Lecture Notes in Macroeconomics

Christian Groth

August 30, 2015

Contents

Chapter 1

Introduction

> The art of successful theorizing is to make the inevitable simplifying assumptions in such a way that the final results are not very sensitive.
>
> −Robert M. Solow (1956, p. 65)

1.1 Macroeconomics

1.1.1 The field

Economics is the social science that studies the production and distribution of goods and services in society. Then, what defines the branch of economics named *macroeconomics?* There are two defining characteristics. First, *macroeconomics* is the systematic study of the economic interactions in society as a whole. This could also be said of *microeconomic* general equilibrium theory, however. The second defining characteristic of macroeconomics is that it aims at understanding the empirical regularities in the behavior of aggregate economic variables such as aggregate production, investment, unemployment, the general price level for goods and services, the inflation rate, the level of interest rates, the level of real wages, the foreign exchange rate, productivity growth etc. Thus macroeconomics focuses on the major lines of the economics of a society.

The aspiration of macroeconomics is three-fold:

1. to *explain* the levels of the aggregate variables as well as their movement over time in the short run and the long run;
2. to make well-founded *forecasts* possible;
3. to provide foundations for rational *economic policy* applicable to macroeconomic problems, be they short-run distress in the form of economic recession or problems of a more long-term, structural character.

We use *economic models* to make our complex economic environment accessible for theoretical analysis. What is an economic model? It is a way of organizing one's thoughts about the economic functioning of a society. A more specific answer is to define an economic model as a conceptual structure based on a set of mathematically formulated assumptions which have an economic interpretation and from which empirically testable predictions can be derived. In particular, a macroeconomic model is an economic model concerned with macroeconomic phenomena, i.e., the short-run fluctuations of aggregate variables as well as their long-run trend.

Any economic analysis is based upon a conceptual framework. Formulating this framework as a precisely stated economic model helps to break down the issue into assumptions about the concerns and constraints of households and firms and the character of the market environment within which these agents interact. The advantage of this approach is that it makes rigorous reasoning possible, lays bare where the underlying disagreements behind different interpretations of economic phenomena are, and makes sensitivity analysis of the conclusions amenable. By being explicit about agents' concerns, the technological constraints, and the social structures (market forms, social conventions, and legal institutions) conditioning their interactions, this approach allows analysis of policy interventions, including the use of well-established tools of welfare economics. Moreover, mathematical modeling is a simple necessity to keep track of the many mutual dependencies and to provide a consistency check of the many accounting relationships involved. And mathematical modeling opens up for use of powerful mathematical theorems from the mathematical toolbox. Without these math tools it would in many cases be impossible to reach any conclusion whatsoever.

Undergraduate students of economics are often perplexed or even frustrated by macroeconomics being so preoccupied with composite theoretical models. Why not study the issues each at a time? The reason is that the issues, say housing prices and changes in unemployment, are not separate, but parts of a complex system of mutually dependent variables. This also suggests that macroeconomics must take advantage of theoretical and empirical knowledge from other branches of economics, including microeconomics, industrial organization, game theory, political economy, behavioral economics, and even sociology and psychology.

At the same time models necessarily give a *simplified* picture of the economic reality. Ignoring secondary aspects and details is indispensable to be able to focus on the essential features of a given problem. In particular macroeconomics deliberately simplifies the description of the individual actors so as to make the analysis of the interaction between different *types* of actors manageable.

The assessment of − and choice between − *competing* simplifying frameworks should be based on how well they perform in relation to the three-fold aim of

macroeconomics listed above, given the problem at hand. A necessary condition for good performance is the empirical tenability of the model's predictions. A guiding principle in the development of useful models therefore lies in confrontation of the predictions as well as the crucial assumptions with data. This can be based on a variety of methods ranging from sophisticated econometric techniques to qualitative case studies.

Three constituents make up an *economic theory:* 1) the union of connected and non-contradictory economic models, 2) the theorems derived from these, and 3) the conceptual system defining the correspondence between the variables of the models and the social reality to which they are to be applied. Being about the interaction of *human* beings in *societies*, the subject matter of economic theory is extremely complex and at the same time history dependent. The overall political, social, and economic institutions ("rules of the game" in a broad sense) evolve. These circumstances explain why economic theory is far from the natural sciences with respect to precision and undisputable empirical foundation. Especially in macroeconomics, to avoid confusion one should be aware of the existence of differing conceptions and in several matters conflicting theoretical schools.

1.1.2 The different "runs"

This textbook is about the macroeconomics of the industrialized market economies of today. We study basic concepts, models, and analytical methods of relevance for understanding macroeconomic processes where sometimes centripetal and sometimes centrifugal forces are dominating. A simplifying device is the distinction between "short-run", "medium-run", and "long-run" analysis. The first concentrates on the behavior of the macroeconomic variables within a time horizon of a few years, whereas "long-run" analysis deals with a considerably longer time horizon − indeed, long enough for changes in the capital stock, population, and technology to have a dominating influence on changes in the level of production. The "medium run" is then something in between.

To be more specific, *long-run macromodels* study the evolution of an economy's productive capacity over time. Typically a time span of at least 15 years is considered. The analytical framework is by and large *supply-dominated*. That is, variations in the employment rate for labor and capital due to demand fluctuations are abstracted away. This can to a first approximation be justified by the fact that these variations, at least in advanced economies, tend to remain within a fairly narrow band. Therefore, under "normal" circumstances the economic outcome after, say, a 30 years' interval reflects primarily the change in supply side factors such as the labor force, the capital stock, and the technology. The fluctuations in demand and monetary factors tend to be of limited quantitative

importance within such a time horizon.

By contrast, when we speak of *short-run macromodels*, we think of models concentrating on mechanisms that determine how fully an economy uses its productive capacity at a given point in time. The focus is on the level of output and employment within a time horizon less than, say, four years. These models are typically *demand-dominated*. In this time perspective the demand side, monetary factors, and price rigidities matter significantly. Shifts in aggregate demand (induced by, e.g., changes in fiscal or monetary policy, exports, interest rates, the general state of confidence, etc.) tend to be accommodated by changes in the produced quantities rather than in the prices of manufactured goods and services. By contrast, variations in the supply of production factors and technology are diminutive and of limited importance within this time span. With Keynes' words the aim of short-run analysis is to explain "what determines the actual employment of the available resources" (Keynes 1936, p. 4).

The short and the long run make up the traditional subdivision of macroeconomics. It is convenient and fruitful, however, to include also a *medium run*, referring to a time interval of, say, four-to-fifteen years.[1] We shall call models attempting to bridge the gap between the short and the long run *medium-run macromodels*. These models deal with the regularities exhibited by *sequences* of short periods. However, in contrast to long-run models which focus on the trend of the economy, medium-run models attempt to understand the pattern characterizing the fluctuations around the trend. In this context, variations at both the demand and supply side are important. Indeed, at the centre of attention is the dynamic interaction between demand and supply factors, the correction of expectations, and the time-consuming adjustment of wages and prices. Such models are also sometimes called *business cycle models*.

