hat{w}$ and g, and functions $v(z)$ and $E(z)$ such that

1. $\forall z$, $v(z)$ solves

$$v(z) = \max_e z - \hat{w}e + \beta\left(\pi v\left(z^{1-\gamma}\frac{1+\varepsilon_H}{1+g}H(e)\right) + (1-\pi)v\left(z^{1-\gamma}\frac{1+\varepsilon_L}{1+g}H(e)\right)\right);$$

2. $\forall z$, $E(z)$ is

$$E(z) = \arg\max_e v(z);$$

3. g solves

$$\pi\int_S z^{1-\gamma}\frac{1+\varepsilon_H}{1+g}H(E(z))d\Gamma(z) + (1-\pi)\int_S z^{1-\gamma}\frac{1+\varepsilon_L}{1+g}H(E(z))d\Gamma(z) = 1;$$

4. $\hat{w}$ solves $\int_S E(z)d\Gamma(z) = e_W$; and

5. $\Gamma(z)$ satisfies

$$\begin{aligned}\Gamma(B) &= \pi\int_{z\in S:z^{1-\gamma}\frac{1+\varepsilon_H}{1+g}H(E(z))\in B} d\Gamma(z) + (1-\pi)\int_{z\in S:z^{1-\gamma}\frac{1+\varepsilon_L}{1+g}H(E(z))\in B} d\Gamma(z) \\ \forall B &\in \beta_S. \end{aligned} \tag{3.22}$$

The first and second conditions give the value function and policy function, respectively. The third condition states that the growth rate g must be such that the integral over all relative TFP levels equals 1. The fourth condition states that the wage $\hat{w}$ must be such that there is market clearing in the market for high-skilled workers. Condition 5 ensures that the probability measure $\Gamma(z)$ is stationary.

7.4 Solution method

The solution method used is similar to that in Aiyagari (1994) albeit with two unknown variables, g and $\hat{w}$. The method involves the following steps. Start with an initial guess for g and $\hat{w}$. Solve the dynamic programming problem using the guess and obtain the policy function $E(z)$. Simulate an individual country's choice of e and the resulting z for T time periods, where T is very large. Use the data generated to check whether condition 4 in (3.22) is satisfied. If not, update the guess for $\hat{w}$ using the bisection method and repeat the procedure until condition 4 holds. Then, check whether condition 3 in (3.22) is satisfied. If not, update the guess for g using the bisection method. Given the new guess for g, find the $\hat{w}$ for which condition 4 is satisfied. Given this combination of $\hat{w}$ and g, check whether condition 3 holds. If not, update the guess for g. Repeat the procedure until both conditions 3 and 4 are satisfied.

7.5 A numerical example

The model with TFP shocks is solved numerically using the parameterization in the example applied to the baseline model. There are three additional parameter values to be set; π, ε_L, and ε_H. π is set to 0.5, implying that the TFP shock is high and low with equal probability. The size of the shock is chosen such that the shock is symmetric; $\varepsilon_L = -0.05$ and $\varepsilon_H = 0.05$. Numerical solutions are then obtained for different values of γ.

Figures 3.5-3.7 depict the world distribution of relative TFP levels z_i corresponding to $\gamma = 0.24$, 0.225, and 0.22, respectively. From the results obtained in this numerical example, the equilibrium appears to be unique for a given value of γ. Figure 3.5 shows the world distribution of z_i for $\gamma = 0.24$. The distribution is single-peaked. As a result of the TFP shocks, there is dispersion around the center value $z = 1$, creating the smooth symmetric shape of the distribution. When γ is decreased to 0.225, as displayed in Figure 3.6, the distribution is still single-peaked but the dispersion has increased, some countries have started to lag behind, while others are moving toward higher relative TFP levels. Figure 3.7 shows that for γ as low as 0.22, the distribution of TFP is asymmetric; it is twin-peaked. One group of countries has settled at a low TFP level, another at a high TFP level. As countries are hit by shocks, it is possible for a given country to move from one of the groups to the other. However, the distribution of countries over TFP levels remains constant. We can conclude that for high values of γ, the distribution is single-peaked, while for sufficiently low values of γ the distribution of countries is twin-peaked.

Even though an individual country's TFP grows at rate g on average, its consumption grows at rate g for all t. The relative levels of consumption are determined by initial conditions. All countries which start at the same initial relative TFP level z_0 have the same consumption level. A country which starts at a higher (lower) initial level has a higher (lower) level of consumption in each time period.

How do the results from the model with TFP shocks compare to the model without shocks? The numerical results show that with shocks, the balanced growth equilibrium appears to be unique for a given value of γ, whereas without shocks, the model produced a continuum of asymmetric equilibria. However, the main results from the baseline model remain; if technological catch-up is important for TFP growth (γ is high) the distribution of TFP is symmetric. If, instead, technological catch-up is less important (γ is low) the distribution of TFP is asymmetric: twin-peaked.

8 Changes in the distribution of world TFP

The distribution of world TFP is not constant over time. Three snapshots of the distribution of labor productivity, measured as log (PPP-adjusted) output per worker, from Acemoglu (2008) are depicted in Figure 3.8.[5] We observed marked twin peaks in 1980 and more dispersion in 2000, though here the peak on the left is less clear, whereas the 1960 distribution shows only weak signs of bimodality. Whether there are two peaks or not, these distributions are different, and the general question we ask here is what fundamental determinants lie behind the shape of the distribution of TFP generated by our model. We will, in turn, and very briefly, discuss some of the determinants: changes in γ, changes in trading costs for e, changes in H, and transitional dynamics.[6]

8.1 The value of γ

First, we saw above that γ is a key determinant of whether bimodality is an outcome in our model; low values are required, ceteris paribus, for bimodality, and generally speaking lower γ tend to make the distribution more dispersed. Thus, a move toward

[5] The distributions are based on kernel density estimates.

[6] Other primitives, such as β and e_w or the nature of the shocks to TFP, can also influence TFP distributions, but we include no analysis of them here.

bimodality/more dispersion might be interpreted as a decrease in γ. What, then, is γ exactly? We have remained deliberately vague on this, and the Nelson-Phelps setting indeed is "ad hoc": it is not derived from first (technological) principles. As specified, γ captures two features at once. First, it captures something that we might label a technology (either physical technology or information technology), i.e., how easily *transferable* TFP knowledge is from the world to a given economy. With this interpretation, a fall in γ means weaker transferability. Given the rise of IT technology and, more generally, globalization, this interpretation would suggest that γ has risen, as opposed to fallen.

The other feature γ captures, however, is how much of the TFP investments are *internalized*, as opposed to how much they are treated as exogenous. Here, in contrast, at least to the extent that countries are able to better coordinate action across countries, and perhaps also better internalize TFP externalities that occur within a given country, one would argue that a fall in γ over this period is a reasonable assumption. A conclusion from this discussion is that one would, perhaps, want a model that goes beyond the simple Nelson-Phelps formulation by making the distinction between transferability (a technology/information concern) and the degree of internalization of the dynamics of TFP (a concern about how well countries are able to do this locally and how well they are able to coordinate between them).

8.2 Trading costs

Globalization was argued above to at least potentially be a determinant of γ. However, since γ also captures other important model features, a cleaner experiment is to consider a decrease in the costs of trading. Our theory so far assumes that e is an input that is traded in world markets at a competitively determined price. Consider instead the polar opposite case, namely that where e cannot be traded at all. Thus, with symmetric endowments, all countries would invest the same amount, e_W, in TFP accumulation. The TFP accumulation equation thus reads

$$T_{i,t+1} = T_{i,t}^{1-\gamma}\bar{T}_t^{\gamma}H(e_W),$$

which can be rewritten as

$$\frac{T_{i,t+1}}{\bar{T}_{t+1}} = \left(\frac{T_{i,t}}{\bar{T}_t}\right)^{1-\gamma} \frac{H(e_W)\bar{T}_t}{\bar{T}_{t+1}} = \left(\frac{T_{i,t}}{\bar{T}_t}\right)^{1-\gamma},$$

since $\bar{T}$ grows at rate $H(e_W)$. Thus, relative TFPs converge to one: there is global convergence. More generally, based on the comparison of the extreme cases, with trade costs in the fundamental input into the accumulation of TFP, we would thus expect to see a force toward convergence. A fall in trade costs would then produce a force away from convergence and toward the bimodal distribution of world TFP.

8.3 Other model parameters and transitional dynamics

We saw in Section 6.1 that the function H, which translates TFP investment input, e, into TFP growth, is an important determinant behind the distribution of world TFP; recall, e.g., that a necessary condition for bimodality in an ABGE based on interior solutions is that $H(e)^{\frac{1-\gamma}{\gamma}}H'(e)$ be non-monotonic. First, notice that if H is flat, i.e., $H(e)$ does not depend on e, asymmetric outcomes are never possible. Thus, if "R&D has become more productive", perhaps due to basic exogenous technical change, we would see a force toward bimodality. The shape of H would matter too, and using our functional form for H, it appears that higher curvature, i.e., more rapidly decreasing returns, also is a force toward bimodality.

Finally, transitional dynamics can of course explain changes in the TFP distribution. That is, suppose that there are no changes in primitives but that these primitives imply a bimodal long-run outcome, and suppose that the initial TFP distribution is a unimodal one with less inequality in TFP than observed currently. Then we would see something like the observed changes play out over time. However, this requires additional understanding of the origins of the initial distribution. These origins, in turn, are likely due to technology, or information, or "country organization" being different in the present than the past. Again take the trade costs as an example: if no trade is possible, we would see an outcome of equal TFP in all countries, and if trade were made possible, there would then be gradual movement from an equal to an unequal TFP distribution. A similar history would likely be generated if TFP shocks were very minor in the past but have become significant now, or if H was flat in the past (R&D was unproductive) but is upward-sloping in modern economies.

9 Concluding comments

This paper tries to answer the question: if there is technological catch-up, as proposed by Nelson and Phelps (1966), and each country takes this into account while maximizing the utility of its citizens, what is the resulting equilibrium distribution of TFP? To this end, the paper presents a dynamic general equilibrium model where individual countries invest in a technology-enhancing input that is traded in world markets, and the accumulation of TFP is modeled according to the Nelson-Phelps specification. Even though all countries are treated symmetrically, the model can generate a nontrivial long-run world distribution of TFP.

The model predicts that if technological catch-up is important for TFP growth, the distribution of countries over TFP is symmetric. If, instead, technological catch-up is less important for TFP growth, the distribution is asymmetric: twin-peaked. There is one group of countries with high TFP in relative terms, with the remainder of the countries operating at a much lower TFP level. All countries grow at the same rate, but the high-TFP countries invest more in technology than the low-TFP countries. The catch-up term thus allows the low-TFP countries to grow at the same rate as the high-TFP countries and not fall further behind in relative terms. More generally, independently of whether bimodality is an important feature of the data, the present model does produce equilibrium world distributions of TFP that display more or less dispersion depending on the fundamental parameters of the model, and the same parameters that tend to generate bimodality tend also to produce more dispersion.

The analysis we present is only a first attempt at a theory of the world distribution of country TFP. A serious quantitative analysis would require moving toward a model with (i) capital accumulation; (ii) some extent of incomplete insurance against TFP shocks; and, last but not least, (iii) geographical detail, so that large countries are different than small countries (and, say, there is heterogeneity in γs). Such extensions would be very interesting to pursue in future work.

Bibliography

[1] Acemoglu, D. (2008), *Introduction to Modern Economic Growth.* Forthcoming at Princeton University Press.

[2] Acemoglu, D. and Ventura, J. (2002), "The World Income Distribution", *The Quarterly Journal of Economics*, Vol. 117, No. 2, pp. 659-694.

[3] Aghion, P. Howitt, P. and Mayer-Foulkes, D. (2005), "The Effect of Financial Development on Convergence: Theory and Evidence", *The Quarterly Journal of Economics*, Vol. 120, No. 1, pp. 173-222.

[4] Aiyagari, S. (1994), "Uninsured Idiosyncratic Risk and Aggregate Saving", *The Quarterly Journal of Economics*, Vol. 109, No. 3, pp. 659-684.

[5] Benhabib, J. and Spiegel, M. (1994), "The Role of Human Capital in Economic Development: Evidence from Aggregate Cross-Country Data", *Journal of Monetary Economics*, Vol. 34, pp. 143-173.

[6] Chari, V. Kehoe, P. and McGrattan, E. (1997), "The Poverty of Nations: A Quantitative Investigation", Federal Reserve Bank of Minneapolis Research Department Staff Report 204/JV.

[7] Hall, R. and Jones, C. (1999), "Why Do Some Countries Produce So Much More Output Per Worker Than Others?", *The Quarterly Journal of Economics*, Vol. 114, No.1, pp. 83-116.

[8] Heston, A. Summers, R. and Aten B. (2006), Penn World Table Version 6.2, Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania, September 2006.

[9] Howitt, P. (2000), "Endogenous Growth and Cross-Country Income Differences", *The American Economic Review*, Vol. 90, No. 4, pp. 829-846.

[10] Howitt, P. and Mayer-Foulkes, D. (2005), "R&D, Implementation, and Stagnation: A Schumpeterian Theory of Convergence Clubs", *Journal of Money, Credit, and Banking*, Vol. 37, No. 1, pp. 147-177.

[11] Jones, C. (2008), *Macroeconomics*, New York: W.W. Norton.

[12] Jones, C. (1998), *Introduction to Economic Growth*, New York: W.W. Norton.

[13] Klenow, P. and Rodríguez-Clare, A. (1997), "The Neoclassical Revival in Growth Economics: Has It Gone Too Far?" *NBER Macroeconomics Annual*, B. Bernanke and J. Rotemberg, eds. Cambridge: MIT Press.

[14] Kremer, M. Onatski, A. and Stock, J. (2001), "Searching for Prosperity", NBER Working Paper no 8250.

[15] Lucas, R. (1988), "On the Mechanics of Economic Development", *Journal of Monetary Economics*, Vol. 22, No. 1, pp. 3-42.

[16] Nelson, R. and Phelps, E. (1966), "Investment in Humans, Technological Diffusion, and Economic Growth", *The American Economic Review*, Vol. 56, No. 1/2, pp. 69-75.

[17] Parente, S. and Prescott, E. (1994), "Barriers to Technology Adoption and Development", *The Journal of Political Economy*, Vol. 102, No. 2, pp. 298-321.

[18] Prescott, E. (1998), "Needed: A Theory of Total Factor Productivity", *International Economic Review*, Vol. 39, No. 3, pp. 525-551.

[19] Quah, D. (1993), "Empirical cross-section dynamics in economic growth", *European Economic Review*, Vol. 37, No. 2/3, pp. 426-434.

[20] Quah, D. (1997), "Empirics for Growth and Distribution: Stratification, Polarization and Convergence Clubs", *Journal of Economic Growth*, Vol. 2, No. 1, pp. 27-59.

[21] Rebelo, S. (1991), "Long-Run Policy Analysis and Long-Run Growth", *The Journal of Political Economy*, Vol. 99, No. 3, pp. 500-521.

[22] Romer, P. (1986), "Increasing Returns and Long-Run Growth", *The Journal of Political Economy*, Vol. 94, No. 5, pp. 1002-1037.

Appendix

A1 Proofs

2-group stability of the BGE

2-group stability is determined by

$$\begin{aligned} g'(z_1) &= (1-\gamma)\, z_1^{-\gamma} \frac{H(E(z_1,z_1))}{f(z_1)} + z_1^{1-\gamma} \frac{H'(E(z_1,z_1))\,(E_1(z_1,z_1)+E_2(z_1,z_1))}{f(z_1)} \\ & \quad - z_1^{1-\gamma} H(E(z_1,z_1)) \frac{f'(z_1)}{f(z_1)^2}, \end{aligned} \tag{3.23}$$

which evaluated at $z_1 = 1$, $f(z_1) = H(e_W)$, and $E(z_1, z_1) = e_W$ equals

$$g'(1) = 1 - \gamma + \frac{H'(e_W)}{H(e_W)} \left(E_1(1,1) + E_2(1,1)\right) - \frac{f'(1)}{H(e_W)}.$$

The derivative $E_2(1,1)$ can be obtained from the market-clearing condition for high-skilled labor:

$$e_W = \varphi E(z_1, z_1) + (1-\varphi) E\left(\frac{1-\varphi z_1}{1-\varphi}, z_1\right).$$

The derivative of this expression with respect to z_1 is

$$\begin{aligned} \frac{\partial e_W}{\partial z_1} &= \varphi\left(E_1(z_1,z_1) + E_2(z_1,z_1)\right) + (1-\varphi) E_2\left(\frac{1-\varphi z_1}{1-\varphi}, z_1\right) \\ & \quad + (1-\varphi) E_1\left(\frac{1-\varphi z_1}{1-\varphi}, z_1\right) \frac{-\varphi}{1-\varphi}. \end{aligned} \tag{3.24}$$

Evaluated at $z_1 = z_2 = 1$ and set to equal zero, this equation gives

$$\varphi\left(E_1(1,1) + E_2(1,1)\right) + (1-\varphi)\left(E_2(1,1) + E_1(1,1)\frac{-\varphi}{1-\varphi}\right) = 0,$$

or, equivalently,

$$\varphi E_1(1,1) + \varphi E_2(1,1) + (1-\varphi) E_2(1,1) - \varphi E_1(1,1) = 0,$$

which implies that $E_2(1,1) = 0$.

The derivative of $f(z_1)$ is

$$\begin{aligned} f'(z_1) &= \varphi(1-\gamma)z_1^{-\gamma}H(E(z_1,z_1)) + \varphi z_1^{1-\gamma}H'(E(z_1,z_1))(E_1(z_1,z_1)+E_2(z_1,z_1)) \\ &\quad -\varphi(1-\gamma)\left(\frac{1-\varphi z_1}{1-\varphi}\right)^{-\gamma} H\left(E\left(\frac{1-\varphi z_1}{1-\varphi},z_1\right)\right) \\ &\quad +(1-\varphi)H'\left(E\left(\frac{1-\varphi z_1}{1-\varphi},z_1\right)\right)E_2\left(\frac{1-\varphi z_1}{1-\varphi},z_1\right)\left(\frac{1-\varphi z_1}{1-\varphi}\right)^{1-\gamma} \\ &\quad +(1-\varphi)H'\left(E\left(\frac{1-\varphi z_1}{1-\varphi},z_1\right)\right)E_1\left(\frac{1-\varphi z_1}{1-\varphi},z_1\right)\left(\frac{-\varphi}{1-\varphi}\right)\left(\frac{1-\varphi z_1}{1-\varphi}\right)^{1-\gamma}. \end{aligned}$$

Evaluated at $z_1 = z_2 = 1$ this expression becomes

$$\begin{aligned} f'(1) &= \varphi H'(e_W)(E_1(1,1)+E_2(1,1)) \\ &\quad +(1-\varphi)H'(e_W)\left(E_2(1,1)+E_1(1,1)\left(\frac{-\varphi}{1-\varphi}\right)\right). \end{aligned}$$

Using the previous result that $E_2(1,1)=0$, the expression reads

$$f'(1) = \varphi H'(e_W)E_1(1,1) + (1-\varphi)H'(e_W)\left(E_1(1,1)\left(\frac{-\varphi}{1-\varphi}\right)\right) = 0.$$

Inserting $E_2(1,1)=0$ and $f'(1)=0$ into $g'(1)$ gives

$$g'(1) = 1-\gamma+\frac{H'(e_W)}{H(e_W)}E_1(1,1),$$

which is equivalent to

$$\frac{\partial z'}{\partial z} = 1-\gamma+\frac{H'(e_W)}{H(e_W)}E_1(1,1).$$

Hence, in the SBGE, 2-group stability is equivalent to measure-zero stability.

We obtain $E_1(1,1)$ as follows. Let the Euler equation in recursive notation be denoted F, where

$$F = \frac{\beta}{\hat{w}}z^{1-\gamma}\frac{H(E(z,z_1))}{f(z_1)} - \frac{H(E(z,z_1))}{H'(E(z,z_1))} + (1-\gamma)\beta\frac{H\left(E\left(z^{1-\gamma}\frac{H(E(z,z_1))}{f(z_1))},g(z_1)\right)\right)}{H'\left(E\left(z^{1-\gamma}\frac{H(E(z,z_1))}{f(z_1)},g(z_1)\right)\right)}. \tag{3.25}$$

Taking the first-order condition gives

$$(E_1(z,z_1)))^2 + pE_1(z,z_1) + q = 0,$$

where

$$p = \frac{\left(\frac{\beta}{\hat{w}}\frac{z^{1-\gamma}}{f(z_1)}H'(E(z,z_1)) - 1 + \frac{H''(E(z,z_1))H(E(z,z_1))}{(H'(E(z,z_1))^2} + (1-\gamma)^2\beta z^{-\gamma}\frac{H(E(z,z_1))}{f(z_1)}\right)}{(1-\gamma)\beta\frac{z^{1-\gamma}}{f(z_1)}\left(H'(E(z,z_1)) - \frac{H'(E(z,z_1))H(E(z,z_1))H''(E(z,z_1))}{(H'(E(z,z_1)))^2}\right)} - \frac{(1-\gamma)^2\beta z^{-\gamma}\frac{(H(E(z,z_1)))^2H''(E(z,z_1))}{f(z_1)(H'(E(z,z_1)))^2}}{(1-\gamma)\beta\frac{z^{1-\gamma}}{f(z_1)}\left(H'(E(z,z_1)) - \frac{H'(E(z,z_1))H(E(z,z_1))H''(E(z,z_1))}{(H'(E(z,z_1)))^2}\right)}$$

and

$$q = \frac{H(E(z,z_1))}{\hat{w}z\left(H'(E(z,z_1)) - \frac{H'(E(z,z_1))H(E(z,z_1))H''(E(z,z_1))}{(H'(E(z,z_1)))^2}\right)}.$$

Hence

$$E_1(z,z_1) = -\frac{p}{2} \pm \sqrt{\left(\frac{p}{2}\right)^2 - q}, \tag{3.26}$$

where p and q are given above. This expression evaluated at $z = z_1 = 1$ gives $E_1(1,1)$.

2-group stability of the ABGE

As in the case of the SBGE, 2-group stability is given by (3.23). In order to evaluate $g'(z_1)$ in a given ABGE, the derivatives $E_1(z_1,z_1)$, $E_2(z_1,z_1)$, and $f'(z_1)$ must be computed. We solve for $E_1(z_1,z_1)$ and $E_1(z_2,z_1)$ by evaluating (3.26) at $z = z_1$, and $z = z_2$, respectively. The expression for $E_1(z_1,z_1)$ corresponding to the stable root is inserted into $g'(z_1)$. Inspection of (3.24) reveals that when evaluated at $z = z_1$ it does not imply that $E_2(z_1,z_1) = 0$. Consequently, $f'(z_1)$ is not zero and both $E_2(z_1,z_1)$ and $f'(z_1)$ must be computed. In order to solve for $g'(z_1)$, we construct the following system of equations. The first two equations are the derivative with respect to z_1 of the recursive version of the Euler equation, (3.25), evaluated at $z = z_1$ and $z = z_2$, respectively. The third equation states that $\frac{\partial e_W}{\partial z_1} = 0$. Combining these equations with $f'(z_1)$ and $g'(z_1)$ yields the following system of five equations:

1. $\left.\frac{\partial F}{\partial z_1}\right|_{z=z_1} = 0;$

2. $\left.\frac{\partial F}{\partial z_1}\right|_{z=z_2} = 0;$

3.

$$\begin{aligned}
&\varphi\left(E_1(z_1,z_1)+E_2(z_1,z_1)\right)+(1-\varphi)E_2\left(\frac{1-\varphi z_1}{1-\varphi},z_1\right)\\
&+(1-\varphi)E_1\left(\frac{1-\varphi z_1}{1-\varphi},z_1\right)\frac{-\varphi}{1-\varphi}\\
=\ &0;
\end{aligned}$$

4.

$$\begin{aligned}
g'(z_1) \;=\;& (1-\gamma)\,z_1^{-\gamma}\frac{H(E(z_1,z_1))}{f(z_1)}+z_1^{1-\gamma}\frac{H'(E(z_1,z_1))\,(E_1(z_1,z_1)+E_2(z_1,z_1))}{f(z_1)}\\
&-z_1^{1-\gamma}H(E(z_1,z_1))\frac{f'(z_1)}{f(z_1)^2};
\end{aligned}$$

and

5.

$$\begin{aligned}
f'(z_1) \;=\;& \varphi\,(1-\gamma)\,z_1^{-\gamma}H(E(z_1,z_1))+\varphi z_1^{1-\gamma}H'(E(z_1,z_1))(E_1(z_1,z_1)+E_2(z_1,z_1))\\
&-\varphi\,(1-\gamma)\left(\frac{1-\varphi z_1}{1-\varphi}\right)^{-\gamma}H\left(E\left(\frac{1-\varphi z_1}{1-\varphi},z_1\right)\right)\\
&+(1-\varphi)H'\left(E\left(\frac{1-\varphi z_1}{1-\varphi},z_1\right)\right)E_2\left(\frac{1-\varphi z_1}{1-\varphi},z_1\right)\left(\frac{1-\varphi z_1}{1-\varphi}\right)^{1-\gamma}\\
&+(1-\varphi)H'\left(E\left(\frac{1-\varphi z_1}{1-\varphi},z_1\right)\right)\\
&E_1\left(\frac{1-\varphi z_1}{1-\varphi},z_1\right)\left(\frac{-\varphi}{1-\varphi}\right)\left(\frac{1-\varphi z_1}{1-\varphi}\right)^{1-\gamma}.
\end{aligned}$$

The system has five unknown variables: $f'(z_1)$, $g'(z_1)$, $E_2(z_1,z_1)$, $E_2(z_2,z_1)$, and $w'(z_1)$. This system yields a second-order equation in $E_2(z_1,z_1)$. Therefore, it has two solutions, and two corresponding expressions for $g'(z_1)$. If, for a given triplet e_1, e_2, and φ, one of the expressions is larger than 1 in absolute value and the other smaller than 1 in absolute value, then the ABGE exhibits saddle-path stability.

A2 Tables and figures

Table 3.1: SBGE for an example economy

γ	*existence*	$\hat{w}$	g	e	*measure-zero stability*	*2-group stability*
0.226	yes	6.18	1.34	0.1	stable	stable
0.224	yes	6.22	1.34	0.1	stable	stable
0.222	yes	6.25	1.34	0.1	unstable	unstable
0.220	yes	6.29	1.34	0.1	unstable	unstable
0.218	yes	6.33	1.34	0.1	unstable	unstable
0.216	yes	6.37	1.34	0.1	unstable	unstable
0.214	no	—	—	—	—	—

Explanatory notes:

$\hat{w}$: TFP-adjusted wage rate
g: TFP growth rate
e: amount of high-skilled workers employed
measure-zero stability: stability with respect to single-country perturbations
2-group stability: stability with respect to two-group perturbations

Table 3.2: Balanced growth equilibria for an example economy

	SBGE		*Range of ABGE*			
γ	*stability*	g	*gap*	$\hat{w}$	g	φ
0.226	stable	1.341	—	—	—	—
0.220	unstable	1.341	0.189-0.202	45.33-44.42	1.342-1.342	0.161-0.603
0.214	—	—	0.093-0.100	49.58-46.85	1.346-1.345	0.330-0.734
0.208	—	—	0.052-0.058	54.47-49.09	1.353-1.350	0.431-0.802
0.202	—	—	0.031-0.035	60.08-51.12	1.363-1.356	0.505-0.846
0.196	—	—	0.019-0.022	66.56-52.95	1.376-1.363	0.562-0.877

Explanatory notes:

g: TFP growth rate
gap: ratio of TFP between low- and high-TFP group
$\hat{w}$: TFP-adjusted wage rate
φ: share of countries in low-TFP group

Table 3.3: Range of ABGE for an example economy

γ	*gap*	$\hat{w}$	g	φ	*measure-zero stability*	*2-group stability*
0.220	0.189-0.202	45.33-44.42	1.342-1.342	0.161-0.603	stable	stable
0.214	0.093-0.100	49.58-46.85	1.346-1.345	0.330-0.734	stable	stable
0.208	0.052-0.058	54.47-49.09	1.353-1.350	0.431-0.802	stable	stable
0.202	0.031-0.035	60.08-51.12	1.363-1.356	0.505-0.846	stable	stable
0.196	0.019-0.022	66.56-52.95	1.376-1.363	0.562-0.877	stable	stable

Explanatory notes:

gap: ratio of TFP between low- and high-TFP group
$\hat{w}$: TFP-adjusted wage rate
g: TFP growth rate
φ: share of countries in low-TFP group
measure-zero stability: stability with respect to single-country perturbations
2-group stability: stability with respect to 2-group perturbations

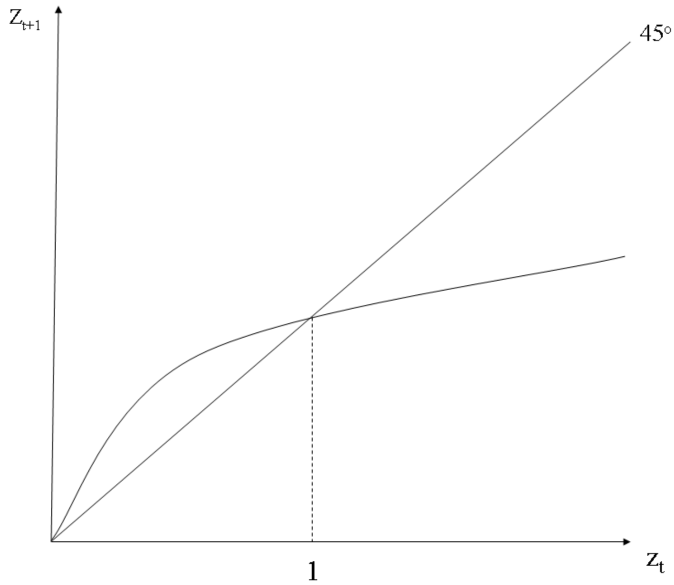

Figure 3.1: Example of a country's policy function.

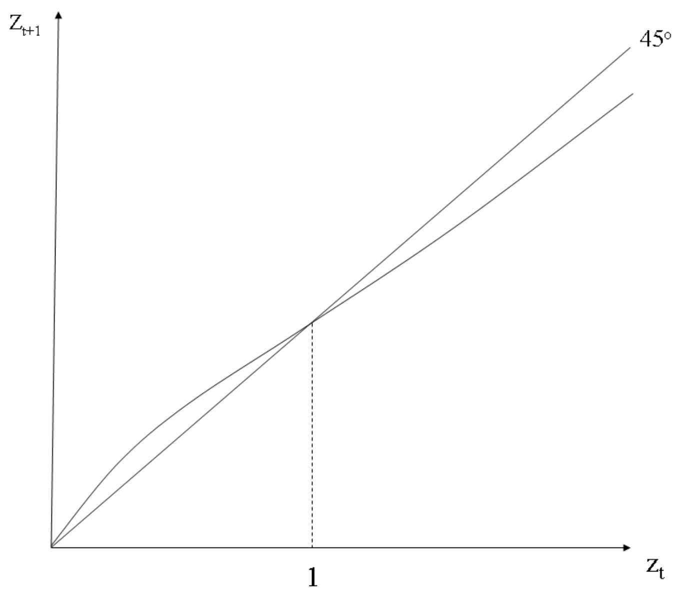

Figure 3.2: Example of a country's policy function.

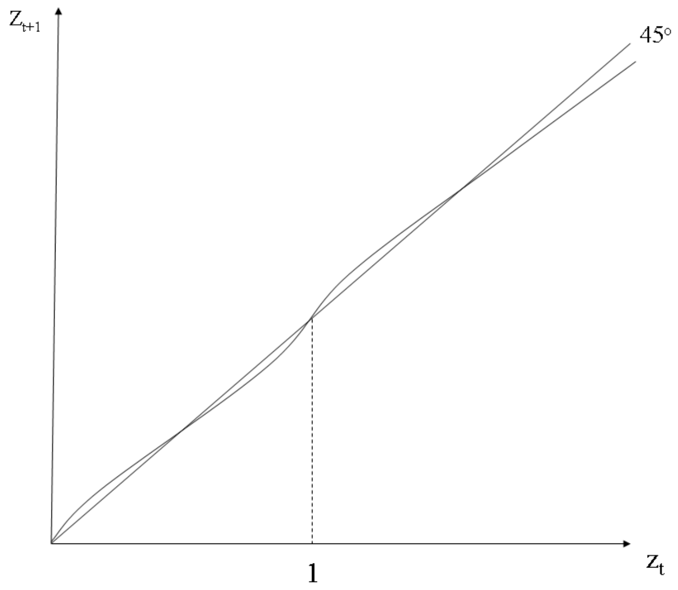

Figure 3.3: Example of a country's policy function.

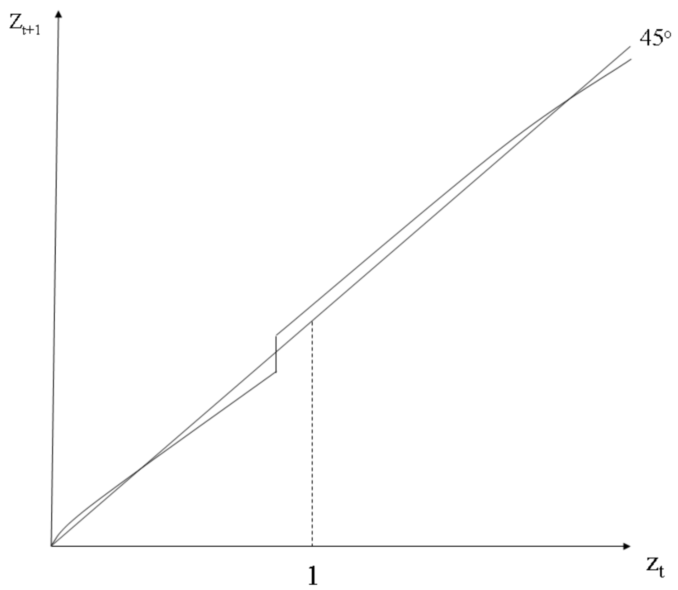

Figure 3.4: Example of a country's policy function.