Returning to the "long run", what does it embrace in this book? Well, since the surge of "new growth theory" or "endogenous growth theory" in the late 1980s and early 1990s, growth theory has developed into a specialized discipline studying the factors and mechanisms that *determine* the evolution of technology and productivity (Paul Romer 1987, 1990; Phillipe Aghion and Peter Howitt, 1992). An attempt to give a systematic account of this expanding line of work within macroeconomics would take us too far. When we refer to "long-run macromodels", we just think of macromodels with a time horizon long enough such that changes in the capital stock, population, and technology matter. Apart from a taste of "new growth theory" in Chapter 11, we leave the *sources* of changes in technology out of consideration, which is tantamount to regarding these changes

[1] These number-of-years figures are only a rough indication. The different "runs" are relative concepts and their appropriateness depends on the specific problem and circumstances at hand.

as exogenous.[2]

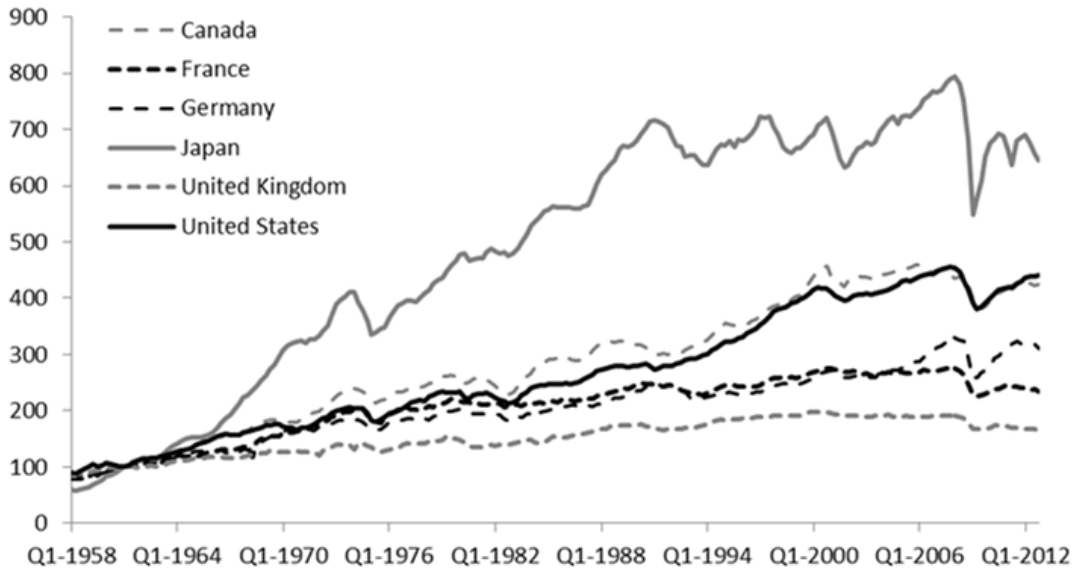

Figure 1.1: Quarterly Industrial Production Index in six major countries (Q1-1958 to Q2-2013; index Q1-1961=100). Source: OECD Industry and Service Statistics. Note: Industrial production includes manufacturing, mining and quarrying, electricity, gas, and water, and construction.

In addition to the time scale dimension, the national-international dimension is important for macroeconomics. Most industrialized economies participate in international trade of goods and financial assets. This results in considerable mutual dependency and co-movement of these economies. Downturns as well as upturns occur at about the same time, as indicated by Fig. 1.1. In particular the economic recessions triggered by the oil price shocks in 1973 and 1980 and by the disruption of credit markets in the outbreak 2007 of the Great Financial Crisis are visible across the countries, as also shown by the evolution of GDP, cf. Fig. 1.2. Many of the models and mechanisms treated in this text will therefore be considered not only in a closed economy setup, but also from the point of view of open economies.

[2]References to textbooks on economic growth are given in *Literature notes* at the end of this chapter.

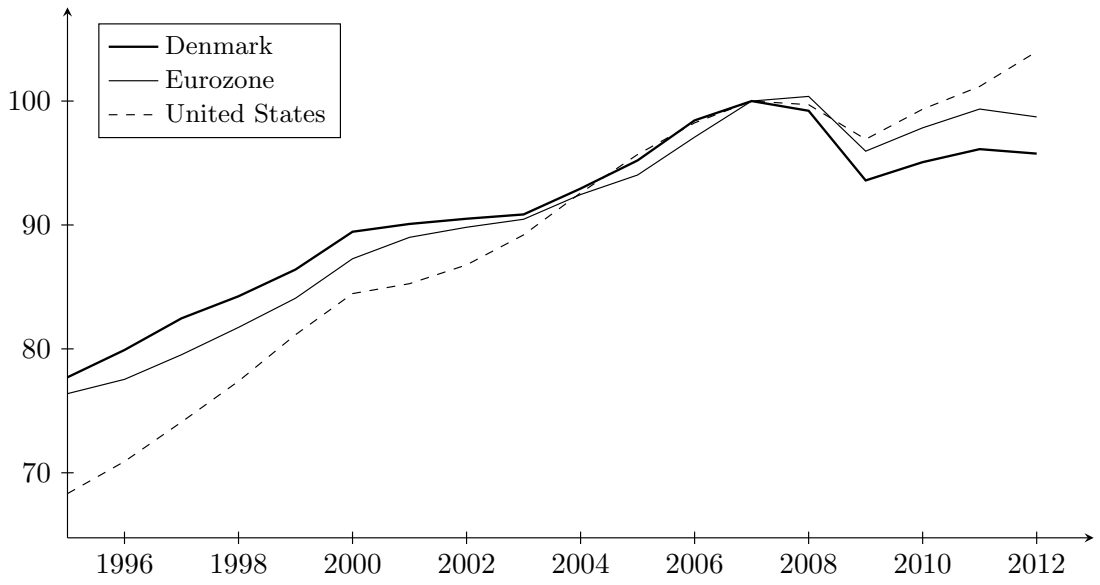

Figure 1.2: Indexed real GDP for Denmark, Eurozone and US, 1995-2012 (2007=100). Source: EcoWin and Statistics Denmark.

1.2 Components of macroeconomic models

1.2.1 Basics

(Incomplete)

Basic categories

- Agents: We use simple descriptions of the economic agents: A *household* is an abstract entity making consumption, saving and labor supply decisions. A *firm* is an abstract entity making decisions about production and sales. The administrative staff and sales personnel are treated along with the production workers as an undifferentiated labor input.
- Technological constraints.
- Goods, labor, and assets markets.
- The institutions and social norms regulating the economic interactions (formal and informal "rules of the game").