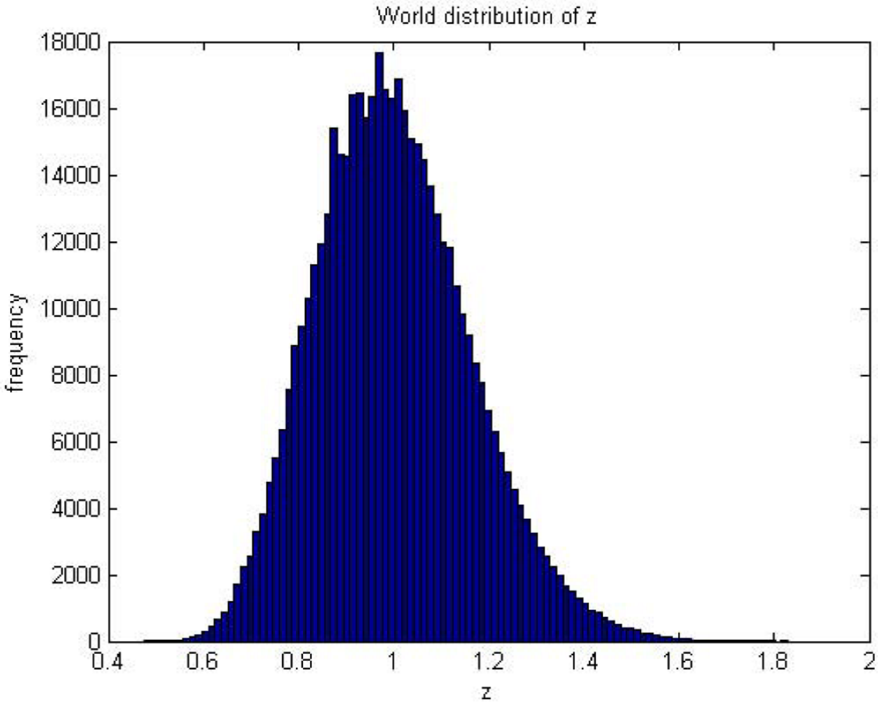

Figure 3.5: World distribution of relative TFP for $\gamma = 0.24$.

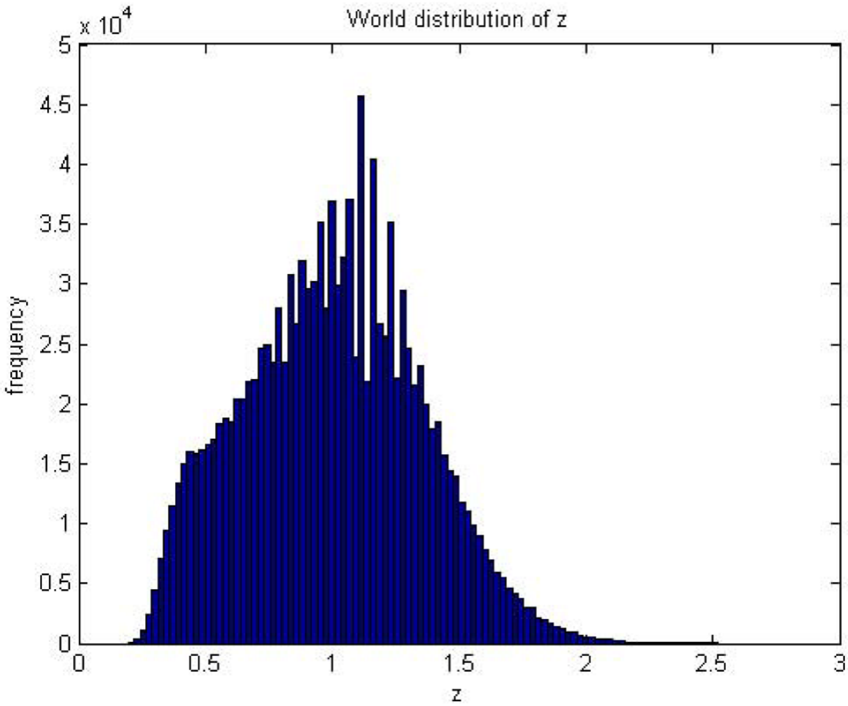

Figure 3.6: World distribution of relative TFP for $\gamma = 0.225$.

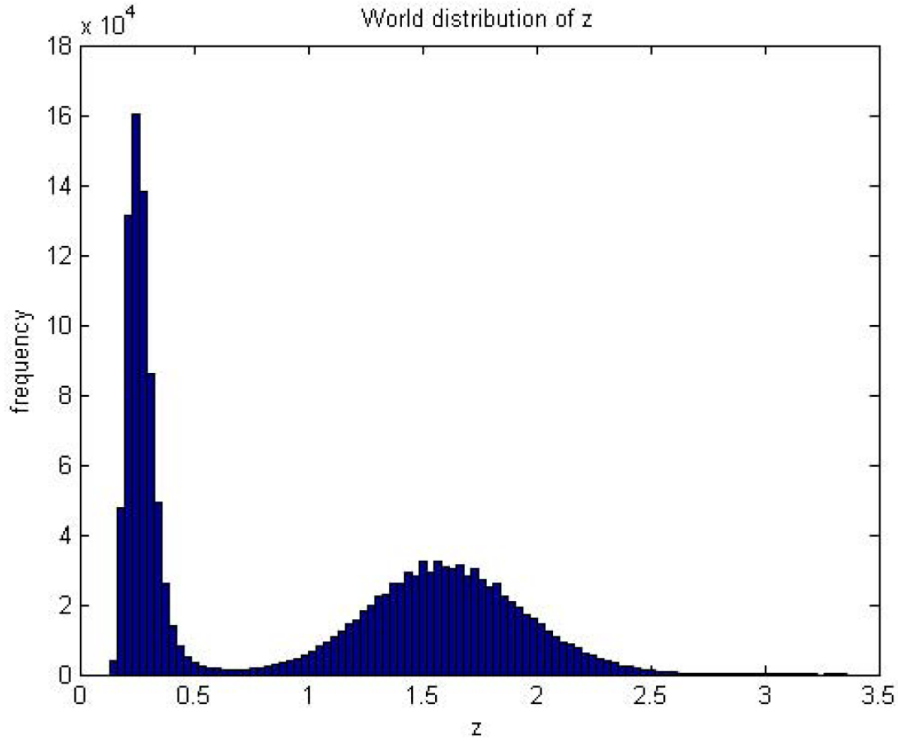

Figure 3.7: World distribution of relative TFP for $\gamma = 0.22$

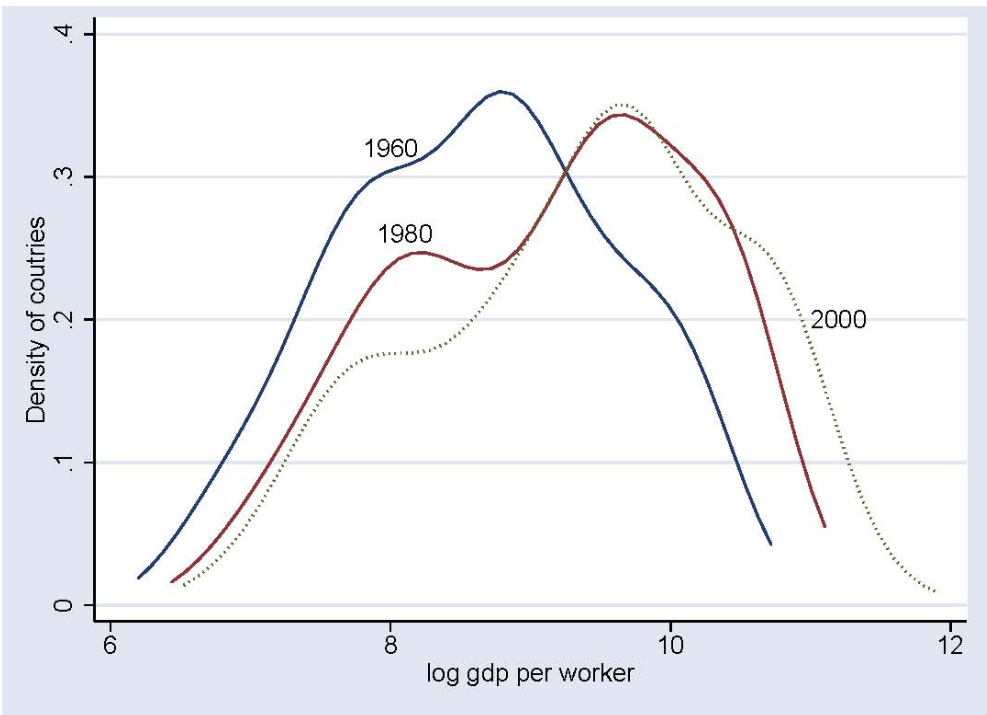

Figure 3.8: Distributions of relative world labor productivity (from Acemoglu 2008).

Chapter 4

Exhaustible Resources, Technology Choice and Industrialization of Developing Countries*

1 Introduction

In recent years, the industrialization of large developing countries, such as China and India, has generated a considerable increase in demand for exhaustible resources, for example copper, aluminum, iron ore and oil. ABARE (2008) reports that China and India accounted for about 35 percent of global steel consumption in 2007 and China alone accounted for about one third of world consumption of aluminum. Between 1990 and 2006, China's and India's total oil consumption increased by about 180 percent.[1] The increase in resource use has resulted in increases in the extraction of resources and through higher resource prices it has had an impact on resource-importing countries in the rest of the world.

* I would like to thank John Hassler and Per Krusell for valuable advice, and Olle Folke and seminar participants at IIES Brown Bag Seminar for comments and suggestions. I would also like to thank Christina Lönnblad for editorial assistance. All remaining errors are my own.

[1] Energy Information Administration, www.eia.doe.gov.

> "China's hunger for natural resources has set off a global commodity boom. Developed countries worry about being left high and dry "
>
> *The Economist, March 13th 2008*

This development has contributed to a renewed interest in the management of exhaustible resources. At what rate should they be extracted? When will exhaustible resources be substituted for renewable resources and how will the transition take place? In order to address such questions, this paper constructs a dynamic model of the world economy that exhibits two production technologies; a resource technology which uses an exhaustible resource as input, and an alternative technology which does not. Both technologies produce using capital and labor. In each time period, the optimal rate of extraction of the exhaustible resource and the optimal technology choice are jointly determined.

The questions asked in this paper are: given that there exist two different technologies in the economy, one which uses an exhaustible resource and an alternative which does not, what will be the optimal path of resource extraction over time? At what point in time will the alternative technology be adopted? How are the time path of extraction and the adoption of the alternative technology affected by the industrialization of developing countries?

The main finding is that the technology choice depends on the relative sizes of capital and resource stocks. If the capital stock is large in relation to the resource stock, the alternative technology is immediately adopted. The two technologies coexist until the resource is abandoned, and there is a complete switch to the alternative technology. If, instead, the capital stock is small, only the resource technology is used initially and the alternative technology is adopted with a delay. In addition, the resource is abandoned at a later point in time. Similarly, the time path of resource extraction depends on the relative sizes of capital and resource stocks. If the capital stock is large in relation to the resource stock, resource extraction is decreasing over time. If, instead, the capital stock is small, resource extraction has the shape of an inverse U; it is first increasing and then decreasing.

The paper also analyzes the effects of industrialization of developing countries on one of our most important exhaustible resources: oil. The model is calibrated to match the world production of crude oil during the period 1990-2000. Industrialization is modeled as an increase in the supply of labor. The calibrated model predicts that the industrialization of developing countries has the following effects on the world economy: the rate of oil extraction increases immediately, the alternative technology is adopted earlier, and oil as an energy source is abandoned earlier.

This paper is organized as follows. The related literature is presented in Section 2. Section 3 describes the model. Section 4 characterizes the equilibrium in a two-period setting. First, a baseline model is constructed, for which it is possible to derive analytical solutions. The restrictive assumptions of the baseline model are then relaxed in the full model and the implications for technology choice and extraction paths are analyzed. Section 5 extends the model to an infinite time horizon. Section 6 entails a calibration of the model to the case of oil and an analysis of the effects of industrialization of developing countries on the choice of technology for energy production and oil extraction. Section 7 concludes.

2 Related literature

One of the first models of optimal extraction of exhaustible resources introducing an alternative to the exhaustible resource, a backstop technology, was formulated by Nordhaus (1973). He described the backstop technology as an ultimate technology using a superabundant resource and capital as inputs. In subsequent articles, several economists have included backstop inputs in models of exhaustible resources. Examples are Dasgupta and Heal (1974), Kamien and Schwartz (1978), and Dasgupta and Stiglitz (1981). In these examples, the backstop input simply delivers a given stream of utility, or is a perfect substitute for the exhaustible resource in production. However, Dasgupta and Heal (1974) argue that there is every reason to suppose that the backstop is *not* a perfect substitute, but the modeling choice is motivated by a desire to keep the presentation simple. The models allow for an uncertain arrival of the backstop input. In Dasgupta and Heal (1974) and Dasgupta and Stiglitz (1981), the process is exogenous, while in Kamien and Schwartz (1978), it is determined by investments in R&D. Nevertheless, as soon as the backstop input has arrived, the exhaustible resource is abandoned.

In contrast, Tsur and Zemel (2005) present a model where the exhaustible resource and the backstop input both exist from the beginning of time. They are perfect substitutes in the production function for output and the cost of using the backstop input can be gradually reduced by investment in R&D. They find that the exhaustible resource and the backstop are used simultaneously until the resource is depleted. The model predicts that the growth path of the economy depends on its production technology and its endowment of physical and human capital.

A weakness of most of the models in this literature is that they either assume that resource extraction is costless or model the cost as a function of the extraction rate only. However, the cost of extraction depends on the remaining stock of resource in a given deposit, as well as on the extraction rate. Several factors account for a negative relationship between extraction cost and the remaining stock. As reported by Young (1992), in the case of minerals, the most accessible parts of a deposit are extracted first. Depletion of the stock forces the firm to move to less accessible parts, where unfavorable roof and/or floor conditions increase the extraction costs. In the case of oil and gas, the increase in extraction costs stems from a decrease in pressure in the oil or gas field, as the amount of remaining oil or gas decreases. A number of empirical studies have found that extraction costs increase as the remaining stock decreases, for example the works by Pesaran (1990) and Lin (2008) on oil extraction. Halvorsen and Smith (1991) and Young (1992) analyze the metal mining industry and find significant stock effects.

Tahvonen and Salo (2001) develop a model of renewable and nonrenewable energy resources. In the model, the extraction cost for the nonrenewable resource depends on the remaining stock of resource. The exhaustible and the renewable resource both exist from the beginning of time and are perfect substitutes. The authors find that the renewable and the nonrenewable resource will be used simultaneously during a transition period and that the use of nonrenewable resources starts at zero, reaches a maximum, and then approaches zero.

This paper constructs a model with two production technologies; one which uses an exhaustible resource and an alternative technology which does not. I argue that in most cases, a substitute for an exhaustible resource requires the use of a different production technology or production method. For example, in the case of fossil fuels, production of energy from solar or hydro power requires a production technology which is quite different from production of energy from oil. This implies that both capital and labor must be separately allocated to each of the technologies. Given that the amount of capital and labor in the economy is not unlimited, the decision of how to allocate capital and labor across the technologies affects equilibrium outcomes. Therefore, this paper features a dynamic model with two distinct technologies and the capital stock and the labor supply are optimally divided between them. In addition, in this paper, the cost of extracting the resource is a function of the remaining stock of resource, which is in accordance with the empirical findings.

3 The Model

The economy has a stock of capital, denoted K, an endowment of labor, denoted L, and a stock of an exhaustible natural resource in the ground, denoted Q. The resource can be extracted at a flow cost $R(M, Q)$, which is a function of the remaining stock of resource and the extraction rate, denoted M. There is no growth in the labor force. The economy has two technologies available for production of output. First, a resource technology, denoted $F(M, K, L)$, which uses the exhaustible resource, capital and labor as inputs and second, an alternative technology, denoted $G(K, L)$, which has capital and labor as inputs. The resource technology has the following functional form

$$F(M, K, L) = B((K - K_A)^{\theta}(L - L_A)^{1-\theta})^{\alpha_1} M^{\alpha_2} \tag{4.1}$$

where K_A is the amount of capital allocated to the alternative technology and L_A is the amount of labor allocated to the alternative technology. $B > 0$, $\alpha_1 \in (0, 1)$, $\alpha_2 \in (0, 1)$ and $\theta \in (0, 1)$.

The alternative technology has the following functional form

$$G(K, L) = AK_A^{\theta} L_A^{1-\theta} \tag{4.2}$$

where $A > 0$. To make the problem interesting, the productivity of the alternative technology, A, must be such that if the resource is abundant, the resource technology is used. Implicit in this formulation are the assumptions that capital is not technology-specific and that capital and labor are used in the same proportions in both technologies.

Both assumptions are made for simplicity. However, as a robustness check, I relax the latter assumption in Section 6 and the main results are robust to this change.

The flow cost of resource extraction, $R(M,Q)$, has the following functional form

$$R(M,Q) = C(Q^{1-\sigma} - (Q-M)^{1-\sigma}) \tag{4.3}$$

where $C > 0$ and $\sigma \in (0,1)$. The cost function has the following properties: $R_Q(M,Q) < 0$ and $R_{QQ}(M,Q) > 0$. The cost of extraction increases as the stock of resource decreases, and the incremental cost due to stock effects rises with the depletion of the stock. In addition, it is assumed that the cost of extracting the very last amount of a resource is prohibitively high, which follows from the specification of the extraction cost function:

$$\lim_{M \to Q} \frac{\partial R(M,Q)}{\partial M} \to \infty. \tag{4.4}$$

The economy has one representative consumer, whose utility function is $U(c)$, where c denotes consumption. The model is solved for the centralized equilibrium. Given initial conditions K_0 and Q_0, the social planner faces the following optimization problem

$$\begin{aligned}
&\max_{\{M_t, K_{A,t}, L_{A,t}, K_{t+1}\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t U(c_t) \\
&s.t. \\
&Q_{t+1} = Q_t - M_t \\
&c_t = AK_{A,t}^{\theta} L_{A,t}^{1-\theta} + B((K_t - K_{A,t})^{\theta}(L - L_{A,t})^{1-\theta})^{\alpha_1} M_t^{\alpha_2} \\
&\qquad -C(Q_t^{1-\sigma} - (Q_t - M_t)^{1-\sigma}) - K_{t+1} \\
&0 \leq M_t \leq Q_t, 0 \leq K_{A,t} \leq K_t, 0 \leq L_{A,t} \leq L \\
&c_t \geq 0 \\
&\forall t.
\end{aligned} \tag{4.5}$$

In each time period t, the social planner chooses extraction rate M_t, the amount of capital allocated to the alternative technology $K_{A,t}$, the amount of labor allocated to the alternative technology $L_{A,t}$, and the aggregate capital stock to enter next period K_{t+1}, so as to maximize the utility of the representative consumer. Savings are chosen in terms of the aggregate capital stock and in the subsequent time period, the social planner determines its division between the two technologies. There is a borrowing constraint, such that savings cannot exceed output, formulated as a non-negativity constraint on consumption.

4 Two time periods

To illustrate the factors affecting the optimal choices of extraction rate and technology, the optimization problem is first cast in a two-period setting and thereafter time is extended to an infinite horizon.

In a two-period setting, the optimization problem in (4.5) is

$$\begin{aligned}
&\max_{\left\{M_t, K_{A,t}, L_{A,t}, K_{t+1}\right\}_{t=1}^{2}} U(c_1) + \beta U(c_2) \\
&s.t. \\
&Q_2 = Q_1 - M_1 \\
&c_1 = AK_{A,1}^{\theta} L_{A,1}^{1-\theta} + B((K_1 - K_{A,1})^{\theta}(L - L_{A,1})^{1-\theta})^{\alpha_1} M_1^{\alpha_2} \\
&\qquad - C(Q_1^{1-\sigma} - (Q_1 - M_1)^{1-\sigma}) - K_2 \\
&c_2 = AK_{A,2}^{\theta} L_{A,2}^{1-\theta} + B((K_2 - K_{A,2})^{\theta}(L - L_{A,2})^{1-\theta})^{\alpha_1} M_2^{\alpha_2} \\
&\qquad - C(Q_2^{1-\sigma} - (Q_2 - M_2)^{1-\sigma}) \\
&0 \leq M_1 \leq Q_1, 0 \leq K_{A,1} \leq K_1, 0 \leq L_{A,1} \leq L \\
&0 \leq M_2 \leq Q_2, 0 \leq K_{A,2} \leq K_2, 0 \leq L_{A,2} \leq L \\
&c_1, c_2 \geq 0
\end{aligned} \tag{4.6}$$

The optimal allocations are now described in turn, starting with the interior solutions.

4.1 Interior solutions

In an interior solution in period 2, $M_2^* \in (0, Q_2)$, $K_{A,2}^* \in (0, K_2)$, and $L_{A,2}^* \in (0, L)$ satisfy the following system of equations

$$\begin{aligned}
&\alpha_2 B((K_2 - K_{A,2})^\theta (L - L_{A,2})^{1-\theta})^{\alpha_1} M_2^{\alpha_2 - 1} = (1-\sigma) C (Q_2 - M_2)^{-\sigma} \\
&A K_{A,2}^{\theta-1} L_{A,2}^{1-\theta} = \alpha_1 B (K_2 - K_{A,2})^{\theta\alpha_1 - 1} (L - L_{A,2})^{(1-\theta)\alpha_1} M_2^{\alpha_2} \\
&A K_{A,2}^{\theta} L_{A,2}^{-\theta} = \alpha_1 B (K_2 - K_{A,2})^{\theta\alpha_1} (L - L_{A,2})^{(1-\theta)\alpha_1 - 1} M_2^{\alpha_2}.
\end{aligned} \tag{4.7}$$

The three equations in (4.7) are conditions for production efficiency in the second period. In the first equation, the term on the left-hand side is the marginal product of the resource. The term on the right-hand side is the marginal cost of extracting the resource. The second equation equalizes the marginal product of capital in period 2 across the two technologies. The third equation equalizes the marginal product of labor in period 2 across the two technologies.

In an interior solution in period 1, $M_1^* \in (0, Q_1)$, $K_{A,1}^* \in (0, K_1)$, and $L_{A,1}^* \in (0, L)$ satisfy the following system of equations

$$\begin{aligned}
&U'(c_1)(\alpha_2 B((K_1 - K_{A,1})^\theta (L - L_{A,1})^{1-\theta})^{\alpha_1} M_1^{\alpha_2 - 1} - (1-\sigma) C (Q_1 - M_1)^{-\sigma}) \\
&\qquad = \beta U'(c_2) C (1-\sigma) \left((Q_1 - M_1 - M_2)^{-\sigma} - (Q_1 - M_1)^{-\sigma}\right) \\
&\qquad A K_{A,1}^{\theta-1} L_{A,1}^{1-\theta} = \alpha_1 B (K_1 - K_{A,1})^{\theta\alpha_1 - 1} (L - L_{A,1})^{(1-\theta)\alpha_1} M_1^{\alpha_2} \\
&\qquad A K_{A,1}^{\theta} L_{A,1}^{-\theta} = \alpha_1 B (K_1 - K_{A,1})^{\theta\alpha_1} (L - L_{A,1})^{(1-\theta)\alpha_1 - 1} M_1^{\alpha_2}.
\end{aligned} \tag{4.8}$$

The first equation in (4.8) determines the optimal allocation of resource extraction across the two periods. The first term on the left-hand side is the marginal product of the resource, weighed by the marginal utility of consumption in period 1. The second term is the marginal cost of extracting the resource in period 1, weighed by marginal utility of consumption in period 1. The term on the right-hand side is the marginal value of not extracting in period 1, which yields a larger resource stock entering period 2. This value is weighed by the marginal utility of consumption in period 2. The second and third equations ensure production efficiency within the period. The second equation equalizes the marginal product of capital in period 1 across the two technologies, while the third equation equalizes the marginal product of labor in period 1 across the two technologies.

4.2 Corner solutions

The functional form of $R(M,Q)$ restricts the corner solutions as follows. As seen from (4.4), $M^* < Q$. Hence, it is never optimal to extract the entire resource stock. However, extraction can be zero. If it is optimal to set resource extraction to zero, it will also be optimal to allocate all capital and labor to the alternative technology. Hence, the allocation $M_t^* = 0$, $K_{A,t}^* = K_t$, and $L_{A,t}^* = L$ is one possible corner solution. The other possible corner solution arises if the alternative technology is not used at all; $K_{A,t}^* = 0$ and $L_{A,t}^* = 0$. Let the optimal extraction in this case be denoted $M_{c,t}$, where subscript c indicates the corner solution. In $t = 1$, $M_{c,t}$ is given by equation 1 in (4.8), with $K_{A,1} = 0$ and $L_{A,1} = 0$. In $t = 2$, $M_{c,t}$ is given by equation 1 in (4.7) with $K_{A,2} = 0$ and $L_{A,2} = 0$.

Each of the two corner solutions can arise in each time period. Consequently, there are several possible technology combinations to be chosen over time. For example, both technologies are used in both periods. Alternatively, only one of the technologies is used in both periods. Another alternative is that there is a switch of technologies between periods. To obtain more precise predictions for the technology choices made across the two time periods, additional assumptions are needed.

4.3 Saving

Savings in period 1 are determined by the following equation

$$\begin{aligned}
&U'(c_1) = \beta\theta\alpha_1 B(K_2 - K_{A,2})^{\theta\alpha_1 - 1}(L - L_{A,2})^{(1-\theta)\alpha_1} M_2^{\alpha_2} U'(c_2) \\
&s.t. \\
&K_2 \leq AK_{A,1}^{\theta} L_{A,1}^{1-\theta} + B((K_1 - K_{A,1})^{\theta}(L - L_{A,1})^{1-\theta})^{\alpha_1} M_1^{\alpha_2} \\
&\qquad -C(Q_1^{1-\sigma} - (Q_1 - M_1)^{1-\sigma}).
\end{aligned}$$

4.4 A baseline model

For the purpose of deriving analytical solutions for the technology choices in the two-period setting, I construct a very simplistic baseline model. The baseline model is one where the following holds: there is no labor, the utility function is linear in consumption, and the discount factor is equal to 1. More formally, it is assumed that the production functions are

$$F(M, K) = B(K - K_A)^{\alpha_1} M^{\alpha_2} \tag{4.9}$$

and

$$G(K) = AK_A, \tag{4.10}$$

and in addition, $U(c_t) = c_t$, and $\beta = 1$. Each of the assumptions will be relaxed in the next section, where a richer model is developed. Given the linear utility function, the parameter restriction $\beta A = 1$ is imposed.

4.4.1 Interior solutions

In an interior solution in period 2, $M_2^* \in (0, Q_2)$ and $K_{A,2}^* \in (0, K_2)$ satisfy the following system of equations

$$\begin{aligned} \alpha_2 B(K_2 - K_{A,2})^{\alpha_1} M_2^{\alpha_2 - 1} &= (1-\sigma)C(Q_2 - M_2)^{-\sigma} \\ A &= \alpha_1 B(K_2 - K_{A,2})^{\alpha_1 - 1} M_2^{\alpha_2}. \end{aligned} \tag{4.11}$$

As can be seen from the second equation in (4.11), the specification of production technologies implies that the marginal product of capital is constant in the alternative technology. In the resource technology, the marginal product is varying with the amount of resource input M. Under the assumption of constant returns to scale in the resource technology, an interior solution implies that the ratio of resource to capital in the resource technology is

$$\frac{M_2}{K_2 - K_{A,2}} = \frac{A^{\frac{1}{1-\alpha_1}}}{\alpha_1^{\frac{1}{1-\alpha_1}} B^{\frac{1}{1-\alpha_1}}}. \tag{4.12}$$

The expression in (4.12) shows that it is optimal to produce using resource and capital in fixed proportions, irrespective of the level of resource input. This implies that if the resource technology is used, it will be optimal to first allocate capital to the resource technology in fixed proportion to the resource and then allocate the remaining capital to the alternative technology. The resource-capital ratio is determined by the relative productivities of the two technologies. An increase in B decreases the resource-capital ratio, as more capital is allocated to the resource technology, for a given amount of resource input. Conversely, an increase in A

increases the resource-capital ratio, as more capital is allocated to the alternative technology.

In an interior solution in period 1, $M_1^* \in (0, Q_1)$ and $K_{A,1}^* \in (0, K_1)$ satisfy the following system of equations

$$\begin{aligned}
&\alpha_2 B(K_1 - K_{A,1})^{\alpha_1} M_1^{\alpha_2 - 1} - (1-\sigma) C (Q_1 - M_1)^{-\sigma} \\
&= \beta C (1-\sigma) \left((Q_1 - M_1 - M_2)^{-\sigma} - (Q_1 - M_1)^{-\sigma} \right) \\
&A = \alpha_1 B (K_1 - K_{A,1})^{\alpha_1 - 1} M_1^{\alpha_2}. \qquad (4.13)
\end{aligned}$$

It can be shown that the economy exhibits constant returns to scale. For details, see Appendix A1.

4.4.2 Corner solutions

In the baseline model, it is possible to solve analytically for the conditions under which the corner solutions arise. Starting with the corner solution $M_t^* = 0$, and $K_{A,t}^* = K_t$ the following holds.

Proposition 1 *If $\alpha_1 + \alpha_2 = 1$, there exists a $\hat{Q} > 0$ such that for $Q \leq \hat{Q}$, $M^* = 0$, where*

$$\hat{Q} = \frac{A^{\frac{\alpha_1}{\sigma(1-\alpha_1)}} C^{\frac{1}{\sigma}} (1-\sigma)^{\frac{1}{\sigma}}}{\alpha_2^{\frac{1}{\sigma}} B^{\frac{1}{\sigma(1-\alpha_1)}} \alpha_1^{\frac{\alpha_1}{\sigma(1-\alpha_1)}}}.$$

Proof. See Appendix A2. ■

Provided that the resource technology exhibits constant returns to scale, it will be optimal to refrain from extracting the resource at all if the resource stock is sufficiently small. The threshold value $\hat{Q}$ is increasing in A as higher productivity of the alternative technology makes it optimal to abandon the resource with more left in the ground. This value is decreasing in B, as higher productivity of the resource technology makes it optimal to extract more before the resource is abandoned. It is increasing in C and decreasing in σ, as higher costs of extraction make it optimal to abandon the resource with more left in the ground. $\hat{Q}$ is independent of the capital stock.

The second corner solution has $K_{A,t}^* = 0$ and $M_t^* = M_{c,t}$. In $t = 1$, $M_{c,t}$ is given by equation 1 in (4.13) with $K_{A,1} = 0$, and in $t = 2$, $M_{c,t}$ is given by the solution to equation 1 in (4.11) with $K_{A,2} = 0$. Under the assumption of constant returns to scale in the resource technology, there is a threshold value $\tilde{M}$ such that for $M_t \geq \tilde{M}$, $K^*_{A,t} = 0$, where

$$\tilde{M} = K \left(\frac{A}{\alpha_1 B} \right)^{\frac{1}{\alpha_2}}. \tag{4.14}$$

The threshold can be reformulated in terms of the capital stock. $K^*_{A,t} = 0$, if $K_t \leq \hat{K}(Q_t)$, where

$$\hat{K}(Q) = \left(\frac{\alpha_1 B}{A} \right)^{\frac{1}{\alpha_2}} \left(Q - \hat{Q} \right). \tag{4.15}$$

If the capital stock is sufficiently small in relation to the resource stock, all capital is allocated to the resource technology. The reason is, as described earlier, that if the resource technology is used, it is optimal to first allocate the available capital to the resource technology, and so that the resource-capital ratio is constant. The threshold $\hat{K}(Q)$ is increasing in Q, as a higher resource stock implies a higher total extraction and hence, a higher capital stock is required. It is decreasing in C and increasing in σ, as higher extraction costs decrease total extraction. It is increasing in B as higher productivity of the resource technology increases total extraction and decreases the resource-capital ratio. Finally, it is decreasing in A as a higher productivity of the alternative technology decreases total extraction and increases the resource-capital ratio.

It is possible to derive conditions under which there is an analytical solution for the total amount of resource extracted over the two periods.

Proposition 2 *If a pair* (Q_2, K_2) *is such that* $K_2 > \hat{K}(Q_2)$*, then* $M_1^* + M_2^* = M^T$*, where*

$$M^T = Q_1 - \hat{Q}.$$

Proof. See Appendix A2. ■

Total extraction is equal to M^T as long as the alternative technology is used at least in period 2. The intuition for this result is that as long as there is enough capital in at least one of the periods to enable the resource technology to be operated at the optimal ratio of resource to capital, extraction will continue until it becomes optimal to abandon the resource, which occurs at $\hat{Q}$. In this case, total extraction is independent of the capital stock.