Types of variables

Endogenous vs. exogenous variables.

Stocks vs. flows.

State variables vs. control variables (decision variables). Closely related to this distinction is that between a *predetermined* variable and a *jump variable.* The former is a variable whose value is determined historically at any point in time. For example, the stock (quantity) of water in a bathtub at time t is historically determined as the accumulated quantity of water stemming from the previous inflow and outflow. But if y_t is a variable which is not tied down by its own past but, on the contrary, can immediately adjust if new conditions or new information emerge, then y_t is a non-predetermined variable, also called a jump variable. A decision about how much to consume and how much to save − or dissave − in a given month is an example of a jump variable. Returning to our bath tub example: in the moment we pull out the waste plug, the outflow of water per time unit will jump from zero to a positive value − it is a jump variable.

Types of basic model relations

Although model relations can take different forms, in macroeconomics they often have the form of equations. A taxonomy for macroeconomic model relations is the following:

1. *Technology equations* describe relations between inputs and output (production functions and similar).

2. *Preference equations* express preferences, e.g. $U = \sum_{t=0}^{T} \frac{u(c_t)}{(1+\rho)^t}$, $\rho > 0, u' > 0, u'' < 0$.

3. *Budget constraints*, whether in the form of an equation or an inequality.

4. *Institutional equations* refer to relationships required by law (e.g., how the tax levied depends on income) and similar.

5. *Behavioral equations* describe the behavioral response to the determinants of behavior. This includes an agent's optimizing behavior written as a function of its determinants. A consumption function is an example. Whether first-order conditions in optimization problems should be considered behavioral equations or just separate first-order conditions is a matter of taste.

6. *Identity equations* are true by definition of the variables involved. National income accounting equations are an example.

7. *Equilibrium equations* define the condition for equilibrium ("state of rest") of some kind, for instance equality of Walrasian demand and Walrasian supply. No-arbitrage conditions for the asset markets also belong under the heading equilibrium condition.

8. *Initial conditions* are equations fixing the initial values of the state variables in a dynamic model

Types of analysis

Statics vs. dynamics. Comparative dynamics vs. study of dynamic effects of a parameter shift in historical time.

Macroeconomics studies processes in real time. The emphasis is on *dynamic* models, that is, models that establishes a link from the state of the economic system to the subsequent state. A dynamic model thus allows a derivation of the evolution over time of the endogenous variables. A *static model* is a model where time does not enter or where all variables refer to the same point in time. Occasionally we consider static models, or more precisely *quasi-static models*. The modifier "quasi-" is meant to indicate that although the model is a framework for analysis of only a single period, the model considers some variables as inherited from *the past* and some variables that involve expectations about the future. What we call *temporary equilibrium models* are of this type. Their role is to serve as a prelude to a more elaborate dynamic model dealing with the same elements.

Dynamic analysis aims at establishing dynamic properties of an economic system: is the system stable or unstable, is it asymptotically stable, if so, is it globally or only locally asymptotically stable, is it oscillatory? If the system is asymptotically stable, how fast is the adjustment?

Partial equilibrium vs. general equilibrium:

We say that a given single market is in *partial equilibrium* at a given point in time if for arbitrarily given prices and quantities in the other markets, the agents' chosen actions in this market are mutually compatible. In contrast the concept of general equilibrium take the mutual dependencies between markets into account. We say that a given economy is in *general equilibrium* at a given point in time if in all markets the actions chosen by all the agents are mutually compatible.

An analyst trying to clarify a partial equilibrium problem is doing *partial equilibrium analysis*. Thus partial equilibrium analysis does not take into account the feedbacks from these actions to the rest of the economy and the feedbacks from these feedbacks − and so on. In contrast, an analyst trying to clarify a general equilibrium problem is doing *general equilibrium analysis*. This requires considering the mutual dependencies in the system of markets as a whole.

Sometimes even the analysis of the constrained maximization problem of a single decision maker is called partial equilibrium analysis. Consider for instance the consumption-saving decision of a household. Then the analytical derivation of the saving function of the household is by some authors included under the heading partial equilibrium analysis, which may seem natural since the real wage and real interest rate appearing as arguments in the derived saving function are

arbitrary. Indeed, what the actual saving of the young will be in the end, depends on the real wage and real interest rate formed in the general equilibrium.

In this book we call the analysis of a single decision maker's problem *partial analysis*, not partial equilibrium analysis. The motivation for this is that transparency is improved if one preserves the notion of equilibrium for a state of a *market* or a state of a *system of markets* .

1.2.2 The time dimension of input and output

In macroeconomic theory the production of a firm, a sector, or the economy as a whole is often represented by a two-inputs-one-output production function,

$$Y = F(K, L), \tag{1.1}$$

where Y is output (value added in real terms), K is capital input, and L is labor input ($K \geq 0$, $L \geq 0$). The idea is that for several issues it is useful to think of output as a homogeneous good which is produced by two inputs, one of which is *capital,* by which we mean a *producible* durable means of production, the other being *labor,* usually considered a *non-producible* human input. Of course, thinking of these variables as representing one-dimensional entities is a drastic abstraction, but may nevertheless be worthwhile in a first approach.

Simple as it looks, an equation like (1.1) is not always interpreted in the right way. A key issue here is: how are the variables entering (1.1) *denominated,* that is, in what units are the variables measured? It is most satisfactory, both from a theoretical and empirical point of view, to think of both outputs and inputs as *flows:* quantities per unit of time. This is generally recognized as far as Y is concerned. Unfortunately, it is less recognized concerning K and L, a circumstance which is probably related to a *tradition in macroeconomic notation*, as we will now explain.

Let the time unit be one year. Then the K appearing in the production function should be seen as the number of machine hours per year. Similarly, L should be seen as the number of labor hours per year. Unless otherwise specified, it should be understood that the rate of utilization of the production factors is constant over time; for convenience, one can then *normalize the rate of utilization of each factor to equal one.* Thus, with one year as our time unit, we imagine that "normally" a machine is in operation in h hours during a year. Then, we define one *machine-year* as the service of a machine in operation h hours a year. If K machines are in operation and on average deliver one machine year per year, then the total capital input is K machine-years per year:

$$K \text{ (machine-yrs/yr)} = K \text{ (machines)} \times 1 \text{ ((machine-yrs/yr)/machine)}, \tag{1.2}$$

where the denomination of the variables is indicated in brackets. Similarly, if the stock of laborers is L men and on average they deliver one *man-year* (say h hours) per year, then the total labor input is L man-years per year:

$$L(\text{man-yrs/yr}) = L(\text{men}) \times 1((\text{man-yrs/yr})/\text{man}). \tag{1.3}$$

One of the reasons that confusion of stocks and flows may arise is the tradition in macroeconomics to use the same symbol, K, for the capital *input* (the number of machine hours per year), in (1.1) as for the capital *stock* in an accumulation equation like

$$K_{t+1} = K_t + I_t - \delta K_t. \tag{1.4}$$

Here the interpretation of K_t is as a capital *stock* (number of machines) at the beginning of period t, I_t is gross investment, and δ is the rate of physical capital depreciation due to wear and tear ($0 \leq \delta \leq 1$). In (1.4) there is no role for the rate of *utilization* of the capital stock, which is, however, of key importance in (1.1). Similarly, there is a tradition in macroeconomics to denote the number of heads in the labor force by L and write, for example, $L_t = L_0(1+n)^t$, where n is a constant growth rate of the labor force. Here the interpretation of L_t is as a stock (number of persons). There is no role for the average rate of utilization in actual employment of this stock over the year.