4.4.3 Saving

Savings in period 1 are determined by

$$1 = \beta\alpha_1 B(K_2 - K_{A,2})^{\alpha_1 - 1} M_2^{\alpha_2}$$
$$s.t.$$
$$K_2 \leq AK_{A,1} + B(K_1 - K_{A,1})^{\alpha_1} M_1^{\alpha_2} - C(Q_1^{1-\sigma} - (Q_1 - M_1)^{1-\sigma}).$$

Under the parameter restriction $\beta A = 1$, the expressions for K_2 given by the above equation and the second equation in (4.11) are identical. Hence, the amount of capital allocated to the resource technology can be determined, but not K_2. Consequently, the social planner will be indifferent over initial capital stocks in period 2, as long as the capital stocks are sufficient to enable production with the resource technology at the optimal resource-capital ratio. If the initial capital stock in period 1 is small in relation to the resource stock, it will be optimal to save. The borrowing constraint implies that for some initial conditions, the amount of savings will be constrained. In this case, the social planner will not be indifferent over K_2; the maximum feasible value of K_2 yields strictly higher utility than any lower value. Hence, if the borrowing constraint binds, the optimal amount of savings is determined, whereas if it does not, savings are indeterminate.

4.4.4 Technology choice

As discussed in the previous section, several technology combinations can occur in the two-period setting. In the simple baseline model, it is possible to distinguish between four cases. The first case is when only the alternative technology is used in both periods. The second case is one in which the following holds. The alternative technology is always used in both periods. Meanwhile, the economy is indifferent between using the resource in the first period only, in the second period only, or in both. In the third case, the following holds. The resource technology is always used in both periods. Meanwhile, the economy is indifferent between using the alternative technology in the first period only, in the second period only, or in both. The fourth case entails the use of the resource technology only. Each case will now be discussed in turn.

4.4.5 Case 1

In case 1, there is no extraction of the resource in either period. This case arises if $Q_1 \leq \hat{Q}$. Naturally, it follows that $Q_2 \leq \hat{Q}$. The allocations are $M_1^* = 0$, $K_{A,1}^* = K_1$, $M_2^* = 0$, and $K_{A,2}^* = K_2$.

4.4.6 Case 2

In case 2, the alternative technology is used in both periods and there is positive resource extraction. Total extraction over both periods is equal to M^T, as defined in Proposition 2. The division of M^T into M_1 and M_2 cannot be determined; hence there are multiple solutions for M_1 and M_2. The allocations in case 2 are $K_{A,1}^* > 0$, $K_{A,2}^* > 0$, $M_1^* \in [0, M^T]$, and $M_2^* \in [0, M^T]$. The conditions under which case 2 arises are the following.

Proposition 3 *Case 2 arises if a pair* (K_1, Q_1) *is such that* $Q_1 > \hat{Q}$ *and* $K_1 > \hat{K}(Q_1)$.

Proof. See Appendix A2. ■

If the initial value of the capital stock is sufficiently large in relation to the resource stock, it follows that the alternative technology is always used in both periods and the social planner is indifferent between the following extraction paths. First, the extraction of M^T in period 1 and zero in period 2, second, the extraction of zero in period 1 and M^T in period 2 and third, any division of total extraction M^T such that extraction is positive in both periods. Hence, it is possible that the resource technology is used in the first period only, the second period only, or in both periods, while the alternative technology is always used in both periods. The indifference is due to the functional form for the cost of resource extraction. The extraction rate affects the extraction costs through its effect on the remaining resource stock and given linear utility and no discounting, this effect is constant across the two periods.

4.4.7 Case 3

In case 3, the resource technology is used in both periods and consequently extraction is positive in both periods; $M_1^* > 0$, and $M_2^* > 0$. However, the economy may use the alternative technology in both periods or switch between using the alternative technology in either of the time periods; $K_{A,1}^* > 0$ and $K_{A,2}^* > 0$, or $K_{A,1}^* > 0$ and $K_{A,2}^* = 0$, or $K_{A,1}^* = 0$ and $K_{A,2}^* > 0$. Using the threshold $\tilde{M}$, it is possible to define the conditions under which case 3 arises as follows.

Proposition 4 *Case 3 arises if a pair* (Q_1, K_1) *is such that* $Q_1 > \hat{Q}$, $K_1 \leq \hat{K}(Q_1)$, *and* $K_2 > \hat{K}(Q_2)$,

where

$Q_2 = Q_1 - \tilde{M}$, *and*

$K_2 = \frac{AK_1}{\alpha_1} - C\left(Q_1^{1-\sigma} - \left(Q_1 - \tilde{M}\right)^{1-\sigma}\right).$

Proof. See Appendix A2. ■

If the capital stock is sufficiently large for the alternative technology to be used in one period, but not in both, case 3 arises. The resource technology is used in both periods. The consumer is indifferent over producing with the alternative technology in the first period only, producing with the alternative technology in the second period only, or producing with the alternative technology in both periods. As in case 2, the indifference is due to the functional form for extraction costs, in combination with linear utility and no discounting.

4.4.8 Case 4

In case 4, the alternative technology is never adopted, while the resource technology is used both periods. The allocations are $K_{A,1}^* = K_{A,2}^* = 0$, $M_1^* > 0$, and $M_2^* > 0$. Again, with the use of $\tilde{M}$, the conditions under which case 4 arises can be stated as follows.

Proposition 5 *Case 4 arises if a pair* (Q_1, K_1) *is such that* $Q_1 > \hat{Q}$, $K_1 \leq \hat{K}(Q_1)$, *and* $K_2 \leq \hat{K}(Q_2)$,

where

$Q_2 = Q_1 - \tilde{M}$, *and*

$K_2 = \frac{AK_1}{\alpha_1} - C\left(Q_1^{1-\sigma} - \left(Q_1 - \tilde{M}\right)^{1-\sigma}\right).$

Proof. See Appendix A2. ■

If the initial capital stock is sufficiently small in relation to the resource stock that even after saving, the alternative technology is not adopted in period 2, case 4 arises.

4.4.9 Numerical example

To illustrate the four cases described in the previous section, the model is solved numerically. The parameterization is as follows. Suppose that the exhaustible resource is fossil fuels. Nordhaus (1992) argues that the share of energy to GDP is roughly 10 percent and therefore I set $\alpha_1 = 0.9$ and $\alpha_2 = 0.1$. As regards the extraction cost function, due to the lack of empirical estimations of σ, it is set to an intermediate value; $\sigma = 0.5$. Parameters B and C scale the functions $F(M, K)$ and $R(M, Q)$ and are chosen so as to ensure that all four cases of technology choice arise within the interval $K_1 \in [0, 6]$ and $Q_1 \in [0, 6]$. Given $A = 1$, this implies that $B = 1.12$ and $C = 0.15$. Varying these parameters does not affect the characteristics of the four cases, but the length of vectors $K_1 \in [K^{min}, K^{max}]$ and $Q_1 \in [Q^{min}, Q^{max}]$ for which all cases are represented.

Figure 4.1 depicts the technology choices made across both time periods given initial conditions K_1 and Q_1. In the figure, area 1 corresponds to case 1, area 2 to case 2, and so forth. The vertical line represents threshold $\hat{Q}$. The line separating areas 2 and 3 is given by threshold $\hat{K}(Q_1)$ and the line separating areas 3 and 4 is given by threshold $\hat{K}(Q_2)$. As seen in the figure, case 1 arises if the initial resource stock is low.

Case 2 arises if the initial capital stock is large in relation to the initial resource stock. Case 3 arises if the initial capital stock is smaller and finally, case 4 arises if the initial capital stock is even smaller.

Figure 4.1 can illustrate some comparative statics with respect to the different cases of technology choice. First, consider an increase in A, the productivity of the alternative technology. An increase in A shifts threshold $\hat{Q}$ to the right and threshold $\hat{K}(Q_1)$ downwards while the effect on $\hat{K}(Q_2)$ is ambiguous. When the alternative technology has a higher productivity, it is used in both periods for a larger set of initial conditions and the resource will be abandoned for a larger set of initial conditions. Second, consider an increase in B, the productivity of the resource technology. An increase in B shifts $\hat{Q}$ to the left and $\hat{K}(Q_1)$ upwards while the effect on $\hat{K}(Q_2)$ is ambiguous. If the productivity of the resource technology increases, the alternative technology is used in both periods for a smaller set of initial conditions. Similarly, the resource is abandoned for a smaller set of initial conditions.

Third, if the cost of resource extraction increases, either by an increase in C or a decrease in σ, $\hat{Q}$ shifts to the right and $\hat{K}(Q_1)$ shifts downwards, while the effect on $\hat{K}(Q_2)$ is ambiguous. An increase in extraction costs implies that the alternative technology will be used in both periods for a larger set of initial conditions and the resource abandoned for a larger set of initial conditions.

In the baseline model, it is possible to characterize the technology choices made for all combinations of initial conditions Q_1 and K_1. The technology choice belongs to one of four possible cases and is determined by the relative sizes of the capital stock and the resource stock. It is also possible to analyze how parameters of the production functions and the extraction cost function affect technology choice.

However, the model entails several assumptions which, while enabling analytical solutions, are quite restrictive. First, the utility function is assumed to be linear, second, the discount factor is equal to one and third, labor is not used for production. In the following section, these assumptions are relaxed and labor is reintroduced into the model. The next step is to investigate whether the characterizations of technology choice derived in the baseline model also hold true in the full model. In the following analysis, it will be necessary to resort to numerical solutions.

4.5 The model with log utility and labor

In this section, the restrictive assumptions of the baseline model are relaxed. The consumer has a log utility function $U(c_t) = \log(c_t)$ and values the future less than the present; $\beta < 1$. In addition, labor is reintroduced into both production functions, as described in Section 3.

The model is solved numerically using the same parameterization as in the baseline model; $\alpha_1 = 0.9$, $\alpha_2 = 0.1$, $\sigma = 0.5$, $B = 1.12$, $C = 0.15$, and $A = 1$. Nordhaus (1992) uses a labor income share of 0.6 and this paper follows his example. The parameter θ is set to $= 0.33$, which yields a capital income share of 0.3 and a labor income share of 0.6 for the resource technology. Following Nordhaus and Yang (1996), the yearly discount rate is 3 percent. One time period is 10 years, which gives $\beta = 0.74$. The supply of labor is set to $L = 16$.

The more realistic assumptions regarding preferences imply that the consumer wishes to smooth consumption and therefore, K_2 is uniquely determined. In addition, the consumer is no longer indifferent over in what time period extraction takes place. As a consequence, the technology choices in cases 2 and 3 can be more precisely described. In case 2, the alternative technology is used in both periods and the resource is extracted in period 1 and possibly in period 2 as well. In case 3, the resource is used in both periods and the alternative technology is not adopted in the first period, but in the second. Let case 1 be denoted the resource-abandonment case, case 2 be denoted the immediate-adoption case and case 3 the delayed-adoption case. Case 4 is then the resource-only case.

The allocations and time paths arising in the two most interesting cases, i.e. the immediate-adoption case and the delayed-adoption case, are illustrated by numerical examples. Figure 4.2 depicts the time path for the example of the immediate-adoption case, with $K_1 = 1$ and $Q_1 = 1$. In this case, the initial capital stock is sufficiently large in relation to the resource stock for the alternative technology to be adopted in the first period. As seen in Figure 4.2, almost the entire total resource extraction takes place in the first period.

Figure 4.3 depicts the time path for the example of the delayed-adoption case, with $K_1 = 1$ and $Q_1 = 8$. In the delayed-adoption case, all capital in period 1 is allocated to the resource technology. The initial capital stock is not sufficiently large for the alternative technology to be immediately adopted, but it is adopted in period 2. The low capital stock implies that a larger share of the total resource extraction is deferred to period 2.

In sum, introducing more realistic preferences as well as labor into the model implies that there are still four cases of technology choice to consider. The two most interesting of these cases are the immediate-adoption case and the delayed-adoption case. In the immediate-adoption case, the alternative technology is adopted in period 1 and both technologies are used simultaneously. In the delayed-adoption case, the alternative technology is adopted in period 2 and a larger share of total resource extraction is deferred to period 2. Although the conditions under which each case arises cannot be explicitly derived, the main features of this model are as those in the baseline model. The immediate-adoption case arises if the initial capital stock is large in relation to the resource stock, while the delayed-adoption case arises if the initial capital stock is relatively small in relation to the resource stock.

5 Infinite time horizon

This section analyzes the properties of the full model when time is extended to an infinite horizon. As will be shown below, there is one main effect of extending the number of time periods, namely that irrespective of the initial conditions, the economy will end up in the resource-abandonment case. However, the initial conditions determine at what point in time this occurs. Consequently, the resource-only case never arises. When the time horizon is extended, the alternative technology will eventually be adopted. Therefore, the economy either begins in the immediate-adoption case, or in the delayed-adoption case, and always ends up in the resource-abandonment case. Which case initially arises is determined by the same conditions as in the two-period model, namely the relative sizes of the capital stock and the resource stock. However, in contrast to the two-period model, the time path of resource extraction can now be increasing.

5.1 Solution method

For the purpose of solving the model for an infinite time horizon, the social planner's maximization problem is recast in a recursive manner, with value function $V(K, Q)$ and policy functions $E_i(K, Q)$, $i \in \{Q', K_A, L_A, K'\}$, where Q' and K' are next period's resource stock and capital stock, respectively. This is a stationary problem and therefore, the value function and the policy functions are independent of time. The optimization problem is

1. $\forall(K,Q)$, $V(K,Q)$ solves

$$\begin{aligned} V(K,Q) = & \max_{Q',K_A,L_A,K'} \log(c) + \beta V(K',Q') \\ s.t. & \\ c = & AK_A^{\theta}L_A^{1-\theta} + B((K-K_A)^{\theta}(L-L_A)^{1-\theta})^{\alpha_1}M^{\alpha_2} \\ & - C(Q^{1-\sigma} - (Q'^{1-\sigma}) - K' \\ M = & Q - Q' \end{aligned}$$

2. $\forall(K,Q)$, $E_i(K,Q) = \arg\max_i V(K,Q)$, $i \in \{Q',K_A,L_A,K'\}$.

The social planner has four choice variables; Q' (which is equivalent to choosing this period's extraction rate M), K_A, L_A, and K'. They are chosen so as to maximize the right-hand side of the expression for $V(K,Q)$ for all values of K and Q.

The model is solved using discretization of the state space and backward induction methods. The discretization of the state space implies that the social planner can choose next period's resource stock Q' and capital stock K' in discrete amounts $Q' \in \boldsymbol{Q}$ and $K' \in \boldsymbol{K}$, where $\boldsymbol{Q}$ and $\boldsymbol{K}$ are grids $[Q^1 < Q^2 < ... < Q^N]$ and $[K^1 < K^2 < ... < K^N]$. There are natural limits on the lower endpoints; $Q^1 = 0$ and $K^1 = 0$. Since the sequence $\{Q_t\}_{t=0}^{\infty}$ is weakly decreasing, Q^N can be set to the maximum initial value for which the model is to be solved. K^N must be set such that it well exceeds the steady-state level of capital when the resource is abandoned.

The backward induction solution method is then applied as follows. Starting from the last time period, T, the value function $V(K,Q,T)$ and the policy functions $E_i(K,Q,T)$ are given by the last period solution to the two-period problem. Using $V(K,Q,T)$ and $E_i(K,Q,T)$, it is possible to solve backwards for the optimal policy functions $E_i(K,Q,T-\tau)$ and value function $V(K,Q,T-\tau)$ for each grid point in the two-dimensional grid, for $\tau = 1,...,T$. As in the two-period model, both interior and corner solutions exist. The backward-solving procedure is repeated until $\|V(K,Q,T-\tau) - V(K,Q,T-\tau-1)\| < \epsilon \; \forall(K,Q)$, where ϵ is a small number.

5.2 Numerical solutions

In the numerical solution of the infinite horizon model, the parameterization is as follows; $\alpha_1 = 0.9$, $\alpha_2 = 0.1$, $\theta = 0.33$, $\sigma = 0.5$, $\beta = 0.74$, $L = 16$, and $A = 1.35$. As described above, parameters $B = 1.8$ and $C = 2.5$ are chosen so as to ensure that the different cases of technology choice arise within the chosen grid for K and Q. Given $A = 1.35$, the steady-state level of capital when the resource is abandoned is 3. The time paths for the economy for examples of different initial conditions K_0 and Q_0 are shown in Figures 4.4-4.6. The figures display the time paths for 10 periods, which is equal to one hundred years.

Figure 4.4 depicts the time path for the economy for $K_0 = 3$ and $Q_0 = 3.2$. A high initial capital stock relative to the resource stock implies that the time path of extraction is downward-sloping over the entire time period. The alternative technology is immediately adopted and the resource is abandoned in period 2. Figure 4.5 depicts the time path for the economy for $K_0 = 1.5$ and $Q_0 = 7$. A lower initial capital stock implies that the time path of resource extraction is initially flat and then downward-sloping. The alternative technology is not adopted until period 4 and the resource is abandoned in period 5.

Figure 4.6 depicts the time path for $K_0 = 0.5$ and $Q_0 = 9$. A substantially lower initial capital stock in relation to the resource stock implies that the time path of extraction is first increasing and then decreasing. The initial increase in resource extraction can be explained as follows. With a low initial capital stock, it is optimal to defer a larger part of the extraction to the future where savings have increased the capital stock, since with a higher capital stock, the amount of capital available to allocate to the resource technology is higher. When the capital stock has become sufficiently high, resource extraction starts to decrease as there is no longer any motive for deferring extraction. As seen in the figure, the alternative technology is not adopted until period 5 and the resource is abandoned in period 6.

Extending the model to an infinite time horizon implies that technology choice can be characterized as follows. As in the two-period model, technology choice is determined by the relative sizes of the resource stock and the capital stock. If the initial capital stock is large in relation to the resource stock, the alternative technology is immediately adopted and the time path of resource extraction is monotonically decreasing. If the capital stock is small, the adoption of the alternative technology is delayed and the resource is abandoned at a later point in time. The intuition for this result is that it is optimal to first allocate capital to the resource technology, such that the resource-capital ratio is constant. If there is little capital available, all of it will be allocated to the resource technology. Over time, as the resource stock decreases and savings increase the capital stock, it will eventually be optimal to adopt the alternative technology.

If the capital stock is small, the time path of resource extraction has the shape of an inverse U, first increasing and then decreasing. The increasing part is explained by the possibility of raising the capital stock by saving, which implies that a larger part of the resource extraction is deferred to the future. The total amount of resource extracted depends on the cost of resource extraction and the productivity of the two technologies, but is independent of the capital stock.

6 Industrialization of developing countries

As described in the introduction, the recent industrialization of large developing countries, such as China and India, has caused an increase in their demand for exhaustible resources, which has had a substantial impact on the world markets for exhaustible resources. In this section, I use the model to try to answer the following questions: what are the effects of the industrialization of large developing countries on one of our most important exhaustible resources; oil? How are the time path of extraction of oil and the adoption of other technologies for energy production affected?

I view the industrialization of developing countries as an increase in their level of human capital, which increases the effective labor supply. Therefore, the industrialization is modeled as a permanent increase in the labor supply of the world economy.[2] The model is calibrated to fit the world production of crude oil, which should be a good proxy for the amount of oil extracted from the ground.

[2] Modeling industrialization as an increase in the efficiency of production technologies would have a comparable effect in this model.

6.1 Calibration

Figure 4.7 displays the world production of crude oil in millions of barrels per day over the period 1970-2007. I calibrate the model to the period 1990-2000. Over that period, the increase in the world production of crude oil was 13.2 percent. Hence, the initial values of the capital stock, the resource stock, and the supply of labor are chosen such that the increase in extraction in the model between 1990 and 2000 equals 13 percent. The combination $K_{1990} = 0.5$, $Q_{1990} = 10.4$, and $L = 15$ generates an increase in the extraction rate equal to 13.3 percent between 1990 and 2000. Given this set of initial conditions, the resulting values of K_{2000} and Q_{2000} are then used to generate time paths for the economy over the period 2000 to 2100 [3] .

Figure 4.8 shows the time path of the world economy when labor supply is at its initial level L. The calibrated model predicts that the rate of extraction of oil peaks in 2000-2010 and that the alternative technology is adopted in 2040-2050. The two technologies are used simultaneously until oil is abandoned in 2050-2060. Is the prediction of the peak in extraction realistic? Numerous studies have made forecasts of the so-called "Peak Oil", the point in time when the maximum rate of world oil production is reached, and the estimates vary (see Witze (2007) for an overview). According to some forecasts, the peak has already occurred or will occur soon. For example, Energy Watch Group (2007) predicts that the peak occurred in 2006. In ASPO (2008), the Association for the Study of Peak Oil and Gas predicts that the peak occurs in 2010. On the other hand, CERA (2006), one of the most optimistic forecasts, projects that world oil production will not peak before 2030. Given the uncertainty surrounding the projections of the peak in oil production, this model's prediction of a peak in 2000-2010 seems plausible.

The calibrated model predicts that the alternative technology is adopted in 2040-2050. Is that prediction realistic? In Sims et al. (2007), the IPCC has reviewed a number of predictions of future energy supply and concludes that they give widely different views. The different scenarios are influenced by uncertainty regarding technological development, particularly in carbon capture and storage technologies, as well as uncertainty regarding energy prices. After acknowledging the uncertainties,

[3] Following from the discretization of the state space, there exist other combinations of initial conditions which generate an increase in the extraction rate of 13 percent between 1990 and 2000, and they result in very similar time paths for the economy.

the IPCC makes the following projections: hydro power can contribute to 17 percent of total electricity generation by 2030, solar power only 1 percent, and wind power about 7 percent. Compared to these projections, the calibrated model predicts that the alternative technology is adopted too late. I argue that given the uncertainty, the model's predictions fall within the range of what is reasonable.

Fisher et al. (2006) predict that under a so-called "partnership technology + CCS scenario", roughly 46 percent of total electricity in the Asia Pacific Partnership countries (USA, Australia, Japan, China, India, and Korea) will be generated by non-fossil fuels in 2050. In comparison to this forecast, the model predicts that the complete switch to the alternative technology occurs too early. However, the model prediction relies on the assumption that capital can immediately and costlessly be transferred from the resource technology to the alternative technology, which is clearly a simplification of reality. Therefore, this result should be interpreted with some caution.

6.2 Results

I use the calibrated model to analyze the effects of the permanent increase in labor supply; $L' > L$, which takes place in the decade 2000-2010. I set $L' = 19$, which corresponds to an increase in labor supply of about 30 percent. Figure 4.9 shows the time path of the world economy over the time period 2000 to 2100 when labor supply is L'. With higher labor supply, the rate of extraction of oil is higher in the first decades and still peaks in the period 2000-2010. In addition, the alternative technology is adopted a decade earlier. The two technologies are then used simultaneously until oil is abandoned, which occurs a decade earlier. The total amount of oil extracted is unaffected by the increase in labor supply.

The increase in the rate of oil extraction is caused by two factors. First, an increase in labor supply makes it possible to produce more energy from oil at any point in time. Second, an increase in labor supply increases the steady-state level of capital in the economy, which increases savings. When more capital is available, oil extraction is higher. Similarly, a higher capital stock and a higher rate of oil extraction both imply that the alternative technology is adopted at an earlier point in time.

One must bear in mind that a factor which can be expected to affect the use of oil for energy production is the emission of carbon dioxide. Attempts to curb global warming, such as introducing taxes on fossil fuel, emissions trading systems etc., will increase the costs of producing energy from oil. They can also be expected to intensify developments of alternative technologies for energy production. Both measures can affect the time path of oil extraction and the adoption of alternative technologies. For computational reasons, emissions of carbon dioxide and the effects of climate change are not included in this model. Nevertheless, the lion's share of the increase in energy production will take place in the developing countries which, at least in a medium-term time perspective, are exempt from introducing measures to reduce emissions of carbon dioxide. Therefore, I argue that the model's predictions for the coming decades are still qualitatively reasonable.

A weakness of this model, which it shares with many other models featuring both exhaustible resources and alternatives, is that it does not account for the exploration of new reserves. The entire stock of resource is known at the beginning of time and the optimal time path of extraction is determined on basis of that stock. However, in the case of oil, there is a constant exploration for new oil reserves. In terms of this model, any change in reserves that was unanticipated in the 1990's would imply that optimal decisions are revised during the time period of simulation.

6.3 Robustness checks

As described earlier, the parameter σ in the extraction cost function is set to 0.5 in the parameterization of the model. This intermediate value was chosen due to lack of empirical estimates. As a robustness check, the model is therefore solved for a range of values: $\sigma \in [0.4, 0.6]$. The main results are unaffected by this variation in σ. A lower value of σ increases threshold $\hat{Q}$ and shifts the time path of extraction downwards. When extraction becomes more costly, it is optimal to leave more of the resource in the ground. Conversely, a higher value of σ reduces threshold $\hat{Q}$ and and shifts the time path of extraction upwards. In addition, I have performed robustness checks with regard to the productivity of the alternative technology, for $A \in [1.25, 1.45]$. A lower productivity implies that the alternative technology is adopted later, oil as an energy source is abandoned later and at a lower level of the remaining stock. Conversely, a higher productivity of the alternative technology implies that it is adopted earlier, oil is abandoned earlier and at a higher level of the remaining stock.

In the original model specification, capital and labor are assumed to be used in the same proportions in the two technologies. Consequently, the coefficient on capital is $\theta\alpha_1$ in the resource technology and θ in the alternative technology, respectively. $\alpha_1 < 1$ implies a higher coefficient on capital in the alternative technology. To investigate whether that difference could affect the results, I calibrate the model using the following functional form for the alternative technology; $G(K, L) = AK_A^{\theta_A} L_A^{1-\theta_A}$, where $\theta_A = \theta\alpha_1$. I find that the main results of the model are robust to this change.

7 Concluding comments

This paper tries to answer the following questions: given that there exist two different technologies for production in an economy, one which uses an exhaustible resource and an alternative which does not, what will be the optimal path of resource extraction over time? At what point in time will the alternative technology be adopted? How are the time path of extraction and the adoption of the alternative technology affected by the industrialization of developing countries? To this end, the paper constructs a dynamic model of the world economy that has two distinct technologies for production; one which uses an exhaustible resource and an alternative technology which does not. Both technologies use capital and labor. In each time period, a social planner decides how much of the resource to extract, which technology or technologies to use, how to allocate capital and labor, and how much to save. The model is first solved in a two-period setting and thereafter extended to an infinite time horizon.

The main finding is that the technology choice depends on the relative sizes of the capital stock and the resource stock. If the capital stock is large in relation to the resource stock, the alternative technology is immediately adopted. The two technologies coexist until the resource is abandoned, and there is a complete switch to the alternative technology. If, instead, the capital stock is small in relation to the resource stock, only the resource technology is used initially and the alternative technology is adopted with a delay. The resource is abandoned at a later point in time. The intuition for this result is that if the resource technology is used, it will be optimal to first allocate capital and labor to the resource technology, such that the ratio of resource input over capital input is constant, and then allocate the remaining capital to the alternative technology. Hence, it is only if the initial capital stock is sufficiently large in relation to the resource stock that the alternative technology will be immediately adopted. The lower is the capital stock, the longer is the delay before the alternative technology is adopted.

Similarly, the time path of resource extraction depends on the relative size of the capital and the resource stocks. If the capital stock is large in relation to the resource stock, resource extraction is decreasing over time. If, instead, the capital stock is small, resource extraction has the shape of an inverse U. With a low initial capital stock, it is optimal to defer a larger part of the extraction to the future, when the capital stock has increased through savings and the amount of capital available to allocate to the resource technology is higher. When the capital stock has become sufficiently high, resource extraction starts to decrease, as the motive for deferring extraction disappears.

The model is then used to analyze the effects of industrialization of developing countries on one of our most important exhaustible resources: oil. The model is calibrated to match the production of crude oil during the period 1990-2000. The industrialization is modeled as a permanent increase in the labor supply of the world economy. The calibrated model predicts that industrialization of developing countries has the following effects: the rate of oil extraction increases immediately, the alternative technology is adopted earlier, and oil as an energy source is abandoned earlier.

The model is solved for the centralized equilibrium. An avenue for future research is therefore to extend the model to a world with two countries, a resource-exporting country and a resource-importing country, and analyze how that affects technology choice and extraction paths. In addition, it would be possible to analyze how industrialization of developing countries affects resource-exporters and resource-importers, respectively.

Another feature of the model is that the productivity of the alternative technology is exogenous. It is reasonable to believe that the productivity can be increased by investment in R&D and that this investment, in turn, depends on the prices of exhaustible resources. The productivity of the alternative technology naturally influences the timing of its adoption. Hence, another avenue for future research is to allow for investment in R&D to increase the productivity of the alternative technology and to analyze how R&D investment affects technology choice.

Bibliography

[1] Association for the Study of Peak Oil and Gas, ASPO, (2008), Newsletter No. 85.

[2] Australian Bureau of Agricultural and Resource Economics, ABARE, (2008), Australian Commodities, Vol. 15, No. 1.

[3] Cambridge Energy Research Associates, CERA, (2006), "Why the "Peak Oil" Theory Falls Down – Myths, Legends, and the Future of Oil Resources".

[4] Dasgupta, P. and Heal, G. (1974), "The Optimal Depletion of Exhaustible Resources", *The Review of Economic Studies*, Vol. 41, pp. 3-28.

[5] Dasgupta, P. and Stiglitz, J. (1981), "Resource Depletion Under Technological Uncertainty", *Econometrica*, Vol. 49, No. 1, pp. 85-104.

[6] Energy Watch Group (2007), "Crude Oil Supply Outlook", EWG-Series No 3/2007.

[7] Fisher, B. Ford, M. Jakeman, G. Gurney, A. Penm, J. Matysek, A. and Gunasekera, D. (2006), "Technological development and economic growth", ABARE Research Report 06.1.

[8] Halvorsen, R. and Smith, T. (1991), "A Test of the Theory of Exhaustible Resources", *The Quarterly Journal of Economics*, Vol. 106, No. 1, pp. 123-140.

[9] Kamien, M. and Schwartz, N. (1978), "Optimal Exhaustible Resource Depletion with Endogenous Technical Change", *The Review of Economic Studies*, Vol. 45, No. 1, pp. 179-196.

[10] Lin, C. (2008), "An Empirical Dynamic Model of OPEC and Non-OPEC", mimeo.

[11] Nordhaus, W. (1973), "The Allocation of Energy Resources", *Brookings Papers on Economic Activity*, Vol. 1973, No. 3, pp. 529-576.

[12] Nordhaus W. (1992), "Lethal Model 2: The Limits to Growth Revisited", *Brookings Papers on Economic Activity*, Vol. 1992, No. 2, pp. 1-59.

[13] Nordhaus W. and Yang, Z. (1996), "A Regional Dynamic General-Equilibrium Model of Alternative Climate-Change Strategies", *The American Economic Review*, Vol. 86, No. 4, pp. 741-765.

[14] Pesaran, M. (1990) "An Econometric Analysis of Exploration and Extraction of Oil in the U.K. Continental Shelf", *The Economic Journal*, Vol. 100, No. 401, pp. 367-390.

[15] R.E.H. Sims, R.N. Schock, A. Adegbululgbe, J. Fenhann, I. Konstantinaviciute, W. Moomaw, H.B. Nimir, B. Schlamadinger, J. Torres-Martínez, C. Turner, Y. Uchiyama, S.J.V. Vuori, N. Wamukonya, and X. Zhang, (2007), Energy supply. In Climate Change 2007: Mitigation. Contribution of Working Group III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change [B. Metz, O.R. Davidson, P.R. Bosch, R. Dave, L.A. Meyer (eds)], Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

[16] Tahvonen, O. and Salo, S. (2001), "Economic growth and transitions between renewable and nonrenewable energy resources", *European Economic Review*, Vol. 45, pp. 1379-1398.

[17] Tsur, Y. and Zemel, A. (2005), "Scarcity, growth and R&D", *Journal of Environmental Economics and Management*, Vol. 49, pp. 484-499.

[18] Witze, A. (2007), "That's oil, folks...", *Nature*, Vol. 445.

[19] Young, D. (1992), "Cost Specification and Firm Behaviour in a Hotelling Model of Resource Extraction", *The Canadian Journal of Economics*, Vol. 25, No. 1, pp. 41-59.