This text will not attempt a break with this tradition of using the same symbol for two in principle different variables. But we insist on interpretations such that the notation is *consistent*. This requires normalization of the utilization rates for capital and labor in the production function to equal one, as indicated in (1.2) and (1.3) above. We are then allowed to use the same symbol for a stock and the corresponding flow because the *values* of the two variables will coincide.

An illustration of the importance of being aware of the distinction between stock and flows appears when we consider the following measure of per capita income in a given year:

$$\frac{GDP}{N} = \frac{GDP}{\#\text{hours of work}} \times \frac{\#\text{hours of work}}{\#\text{employed workers}} \times \frac{\#\text{employed workers}}{\#\text{workers}} \times \frac{\#\text{workers}}{N}, \tag{1.5}$$

where N, #workers, and #employed workers indicate, say, the average size of the population, the workforce (including the unemployed), and the employed workforce, respectively, during the year. That is, aggregate per capita income equals average labor productivity times average labor intensity times the crude employment rate times the workforce participation rate.[3] An increase from one year to

[3]By the crude employment rate is meant the number of employed individuals, without weighting by the number of hours they work per week, divided by the total number of individuals in the labor force.

the next in the ratio on the left-hand side of the equation reflects the net effect of changes in the four ratios on the right-hand side. Similarly, a fall in per capita income (a ratio between a flow and a stock) need not reflect a fall in productivity (GDP/#hours of work, a ratio of two flows), but may reflect, say, a fall in the number of hours per member of the workforce (#hours of work/#workers) due to a rise in unemployment (fall in #employed workers/workers) or an ageing population (fall in #workers/N).

A *second* conceptual issue concerning the production function in (1.1) relates to the question: what about land and other natural resources? As farming requires land and factories and office buildings require building sites, a third argument, a natural resource input, should in principle appear in (1.1). In theoretical macroeconomics for industrialized economies this third factor is often left out because it does not vary much as an input to production and tends to be of secondary importance in value terms.

A *third* conceptual issue concerning the production function in (1.1) relates to the question: what about *intermediate goods*? By intermediate goods we mean non-durable means of production like raw materials and energy. Certainly, raw materials and energy are generally necessary inputs at the micro level. Then it seems strange to regard output as produced by only capital and labor. The point is that in macroeconomics we often abstract from the engineering input-output relations, involving intermediate goods. We imagine that at a lower stage of production, raw materials and energy are continuously produced by capital and labor, but are then immediately used up at a higher stage of production, again using capital and labor. The value of these materials are not part of value added in the sector or in the economy as a whole. Since value added is what macroeconomics usually focuses at and what the Y in (1.1) represents, materials therefore are often not explicit in the model.

On the other hand, if of interest for the problems studied, the analysis *should,* of course, take into account that at the aggregate level in real world situations, there will generally be a minor difference between produced and used-up raw materials which then constitute net investment in inventories of materials.

To further clarify this point as well as more general aspects of how macroeconomic models are related to national income and product accounts, the next section gives a review of national income accounting.

1.3 Macroeconomic models and national income accounting

Stylized national income and product accounts

(very incomplete)

We give here a stylized picture of national income and product accounts with emphasis on the conceptual structure. The basic point to be aware of is that national income accounting looks at output from *three sides*:

- the production side (value added),
- the use side,
- the income side.

These three "sides" refer to different approaches to the practical measurement of production and income: the "output approach", the "expenditure approach", and the "income approach".

Consider a closed economy with three production sectors. Sector 1 produces raw materials (or energy) in the amount Q_1 per time unit, Sector 2 produces durable capital goods in the amount Q_2 per time unit, and the third sector produces consumption goods in the amount Q_3 per time unit. It is common to distinguish between three basic *production factors* available ex ante a given production process. These are *land* (or, more generally, non-producible natural resources), *labor*, and *capital* (producible durable means of production). In practice also raw materials are a necessary production input. Traditionally, this input has been regarded as itself produced at an early stage within the production process and then used up during the remainder of the production process. In formal dynamic analysis, however, both capital and raw materials are considered produced prior to the production process in which the latter are used up. This is why we include raw materials as a fourth production factor in the production functions of the three sectors.

....

1.4 Some terminological points

On the vocabulary used in this book:

(Incomplete)

Economic terms

Physical capital refers to stocks of *reproducible durable* means of production such as machines and structures. Reproducible *non-durable* means of production include raw materials and energy and are sometimes called intermediate goods. Non-reproducible means of production, such as land and other natural resources, are in this book not included under the heading "capital" but just called *natural resources*.

We follow the convention in macroeconomics and, unless otherwise specified, use "capital" for physical capital, that is, a production factor. In other branches of economics and in everyday language "capital" may mean the funds (sometimes called "financial capital") that finance purchases of physical capital.

By a household's *wealth* (sometimes denoted *net wealth*), W, we mean the value of the total stock of resources possessed by the household at a given point in time. This wealth generally has two main components, the *human wealth*, which is the present value of the stream of future labor income, and the *non-human wealth.* The latter is the sum of the value of the household's *physical assets* (also called *real* assets) and its *net financial assets.* Typically, housing wealth is the dominating component in households' physical assets. By *net financial assets* is meant the difference between the value of financial assets and the value of financial liabilities. *Financial assets* include cash as well as paper claims that entitles the owner to future transfers from the issuer of the claim, perhaps conditional on certain events. Bonds and shares are examples. And a *financial liability* of a household (or other type of agent) is an obligation to transfer resources to others in the future. A mortgage loan is an example.

In spite of this distinction between what is called physical assets and what is called financial assets, often in macroeconomics (and in this book unless otherwise indicated) the household's "financial wealth" is used synonymous with its non-human wealth, that is, including purely physical assets like land, house, car, machines, and other equipment. Somewhat at odds with this convention macroeconomics (including this book) generally uses "investment" as synonymous with "physical capital investment", that is, procurement of new machines and plants by firms and new houses or apartments by households. Then, when having purchases of *financial* assets in mind, macroeconomists talk of *financial investment.*

...

Saving (flow) vs. savings (stock).

...

1.5 Brief history of macroeconomics

Text not yet available.

—

Akerlof and Shiller (2009)

Gali (2008)

1.6 Literature notes

....

The modern theory of economic growth ("new growth theory", "endogenous growth theory") is extensively covered in dedicated textbooks like Aghion and Howitt (1998), Jones (2002), Barro and Sala-i Martin (2004), Acemoglu (2009), and Aghion and Howitt (2009). A good introduction to analytical development economics is Basu (1997).

Snowdon and Vane (1997), Blanchard (2000), and Woodford (2000) present useful overviews of the history of macroeconomics. For surveys on recent developments on the research agenda within theory as well as practical policy analysis, see Mankiw (2006), Blanchard (2008), and Woodford (2009). Somewhat different perspectives, from opposite poles, are offered by Chari et al. (2009) and Colander et al. (2008).

To be incorporated in the preface:

Two textbooks that have been a great inspiration for the one in your hands are Blanchard and Fischer, *Lectures in Macroeconomics*, 1989, and Malinvaud, *Macroeconomic Theory,* vol. A and B, 1998, both of which dig deeper into a lot of the stuff. Compared with Blanchard and Fischer the present book on the one hand of course includes some more recent contributions to macroeconomics, while on ther hand it is more elementary. It is intended to be accessible for third-year undergraduates with a good background in calculus and first-year graduate students. Compared with Malinvaud the emphasis in this book is more on formulating complete dynamic models and analyze their applications and implications.

Chapter 2

Review of technology and firms

The aim of this chapter is threefold. First, we shall introduce this book's vocabulary concerning firms' technology and technological change. Second, we shall refresh our memory of key notions from microeconomics relating to firms' behavior and factor market equilibrium under simplifying assumptions, including perfect competition. Finally, to prepare for the many cases where perfect competition and other simplifying assumptions are not good approximations to reality, we give an introduction to firms' behavior under more realistic conditions including monopolistic competition.

The vocabulary pertaining to other aspects of the economy, for instance households' preferences and behavior, is better dealt with in close connection with the specific models to be discussed in the subsequent chapters. Regarding the distinction between discrete and continuous time analysis, most of the definitions contained in this chapter are applicable to both.

2.1 The production technology

Consider a two-input-one-output production function given by

$$Y = F(K, L), \tag{2.1}$$

where Y is output (value added) per time unit, K is capital input per time unit, and L is labor input per time unit ($K \geq 0$, $L \geq 0$). We may think of (2.1) as describing the output of a firm, a sector, or the economy as a whole. It is in any case a very simplified description, ignoring the heterogeneity of output, capital, and labor. Yet, for many macroeconomic questions it may be a useful first approach.

Note that in (2.1) not only Y but also K and L represent *flows,* that is, quantities per unit of time. If the time unit is one year, we think of K as

measured in machine hours per year. Similarly, we think of L as measured in labor hours per year. Unless otherwise specified, it is understood that the rate of utilization of the production factors is constant over time and normalized to one for each production factor. As explained in Chapter 1, we can then use the same symbol, K, for the *flow* of capital services as for the *stock* of capital. Similarly with L.

2.1.1 A neoclassical production function

By definition, Y, K and L are non-negative. It is generally understood that a production function, $Y = F(K, L)$, is *continuous* and that $F(0, 0) = 0$ (no input, no output). Sometimes, when a production function is specified by a certain formula, that formula may not be defined for $K = 0$ or $L = 0$ or both. In such a case we adopt the convention that the domain of the function is understood extended to include such boundary points whenever it is possible to assign function values to them such that continuity is maintained. For instance the function $F(K, L) = \alpha L + \beta KL/(K + L)$, where $\alpha > 0$ and $\beta > 0$, is not defined at $(K, L) = (0, 0)$. But by assigning the function value 0 to the point $(0, 0)$, we maintain both continuity and the "no input, no output" property, cf. Exercise 2.4.

We call the production function *neoclassical* if for all (K, L), with $K > 0$ and $L > 0$, the following additional conditions are satisfied:

(a) $F(K, L)$ has continuous first- and second-order partial derivatives satisfying:

$$F_K > 0, \quad F_L > 0, \tag{2.2}$$

$$F_{KK} < 0, \quad F_{LL} < 0. \tag{2.3}$$

(b) $F(K, L)$ is strictly quasiconcave (i.e., the level curves, also called isoquants, are strictly convex to the origin).

In words: (a) says that a neoclassical production function has continuous substitution possibilities between K and L and the *marginal productivities* are positive, but diminishing in own factor. Thus, for a given number of machines, adding one more unit of labor, adds to output, but less so, the higher is already the labor input. And (b) says that every isoquant, $F(K, L) = \bar{Y}$, has a strictly convex form qualitatively similar to that shown in Fig. 2.1.[1] When we speak of for example F_L as the marginal *productivity* of labor, it is because the "pure"

[1]For any fixed $\bar{Y} \geq 0$, the associated *isoquant* is the level set $\{(K, L) \in \mathbb{R}_+ |\ F(K, L) = \bar{Y}\}$. A refresher on mathematical terms such as *level set, boundary point, convex function,* etc. is contained in Math Tools.

partial derivative, $\partial Y/\partial L = F_L$, has the denomination of a productivity (output units/yr)/(man-yrs/yr). It is quite common, however, to refer to F_L as the marginal *product* of labor. Then a unit marginal increase in the labor input is understood: $\Delta Y \approx (\partial Y/\partial L)\Delta L = \partial Y/\partial L$ when $\Delta L = 1$. Similarly, F_K can be interpreted as the marginal *productivity* of capital or as the marginal *product* of capital. In the latter case it is understood that $\Delta K = 1$, so that $\Delta Y \approx (\partial Y/\partial K)\Delta K = \partial Y/\partial K$.

The definition of a neoclassical production function can be extended to the case of n inputs. Let the input quantities be $X_1, X_2, \ldots, X_n$ and consider a production function $Y = F(X_1, X_2, \ldots, X_n)$. Then F is called neoclassical if all the marginal productivities are positive, but diminishing in own factor, and F is strictly quasiconcave (i.e., the upper contour sets are strictly convex, cf. Appendix A). An example where $n = 3$ is $Y = F(K, L, J)$, where J is land, an important production factor in an agricultural economy.

Returning to the two-factor case, since $F(K, L)$ presumably depends on the level of technical knowledge and this level depends on time, t, we might want to replace (2.1) by

$$Y_t = F(K_t, L_t, t), \tag{2.4}$$

where the third argument indicates that the production function may shift over time, due to changes in technology. We then say that F is a neoclassical production function if for all t in a certain time interval it satisfies the conditions (a) and (b) w.r.t its first two arguments. *Technological progress* can then be said to occur when, for K_t and L_t held constant, output increases with t.