Appendix

A1 Constant returns to scale

Suppose that the initial amount of capital in the economy is multiplied by $N > 0$ and there are N identical deposits of the exhaustible resource, all with initial stock Q_1. Total extraction from all deposits in period t is $\bar{M}_t$, for $t \in \{1, 2\}$. Given that all deposits are identical, the extraction from each deposit $n \in N$ will be $M_{t,n} = \frac{\bar{M}_t}{N}$. The total flow cost of extraction can then be written as $R(M, Q) = NC\left(Q_t^{1-\sigma} - \left(Q_t - \frac{\bar{M}_t}{N}\right)^{1-\sigma}\right)$. The stock of resource Q_t is now the stock per deposit at time t, while the total resource stock in the economy is NQ_t. Therefore, the law of motion for the stock of resource per deposit is $Q_{t+1} = Q_t - \frac{\bar{M}_t}{N}$. In an interior solution in period 2, $\bar{M}_2^* \in (0, NQ_2)$ and $K_{A,2}^* \in (0, K_2)$ satisfy the following system of equations

$$
\begin{aligned}
\alpha_2 B(K_2 - K_{A,2})^{\alpha_1} \bar{M}_2^{\alpha_2 - 1} &= (1-\sigma)C\left(Q_2 - \frac{\bar{M}_2}{N}\right)^{-\sigma} \\
A &= \alpha_1 B(K_2 - K_{A,2})^{\alpha_1 - 1} \bar{M}_2^{\alpha_2}.
\end{aligned}
$$

Given $\alpha_1 + \alpha_2 = 1$, inserting the second equation into the first yields

$$
\frac{\alpha_2 B^{\frac{1}{1-\alpha_1}} \alpha_1^{\frac{\alpha_1}{1-\alpha_1}}}{A^{\frac{\alpha_1}{1-\alpha_1}}} = (1-\sigma)C\left(Q_2 - \frac{\bar{M}_2}{N}\right)^{-\sigma}.
$$

Since $M_{2,n} = \frac{\bar{M}_2}{N}$, this condition is identical to the corresponding condition given by (4.11) and hence, the allocation is the same. In an interior solution in period 1, $\bar{M}_1^* \in (0, NQ_1)$ and $K_{A,1}^* \in (0, K_1)$ satisfy the following system of equations

$$
\begin{aligned}
& \alpha_2 B(NK_1 - K_{A,1})^{\alpha_1} \bar{M}_1^{\alpha_2 - 1} - (1-\sigma)C\left(Q_1 - \frac{\bar{M}_1}{N}\right)^{-\sigma} \\
= \;& \beta C(1-\sigma)\left(\left(Q_1 - \frac{\bar{M}_1}{N} - \frac{\bar{M}_2}{N}\right)^{-\sigma} - \left(Q_1 - \frac{\bar{M}_1}{N}\right)^{-\sigma}\right) \\
& A = \alpha_1 B(NK_1 - K_{A,1})^{\alpha_1 - 1} \bar{M}_1^{\alpha_2}.
\end{aligned}
$$

Given $\alpha_1 + \alpha_2 = 1$, inserting the second equation into the first yields

$$\frac{\alpha_2 B^{\frac{1}{1-\alpha_1}} \alpha_1^{\frac{\alpha_1}{1-\alpha_1}}}{A^{\frac{\alpha_1}{1-\alpha_1}}} - (1-\sigma)C\left(Q_1 - \frac{\bar{M}_1}{N}\right)^{-\sigma} = \beta C(1-\sigma)\left(\begin{array}{c}\left(Q_1 - \frac{\bar{M}_1}{N} - \frac{\bar{M}_2}{N}\right)^{-\sigma} \\ -\left(Q_1 - \frac{\bar{M}_1}{N}\right)^{-\sigma}\end{array}\right).$$

Since $M_{1,n} = \frac{\bar{M}_1}{N}$, this condition is identical to the corresponding condition given by (4.13) and hence, the allocation is the same.

A2 Proofs

Proof of Proposition 1

Suppose that extraction is $M = \epsilon$ where $\epsilon \to 0$. Inserting equation 2 in (4.13) into the expression for consumption gives

$$c = A\left(K - \left(\frac{A}{\alpha_1 B \epsilon^{\alpha_2}}\right)^{\frac{1}{\alpha_1 - 1}}\right) + B\left(\frac{A}{\alpha_1 B \epsilon^{\alpha_2}}\right)^{\frac{\alpha_1}{\alpha_1 - 1}} \epsilon^{\alpha_2} - R(\epsilon, Q).$$

By defining

$$\begin{aligned} D_0 &= AK \\ D_1 &= -A\left(\frac{A}{\alpha_1 B}\right)^{\frac{1}{\alpha_1 - 1}} + B\left(\frac{A}{\alpha_1 B}\right)^{\frac{\alpha_1}{\alpha_1 - 1}} \end{aligned}$$

consumption can be written as $c = D_0 + D_1 \epsilon^{\frac{\alpha_2}{1-\alpha_1}} - R(\epsilon, Q)$. The derivative of this expression w.r.t ϵ is

$$\frac{\partial c}{\partial \epsilon} = \frac{\alpha_2}{1-\alpha_1} D_1 \epsilon^{\frac{\alpha_2}{1-\alpha_1} - 1} - \frac{\partial R(\epsilon, Q)}{\partial \epsilon}$$

Now, if $\alpha_1 + \alpha_2 = 1$, then

$$\frac{\partial c}{\partial \epsilon} = \frac{\alpha_2}{1-\alpha_1} D_1 \epsilon^0 - \frac{\partial R(\epsilon, Q)}{\partial \epsilon}$$

If $\lim_{\epsilon \to 0} \frac{\partial R(\epsilon,Q)}{\partial \epsilon} > \frac{\alpha_2}{1-\alpha_1} D_1$, it follows that $\frac{\partial c}{\partial \epsilon} < 0$, and it is optimal to set $M^* = 0$. $\lim_{\epsilon \to 0} \frac{\partial R(\epsilon,Q)}{\partial \epsilon} = \frac{C(1-\sigma)}{Q^\sigma}$. Hence, there exists a $\hat{Q} > 0$, such that $\frac{C(1-\sigma)}{\hat{Q}^\sigma} > \frac{\alpha_2}{1-\alpha_1} D_1$. An interior solution to the optimization problem in period 1 satisfies the following system of equations

$$\begin{aligned} & \alpha_2 B(K_1 - K_{A,1})^{\alpha_1} M_1^{\alpha_2 - 1} - (1-\sigma) C (Q_1 - M_1)^{-\sigma} \\ = \; & \beta C (1-\sigma) \left((Q_1 - M_1 - M_2)^{-\sigma} - (Q_1 - M_1)^{-\sigma} \right) \\ & A = \alpha_1 B (K_1 - K_{A,1})^{\alpha_2 - 1} M_1^{\alpha_2}. \end{aligned}$$

Solving for $K_1 - K_{A,1}$ from equation 2 above and inserting it in equation 1 gives

$$\begin{aligned} \frac{\alpha_2 B^{\frac{1}{1-\alpha_1}} \alpha_1^{\frac{\alpha_1}{1-\alpha_1}} M^{\frac{\alpha_1+\alpha_2-1}{1-\alpha_1}}}{A^{\frac{\alpha_1}{1-\alpha_1}}} \; &= \; (1-\sigma) C (Q_1 - M_1)^{-\sigma} + \beta C (1-\sigma)(Q_1 - M_1 - M_2)^{-\sigma} \\ & \quad -\beta C (1-\sigma)(Q_1 - M_1)^{-\sigma}. \end{aligned}$$

Now, if $\alpha_1 + \alpha_2 = 1$ and $\beta = 1$, the expression simplifies to

$$\frac{\alpha_2 B^{\frac{1}{1-\alpha_1}} \alpha_1^{\frac{\alpha_1}{1-\alpha_1}}}{A^{\frac{\alpha_1}{1-\alpha_1}}} = C(1-\sigma)(Q_1 - M_1 - M_2)^{-\sigma}.$$

Solving for $M_1^* + M_2^*$ yields

$$M_1^* + M_2^* = Q_1 - \frac{A^{\frac{\alpha_1}{\sigma(1-\alpha_1)}} C^{\frac{1}{\sigma}} (1-\sigma)^{\frac{1}{\sigma}}}{\alpha_2^{\frac{1}{\sigma}} B^{\frac{1}{\sigma(1-\alpha_1)}} \alpha_1^{\frac{\alpha_1}{\sigma(1-\alpha_1)}}}.$$

$M_1^* + M_2^* = 0$ if $Q_1 = \hat{Q}$, where

$$\hat{Q} = \frac{A^{\frac{\alpha_1}{\sigma(1-\alpha_1)}} C^{\frac{1}{\sigma}} (1-\sigma)^{\frac{1}{\sigma}}}{\alpha_2^{\frac{1}{\sigma}} B^{\frac{1}{\sigma(1-\alpha_1)}} \alpha_1^{\frac{\alpha_1}{\sigma(1-\alpha_1)}}}.$$

Proof of Proposition 2

If $K_2 > \hat{K}(Q_2)$, it follows that $K_{A,2}^* > 0$ and $K_{A,1}^* = 0$ or $K_{A,1}^* > 0$. As long as $K_{A,1}^* > 0$, it follows from Proposition 1 that $M_1^* + M_2^* = Q_1 - \hat{Q}$. Now, suppose that $K_{A,1}^* = 0$. M_2^* is given by

$$M_2^* = Q_2 - \frac{A^{\frac{\alpha_1}{\sigma(1-\alpha_1)}} C^{\frac{1}{\sigma}} (1-\sigma)^{\frac{1}{\sigma}}}{\alpha_2^{\frac{1}{\sigma}} B^{\frac{1}{\sigma(1-\alpha_1)}} \alpha_1^{\frac{\alpha_1}{\sigma(1-\alpha_1)}}}$$

but by definition, $Q_2 = Q_1 - M_1^*$. Consequently, $M_1^* + M_2^*$ is given by

$$M_1^* + M_2^* = Q_1 - \frac{A^{\frac{\alpha_1}{\sigma(1-\alpha_1)}} C^{\frac{1}{\sigma}} (1-\sigma)^{\frac{1}{\sigma}}}{\alpha_2^{\frac{1}{\sigma}} B^{\frac{1}{\sigma(1-\alpha_1)}} \alpha_1^{\frac{\alpha_1}{\sigma(1-\alpha_1)}}}.$$

Proof of Proposition 3

If $K_1 > \hat{K}(Q_1)$, it follows that $K_{A,1}^* > 0$. Hence, $M_1^* + M_2^* = M^T$. If $Q_1 > \hat{Q}$ it follows that $M^T > 0$. Given that $Q_2 \leq Q_1$, it must be the case that $K_{A,2}^* > 0$. Since the optimal solution only determines M^T and not the division into M_1^* and M_2^*, M_1^* can take any value in $[0, M^T]$ and the utility of the consumer is maximized. Then, $M_2^* = M^T - M_1^*$.

Proof of Proposition 4

The condition $Q_1 > \hat{Q}$ implies that $M_1^* + M_2^* > 0$. The condition $K_2 > \hat{K}(Q_2)$ implies that $M_1^* + M_2^* = M^T$. The optimal amount of saving given $M_1 = \tilde{M}$ is: $K_2 = \frac{AK_1}{\alpha_1} - C\left(Q_1^{1-\sigma} - \left(Q_1 - \tilde{M}\right)^{1-\sigma}\right)$, and next period's resource stock is $Q_2 = Q_1 - \tilde{M}$. For $M_1 = \tilde{M}$, $K_{A,1}^* = 0$ and since $K_2 > \hat{K}(Q_2)$, we have that $K_{A,2}^* > 0$. For $M_1 < \tilde{M}$, $K_{A,1}^* > 0$. If the resulting K_2 and Q_2 are such that $K_2 < \hat{K}(Q_2)$, then $K_{A,2}^* = 0$. Otherwise, $K_{A,2}^* > 0$. Hence, the allocations are either $K_{A,1}^* = 0$, $K_{A,2}^* > 0$ or $K_{A,1}^* > 0$, $K_{A,2}^* = 0$ or $K_{A,1}^* > 0$, $K_{A,2}^* > 0$. $K_1 \leq \hat{K}(Q_1)$ is equivalent to $\tilde{M} < M^T$. This implies that $M_2^* > 0$ for all M_1, since $M_2^* = M^T - M_1^*$. Is $M_1^* > 0$? Suppose that $M_1^* = 0$, then $Q_2 = Q_1$ and, given $A = 1$, $K_2 = K_1$. But then $K_2 < \hat{K}(Q_2)$. Consequently, $M_1 = 0$ implies that $K_{A,2} = 0$ and then utility is not maximized. Therefore, $M_1^* > 0$ and $M_2^* > 0$.

Proof of Proposition 5

The condition $Q_1 > \hat{Q}$ implies that $M_1^* + M_2^* > 0$. Since $K_2 \leq \hat{K}(Q_2)$, Proposition 2 does not hold, and we have that $M_1^* + M_2^* < M^T$. Therefore, it follows from the first-order conditions that it cannot be optimal to set $K_{A,1}^* > 0$. Hence, $M_1^* > 0$ and $K_{A,1}^* = 0$. Similarly, it cannot be optimal to allocate capital to the alternative technology in period 2, and $M_2^* > 0$ and $K_{A,2}^* = 0$.

A3 Figures

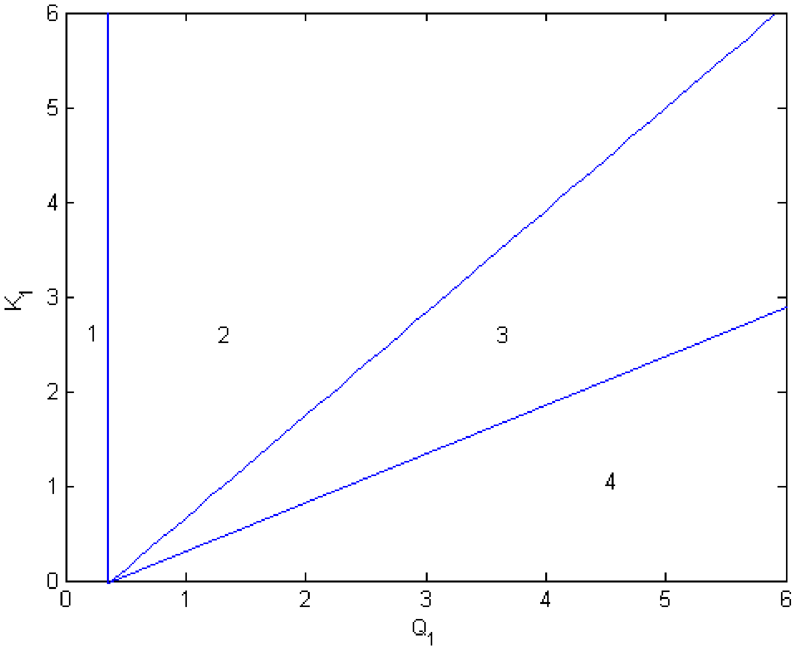

Figure 4.1: Technology choice in the baseline model. Area 1: case 1, area 2: case 2, area 3: case 3, and area 4: case 4.

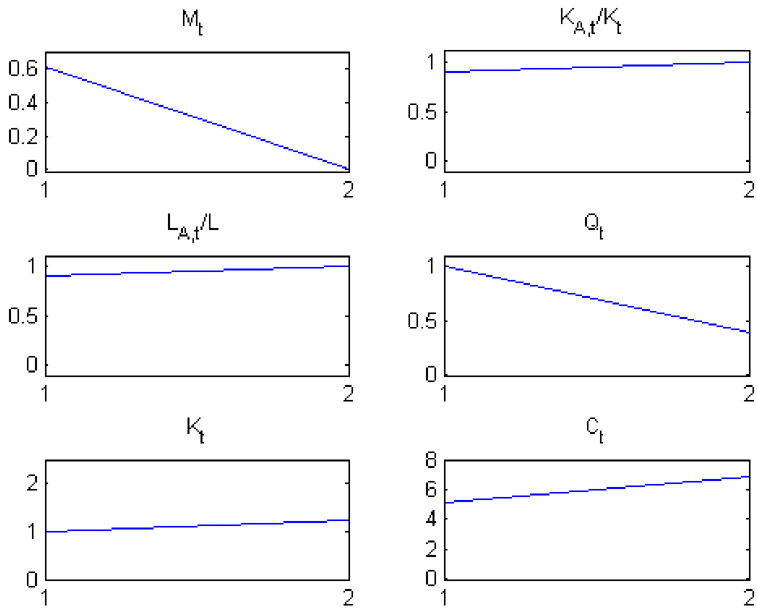

Figure 4.2: Example of the immediate-adoption case. M_t : resource extraction, $K_{A,t}/K_t$: share of capital stock allocated to alternative technology, $L_{A,t}/L_t$: share of labor force allocated to alternative technology, Q_t : resource stock, K_t : capital stock, C_t : consumption.

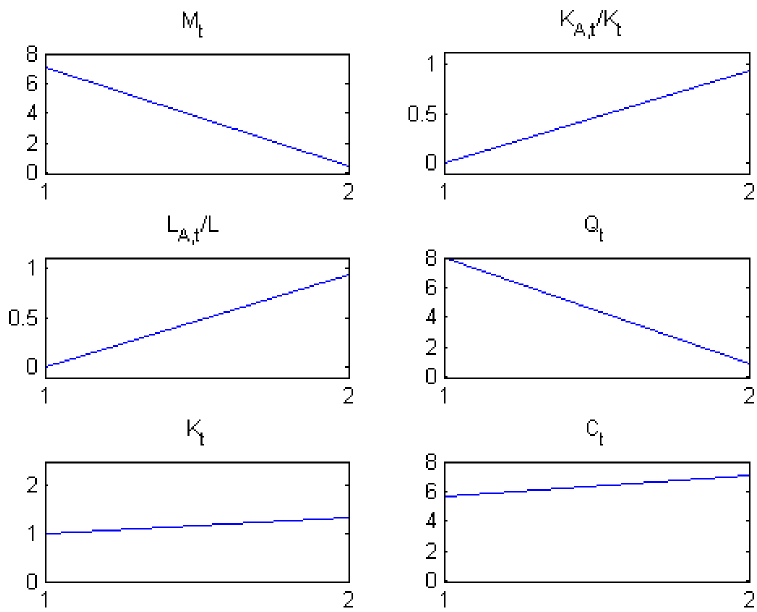

Figure 4.3: Example of the delayed-adoption case.

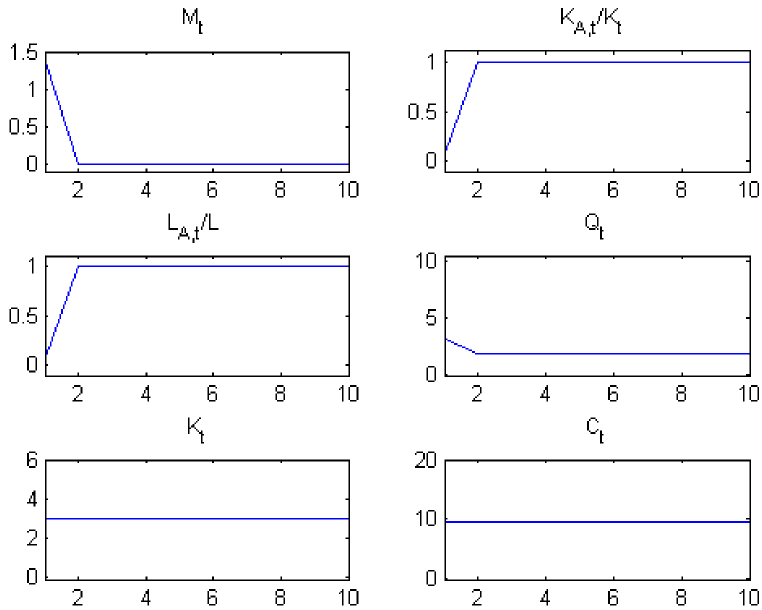

Figure 4.4: Time path for initial conditions $K_0 = 3$ and $Q_0 = 3.2$.

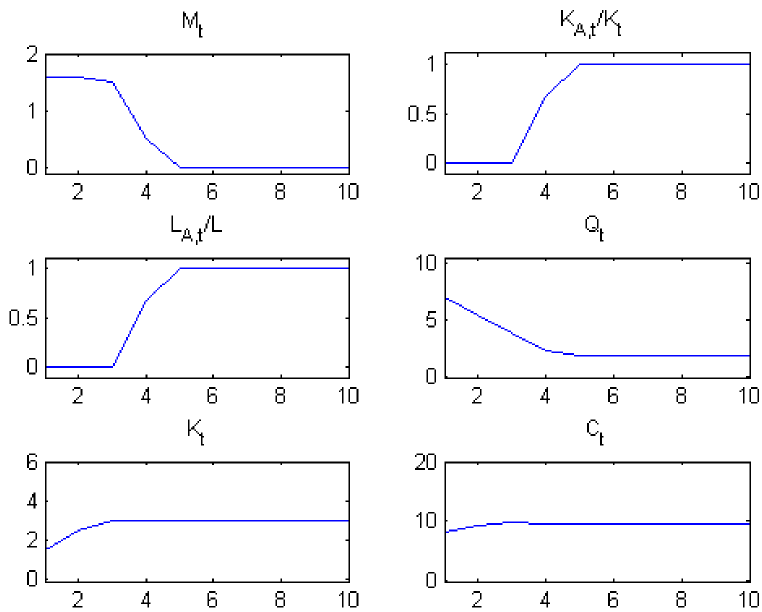

Figure 4.5: Time path for initial conditions $K_0 = 1.5$ and $Q_0 = 7$.

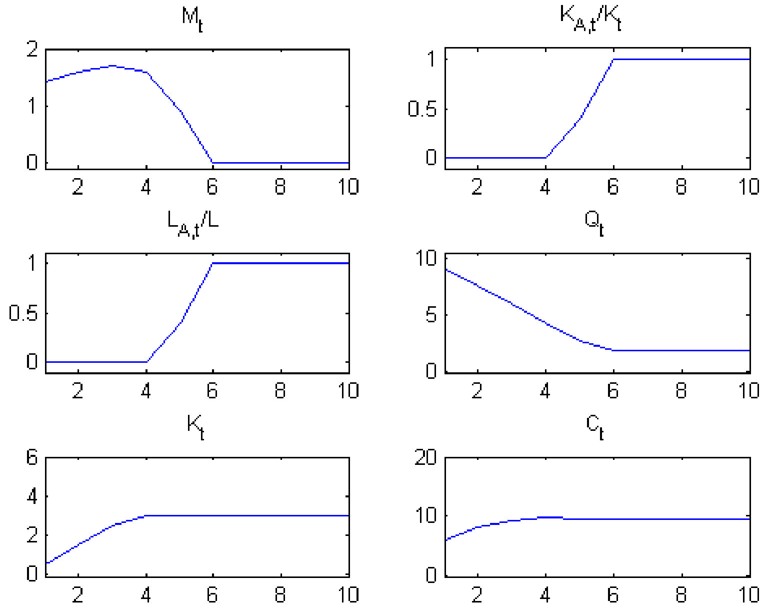

Figure 4.6: Time path for initial conditions $K_0 = 0.5$ and $Q_0 = 9$.

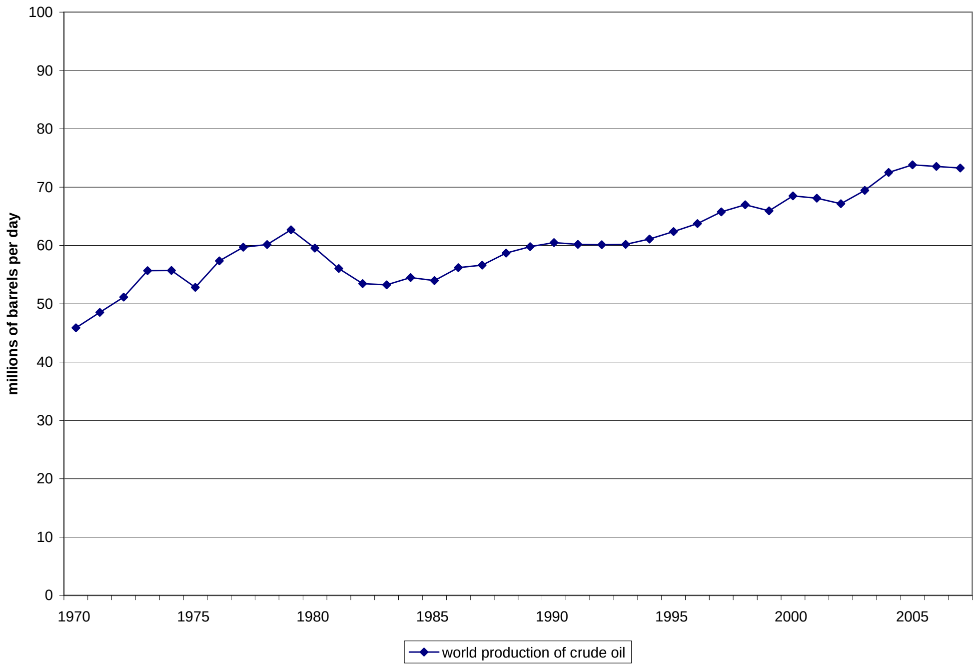

Figure 4.7: World production of crude oil 1970-2007. Source: International Petroleum Monthly, April 2008, Energy Information Administration.

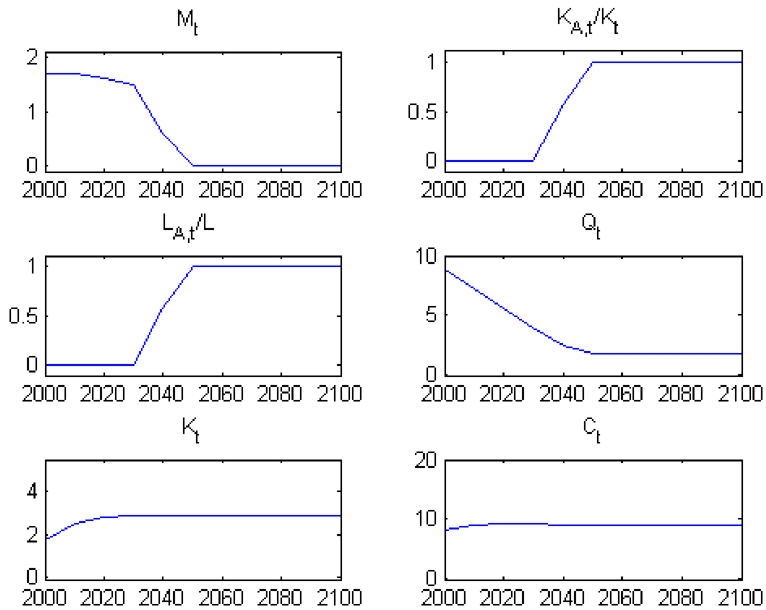

Figure 4.8: Time path of the world economy when the labor supply is L.

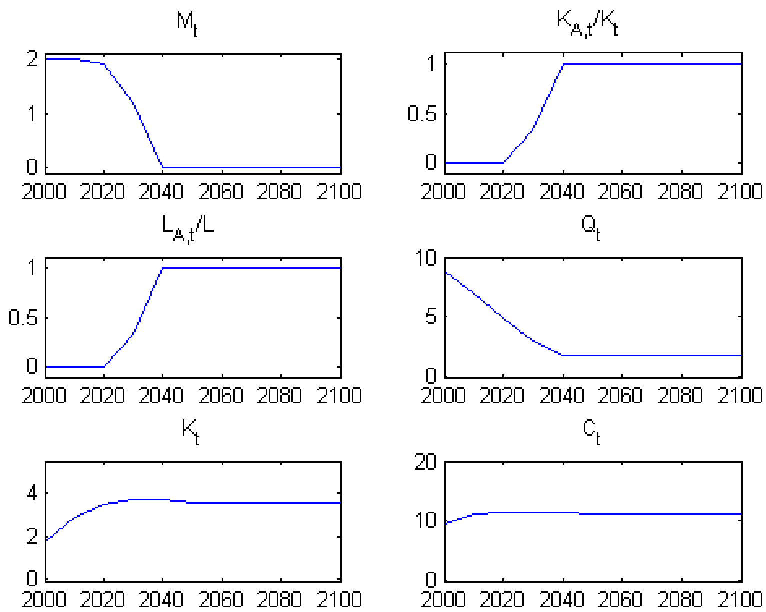

Figure 4.9: Time path of the world economy when the labor supply is L'.

IIES Monograph Series

1. Michaely, Michael *The Theory of Commercial Policy: Trade and Protection*, 1973
2. Söderström, Hans Tson *Studies in the Microdynamics of Production and Productivity Change*, 1974
3. Hamilton, Carl B. *Project Analysis in the Rural Sector with Special Reference to the Evaluation of Labour Cost*, 1974
4. Nyberg, Lars and Staffan Viotti *A Control Systems Approach to Macroeconomic Theory and Policy in the Open Economy*, 1975
5. Myhrman, Johan *Monetary Policy in Open Economies*, 1975
6. Krauss, Melvyn *International Trade and Economic Welfare*, 1975
7. Wihlborg, Clas *Capital Market Integration and Monetary Policy under Different Exchange Rate Regimes*, 1976
8. Svensson, Lars E.O. *On Competitive Markets and Intertemporal Resources Allocation*, 1976
9. Yeats, Alexander J. *Trade Barriers Facing Developing Countries*, 1978
10. Calmfors, Lars *Prices, Wages and Employment in the Open Economy*, 1978
11. Kornai, János *Economics of Shortage*, Vols I and II, 1979
12. Flam, Harry *Growth, Allocation and Trade in Sweden. An Empirical Application of the Heckscher-Ohlin Theory*, 1981
13. Persson, Torsten *Studies of Alternative Exchange Rate Systems. An Intertemporal General Equilibrium Approach*, 1982
14. Erzan, Refik *Turkey's Comparative Advantage, Production and Trade Patterns in Manufactures. An Application of the Factor Proportions Hypothesis with Some Qualifications*, 1983
15. Horn af Rantzien, Henrik *Imperfect Competition in Models of Wage Formation and International Trade*, 1983
16. Nandakumar, Parameswar *Macroeconomic Effects of Supply Side Policies and Disturbances in Open Economies*, 1985
17. Sellin, Peter *Asset Pricing and Portfolio Choice with International Investment Barriers*, 1990
18. Werner, Ingrid *International Capital Markets: Controls, Taxes and Resources Allocation*, 1990
19. Svedberg, Peter *Poverty and Undernutrition in Sub-Saharan Africa: Theory, Evidence, Policy*, 1991
20. Nordström, Håkan *Studies in Trade Policy and Economic Growth*, 1992
21. Hassler, John, Lundvik, Petter, Persson, Torsten and Söderlind, Paul *The Swedish Business Cycle: Stylized facts over 130 years*, 1992

22. Lundvik, Petter *Business Cycles and Growth*, 1992
23. Söderlind, Paul *Essays in Exchange Rates, Business Cycles and Growth*, 1993
24. Hassler, John A.A. *Effects of Variations in Risk on Demand and Measures of Business Cycle Comovements*, 1994
25. Daltung, Sonja *Risk, Efficiency, and Regulation of Banks*, 1994
26. Lindberg, Hans *Exchange Rates: Target Zones, Interventions and Regime Collapses*, 1994
27. Stennek, Johan *Essays on Information-Processing and Competition*, 1994
28. Jonsson, Gunnar *Institutions and Incentives in Monetary and Fiscal Policy*, 1995
29. Dahlquist, Magnus *Essays on the Term Structure of Interest Rates and Monetary Policy*, 1995
30. Svensson, Jakob *Political Economy and Macroeconomics: On Foreign Aid and Development*, 1996
31. Blix, Mårten *Rational Expectations and Regime Shifts in Macroeconometrics*, 1997
32. Lagerlöf, Nils-Petter *Intergenerational Transfers and Altruism*, 1997
33. Klein, Paul *Papers on the Macroeconomics of Fiscal Policy*, 1997
34. Jonsson, Magnus *Studies in Business Cycles*, 1997
35. Persson, Lars *Asset Ownership in Imperfectly Competitive Markets*, 1998
36. Persson, Joakim *Essays on Economic Growth*, 1998
37. Domeij, David *Essays on Optimal Taxation and Indeterminacy*, 1998
38. Flodén, Martin *Essays on Dynamic Macroeconomics*, 1999
39. Tangerås, Thomas *Essays in Economics and Politics: Regulation, Elections and International Conflict*, 2000
40. Lidbom, Per Pettersson *Elections, Party Politics and Economic Policy*, 2000
41. Vestin, David *Essays on Monetary Policy*, 2001
42. Olofsgård, Anders *Essays on Interregional and International Political Economics*, 2001
43. Johansson, Åsa *Essays on Macroeconomic Fluctuations and Nominal Wage Rigidity*, 2002
44. Groth, Charlotta *Topics on Monetary Policy*, 2002
45. Gancia, Gino A. *Essays on Growth, Trade and Inequality*, 2003
46. Harstad, Bård *Organizing Cooperation: Bargaining, Voting and Control*, 2003
47. Kohlscheen, Emanuel *Essays on Debts and Constitutions*, 2004
48. Olovsson, Conny *Essays on Dynamic Macroeconomics*, 2004
49. Stavlöt, Ulrika *Essays on Culture and Trade*, 2005
50. Herzing, Mathias *Essays on Uncertainty and Escape in Trade Agreements*, 2005

51. Bonfiglioli, Alessandra *Essays on Financial Markets and Macroeconomics*, 2005
52. Pienaar, Natalie *Economic Applications of Product Quality Regulation in WTO Trade Agreements*, 2005
53. Song, Zheng *Essays on Dynamic Political Economy*, 2005
54. Eisensee, Thomas *Essays on Public Finance: Retirement Behavior and Disaster Relief*, 2005
55. Favara, Giovanni *Credit and Finance in the Macroeconomy*, 2006
56. Björkman, Martina *Essays on Empirical Development Economics: Education, Health and Gender*, 2006
57. Larsson, Anna *Real Effects of Monetary Regimes*, 2007
58. Prado, Jr., Jose Mauricio *Essays on Public Macroeconomic Policy*, 2007
59. Tonin, Mirco *Essays on Labor Market Structures and Policies*, 2007
60. Queijo von Heideken, Virgina *Essays on Monetary Policy and Asset Markets*, 2007
61. Finocchiaro, Daria *Essays on Macroeconomics*, 2007
62. Waisman, Gisela *Essays on Discrimination and Corruption*, 2008
63. Holte, Martin Bech *Essays on Incentives and Leadership*, 2008
64. Färnstrand Damsgaard, Erika *Essays on Technology Choice and Spillovers*, 2008