For convenience, to begin with we skip the explicit reference to time and level of technology.

The marginal rate of substitution Given a neoclassical production function F, we consider the isoquant defined by $F(K, L) = \bar{Y}$, where $\bar{Y}$ is a positive constant. The *marginal rate of substitution*, MRS_{KL}, of K for L at the point (K, L) is defined as the absolute slope of the isoquant $\left\{(K, L) \in \mathbb{R}^2_{++} \middle| F(K, L) = \bar{Y}\right\}$ at that point, cf. Fig. 2.1. For some reason (unknown to this author) the tradition in macroeconomics is to write $Y = F(K, L)$ and in spite of ordering the arguments of F this way, nonetheless have K on the vertical and L on the horizontal axis when considering an isoquant. At this point we follow the tradition.

The equation $F(K, L) = \bar{Y}$ defines K as an implicit function $K = \varphi(L)$ of L. By implicit differentiation we get $F_K(K, L)dK/dL + F_L(K, L) = 0$, from which follows

$$MRS_{KL} \equiv -\frac{dK}{dL}\Big|_{Y=\bar{Y}} = -\varphi'(L) = \frac{F_L(K, L)}{F_K(K, L)} > 0. \tag{2.5}$$

So MRS_{KL} equals the ratio of the marginal productivities of labor and capital, respectively.[2] The economic interpretation of MRS_{KL} is that it indicates (approximately) the amount of K that can be saved by applying an extra unit of labor.

Since F is neoclassical, by definition F is strictly quasi-concave and so the marginal rate of substitution is diminishing as substitution proceeds, i.e., as the labor input is further increased along a given isoquant. Notice that this feature characterizes the marginal rate of substitution for any neoclassical production function, whatever the returns to scale (see below).

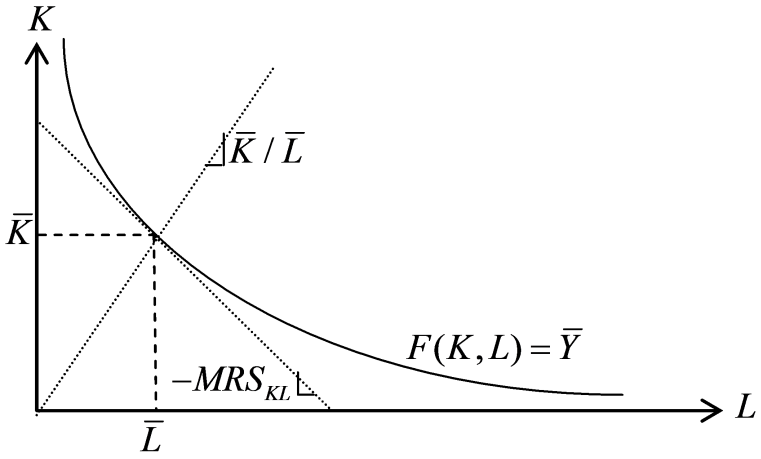

Figure 2.1: MRS_{KL} as the absolute slope of the isoquant representing $F(K, L) = \bar{Y}$.

When we want to draw attention to the dependency of the marginal rate of substitution on the factor combination considered, we write $MRS_{KL}(K, L)$. Sometimes in the literature, the marginal rate of substitution between two production factors, K and L, is called the *technical* rate of substitution (or the technical rate of transformation) in order to distinguish from a consumer's marginal rate of substitution between two consumption goods.

As is well-known from microeconomics, a firm that minimizes production costs for a given output level and given factor prices, will choose a factor combination such that MRS_{KL} equals the ratio of the factor prices. If $F(K, L)$ is homogeneous of degree q, then the marginal rate of substitution depends only on the factor proportion and is thus the same at any point on the ray $K = (\bar{K}/\bar{L})L$. In this case the expansion path is a straight line.

[2] The subscript $\big|_{Y=\bar{Y}}$ in (2.5) signifies that "we are moving along a given isoquant $F(K, L) = \bar{Y}$", i.e., we are considering the relation between K and L under the restriction $F(K, L) = \bar{Y}$. Expressions like $F_L(K, L)$ or $F_2(K, L)$ mean the partial derivative of F w.r.t. the second argument, evaluated at the point (K, L).

The Inada conditions A continuously differentiable production function is said to satisfy the *Inada conditions*[3] if

$$\lim_{K\to 0} F_K(K,L) = \infty, \lim_{K\to\infty} F_K(K,L) = 0, \tag{2.6}$$

$$\lim_{L\to 0} F_L(K,L) = \infty, \lim_{L\to\infty} F_L(K,L) = 0. \tag{2.7}$$

In this case, the marginal productivity of either production factor has no upper bound when the input of the factor becomes infinitely small. And the marginal productivity is gradually vanishing when the input of the factor increases without bound. Actually, (2.6) and (2.7) express *four* conditions, which it is preferable to consider separately and label one by one. In (2.6) we have two *Inada conditions for* MPK (the marginal productivity of capital), the first being a *lower*, the second an *upper* Inada condition for MPK. And in (2.7) we have two *Inada conditions for* MPL (the marginal productivity of labor), the first being a *lower*, the second an *upper* Inada condition for MPL. In the literature, when a sentence like "the Inada conditions are assumed" appears, it is sometimes not made clear which, and how many, of the four are meant. Unless it is evident from the context, it is better to be explicit about what is meant.

The definition of a neoclassical production function we have given is quite common in macroeconomic journal articles and convenient because of its flexibility. Yet there are textbooks that define a neoclassical production function more narrowly by including the Inada conditions as a requirement for calling the production function neoclassical. In contrast, in this book, when in a given context we need one or another Inada condition, we state it explicitly as an additional assumption.

2.1.2 Returns to scale

If all the inputs are multiplied by some factor, is output then multiplied by the same factor? There may be different answers to this question, depending on circumstances. We consider a production function $F(K,L)$ where $K > 0$ and $L > 0$. Then F is said to have *constant returns to scale* (CRS for short) if it is homogeneous of degree one, i.e., if for all $(K,L) \in \mathbb{R}^2_{++}$ and all $\lambda > 0$,

$$F(\lambda K, \lambda L) = \lambda F(K,L).$$

As all inputs are scaled up or down by some factor, output is scaled up or down by the same factor.[4] The assumption of CRS is often defended by the *replication*

[3] After the Japanese economist Ken-Ichi Inada, 1925-2002.

[4] In their definition of a neoclassical production function some textbooks add constant returns to scale as a requirement besides (a) and (b) above. This book follows the alternative

argument saying that "by doubling all inputs we are always able to double the output since we are essentially just replicating a viable production activity". Before discussing this argument, lets us define the two alternative "pure" cases.

The production function $F(K, L)$ is said to have *increasing returns to scale* (IRS for short) if, for all $(K, L) \in \mathbb{R}^2_{++}$ and all $\lambda > 1$,

$$F(\lambda K, \lambda L) > \lambda F(K, L).$$

That is, IRS is present if, when increasing the *scale* of operations by scaling up every input by some factor > 1, output is scaled up by *more* than this factor. One argument for the plausibility of this is the presence of equipment indivisibilities leading to high unit costs at low output levels. Another argument is that gains by specialization and division of labor, synergy effects, etc. may be present, at least up to a certain level of production. The IRS assumption is also called the *economies of scale* assumption.

Another possibility is *decreasing returns to scale* (DRS). This is said to occur when for all $(K, L) \in \mathbb{R}^2_{++}$ and all $\lambda > 1$,

$$F(\lambda K, \lambda L) < \lambda F(K, L).$$

That is, DRS is present if, when all inputs are scaled up by some factor, output is scaled up by *less* than this factor. This assumption is also called the *diseconomies of scale* assumption. The underlying hypothesis may be that control and coordination problems confine the expansion of size. Or, considering the "replication argument" below, DRS may simply reflect that behind the scene there is an additional production factor, for example land or a irreplaceable quality of management, which is tacitly held fixed, when the factors of production are varied.

EXAMPLE 1 The production function

$$Y = AK^{\alpha}L^{\beta}, \qquad A > 0, 0 < \alpha < 1, 0 < \beta < 1, \tag{2.8}$$

where A, α, and β are given parameters, is called a *Cobb-Douglas production function*. The parameter A depends on the choice of measurement units; for a given such choice it reflects efficiency, also called the "total factor productivity". Exercise 2.2 asks the reader to verify that (2.8) satisfies (a) and (b) above and is therefore a neoclassical production function. The function is homogeneous of degree $\alpha + \beta$. If $\alpha + \beta = 1$, there are CRS. If $\alpha + \beta < 1$, there are DRS, and if

terminology where, if in a given context an assumption of constant returns to scale is needed, this is stated as an additional assumption and we talk about a *CRS-neoclassical production function*.

$\alpha + \beta > 1$, there are IRS. Note that α and β must be less than 1 in order not to violate the diminishing marginal productivity condition. □

EXAMPLE 2 The production function

$$Y = A\left[\alpha K^{\beta} + (1-\alpha)L^{\beta}\right]^{\frac{1}{\beta}}, \tag{2.9}$$

where A, α, and β are parameters satisfying $A > 0$, $0 < \alpha < 1$, and $\beta < 1$, $\beta \neq 0$, is called a *CES production function* (CES for Constant Elasticity of Substitution). For a given choice of measurement units, the parameter A reflects efficiency (or "total factor productivity") and is thus called the *efficiency parameter.* The parameters α and β are called the *distribution parameter* and the *substitution parameter,* respectively. The latter name comes from the property that the higher is β, the more sensitive is the cost-minimizing capital-labor ratio to a rise in the relative factor price. Equation (2.9) gives the CES function for the case of constant returns to scale; the cases of increasing or decreasing returns to scale are presented in Chapter 4.5. A limiting case of the CES function (2.9) gives the Cobb-Douglas function with CRS. Indeed, for fixed K and L,

$$\lim_{\beta \to 0} A\left[\alpha K^{\beta} + (1-\alpha)L^{\beta}\right]^{\frac{1}{\beta}} = AK^{\alpha}L^{1-\alpha}.$$

This and other properties of the CES function are shown in Chapter 4.5. The CES function has been used intensively in empirical studies. □

EXAMPLE 3 The production function

$$Y = \min(AK, BL), \qquad A > 0, B > 0, \tag{2.10}$$

where A and B are given parameters, is called a *Leontief production function*[5] (or a *fixed-coefficients production function; A* and B are called the *technical coefficients.* The function is not neoclassical, since the conditions (a) and (b) are not satisfied. Indeed, with this production function the production factors are not substitutable at all. This case is also known as the case of *perfect complementarity* between the production factors. The interpretation is that already installed production equipment requires a fixed number of workers to operate it. The inverse of the parameters A and B indicate the required capital input per unit of output and the required labor input per unit of output, respectively. Extended to many inputs, this type of production function is often used in multi-sector input-output models (also called Leontief models). In aggregate analysis neoclassical production functions, allowing substitution between capital and labor, are more popular

[5] After the Russian-American economist and Nobel laureate Wassily Leontief (1906-99) who used a generalized version of this type of production function in what is known as *input-output analysis.*

than Leontief functions. But sometimes the latter are preferred, in particular in short-run analysis with focus on the use of already installed equipment where the substitution possibilities tend to be limited.[6] As (2.10) reads, the function has CRS. A generalized form of the Leontief function is $Y = \min(AK^\gamma, BL^\gamma)$, where $\gamma > 0$. When $\gamma < 1$, there are DRS, and when $\gamma > 1$, there are IRS. □

The replication argument The assumption of CRS is widely used in macroeconomics. The model builder may appeal to the *replication argument*. This is the argument saying that by doubling all the inputs, we should always be able to double the output, since we are just "replicating" what we are already doing. Suppose we want to double the production of cars. We may then build another factory identical to the one we already have, man it with identical workers and deploy the same material inputs. Then it is reasonable to assume output is doubled.

In this context it is important that the CRS assumption is about *technology* in the sense of functions linking outputs to inputs. Limits to the *availability* of input resources is an entirely different matter. The fact that for example managerial talent may be in limited supply does not preclude the thought experiment that *if* a firm could double all its inputs, including the number of talented managers, then the output level could also be doubled.

The replication argument presupposes, first, that *all* the relevant inputs are explicit as arguments in the production function; second, that these are changed equiproportionately. This, however, exhibits the weakness of the replication argument as a defence for assuming CRS of our present production function, F. One could easily make the case that besides capital and labor, also land is a necessary input and should appear as a separate argument.[7] If an industrial firm decides to duplicate what it has been doing, it needs a piece of land to build another plant like the first. Then, on the basis of the replication argument, we should in fact expect DRS w.r.t. capital and labor alone. In manufacturing and services, empirically, this and other possible sources for departure from CRS w.r.t. capital and labor may be minor and so many macroeconomists feel comfortable enough with assuming CRS w.r.t. K and L alone, at least as a first approximation. This approximation is, however, less applicable to poor countries, where natural resources may be a quantitatively important production factor.

There is a further problem with the replication argument. By definition, CRS is present if and only if, by changing all the inputs equiproportionately by *any* positive factor λ (not necessarily an integer), the firm is able to get output changed

[6] Cf. Section 2.5.2.

[7] Recall from Chapter 1 that we think of "capital" as producible means of production, whereas "land" refers to non-producible natural resources, including for instance building sites.

by the same factor. Hence, the replication argument requires that indivisibilities are negligible, which is certainly not always the case. In fact, the replication argument is more an argument *against* DRS than *for* CRS in particular. The argument does not rule out IRS due to synergy effects as scale is increased.

Sometimes the replication line of reasoning is given a more subtle form. This builds on a useful *local* measure of returns to scale, named the *elasticity of scale*.

The elasticity of scale* [8] To allow for indivisibilities and mixed cases (for example IRS at low levels of production and CRS or DRS at higher levels), we need a local measure of returns to scale. One defines the *elasticity of scale*, $\eta(K,L)$, of F at the point (K,L), where $F(K,L)>0$, as

$$\eta(K,L)=\frac{\lambda}{F(K,L)}\frac{dF(\lambda K,\lambda L)}{d\lambda}\approx\frac{\Delta F(\lambda K,\lambda L)/F(K,L)}{\Delta\lambda/\lambda},\text{ evaluated at }\lambda=1. \tag{2.11}$$

So the elasticity of scale at a point (K,L) indicates the (approximate) percentage increase in output when both inputs are increased by 1 percent. We say that

$$\text{if }\eta(K,L)\begin{cases}>1,\text{ then there are locally }\textit{IRS,}\\=1,\text{ then there are locally }\textit{CRS,}\\<1,\text{ then there are locally }\textit{DRS.}\end{cases} \tag{2.12}$$

The production function *may* have the same elasticity of scale everywhere. This is the case if and only if the production function is homogeneous of some degree $h>0$. In that case $\eta(K,L)=h$ for all (K,L) for which $F(K,L)>0$, and h indicates the *global elasticity of scale*. The Cobb-Douglas function, cf. Example 1, is homogeneous of degree $\alpha+\beta$ and has thereby global elasticity of scale equal to $\alpha+\beta$.

Note that the elasticity of scale at a point (K,L) will always equal the sum of the partial output elasticities at that point:

$$\eta(K,L)=\frac{F_K(K,L)K}{F(K,L)}+\frac{F_L(K,L)L}{F(K,L)}. \tag{2.13}$$

This follows from the definition in (2.11) by taking into account that

$$\begin{aligned}\frac{dF(\lambda K,\lambda L)}{d\lambda}&=F_K(\lambda K,\lambda L)K+F_L(\lambda K,\lambda L)L\\&=F_K(K,L)K+F_L(K,L)L,\text{ when evaluated at }\lambda=1.\end{aligned}$$

[8] A section headline marked by * indicates that in a first reading the section can be skipped - or at least just skimmed through.

Fig. 2.2 illustrates a popular case from introductory economics, an average cost curve which from the perspective of the individual firm is U-shaped: at low levels of output there are falling average costs (thus IRS), at higher levels rising average costs (thus DRS).[9] Given the input prices w_K and w_L and a specified output level $F(K,L) = \bar{Y}$, we know that the cost-minimizing factor combination $(\bar{K}, \bar{L})$ is such that $F_L(\bar{K}, \bar{L})/F_K(\bar{K}, \bar{L}) = w_L/w_K$. It is shown in Appendix A that the elasticity of scale at $(\bar{K}, \bar{L})$ will satisfy:

$$\eta(\bar{K}, \bar{L}) = \frac{LAC(\bar{Y})}{LMC(\bar{Y})}, \tag{2.14}$$

where $LAC(\bar{Y})$ is average costs (the minimum unit cost associated with producing $\bar{Y}$) and $LMC(\bar{Y})$ is marginal costs at the output level $\bar{Y}$. The L in LAC and LMC stands for "long-run", indicating that both capital and labor are considered variable production factors within the period considered. At the optimal plant size, Y^*, there is equality between LAC and LMC, implying a unit elasticity of scale. That is, locally we have CRS. That the long-run average costs are here portrayed as rising for $\bar{Y} > Y^*$, is not essential for the argument but may reflect either that coordination difficulties are inevitable or that some additional production factor, say the building site of the plant, is tacitly held fixed.

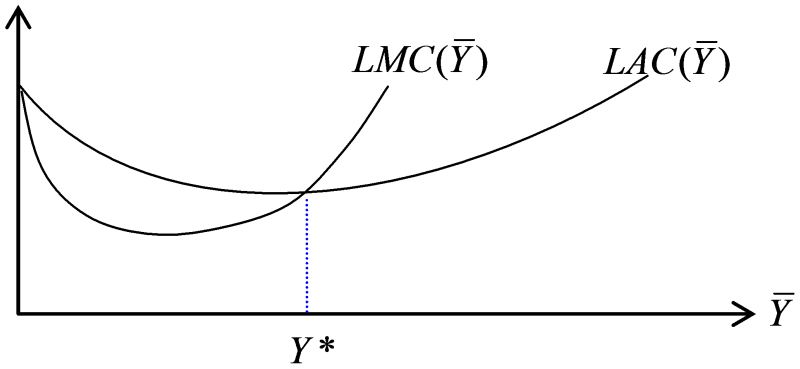

Figure 2.2: Locally CRS at optimal plant size.

Anyway, on this basis Robert Solow (1956) came up with a more subtle replication argument for CRS at the aggregate level. Even though technologies may differ across plants, the surviving plants in a competitive market will have the same average costs at the optimal plant size. In the medium and long run, changes in aggregate output will take place primarily by entry and exit of optimal-size

[9] By a "firm" is generally meant the company as a whole. A company may have several "manufacturing plants" placed at different locations.

plants. Then, with a large number of relatively small plants, each producing at approximately constant unit costs for small output variations, we can without substantial error assume constant returns to scale at the aggregate level. So the argument goes. Notice, however, that even in this form the replication argument is not entirely convincing since the question of indivisibility remains. The optimal, i.e., cost-minimizing, plant size may be large relative to the market − and is in fact so in many industries. Besides, in this case also the perfect competition premise breaks down.

2.1.3 Properties of the production function under CRS

The empirical evidence concerning returns to scale is mixed (see the literature notes at the end of the chapter). Notwithstanding the theoretical and empirical ambiguities, the assumption of CRS w.r.t. capital and labor has a prominent role in macroeconomics. In many contexts it is regarded as an acceptable approximation and a convenient simple background for studying the question at hand.

Expedient inferences of the CRS assumption include:

(i) marginal costs are constant and equal to average costs (so the right-hand side of (2.14) equals unity);

(ii) if production factors are paid according to their marginal productivities, factor payments exactly exhaust total output so that pure profits are neither positive nor negative (so the right-hand side of (2.13) equals unity);

(iii) a production function known to exhibit CRS and satisfy property (a) from the definition of a neoclassical production function above, will automatically satisfy also property (b) and consequently *be* neoclassical;

(iv) a neoclassical two-factor production function with CRS has always $F_{KL} > 0$, i.e., it exhibits "direct complementarity" between K and L;

(v) a two-factor production function that has CRS and is twice continuously differentiable with positive marginal productivity of each factor everywhere in such a way that all isoquants are strictly convex to the origin, *must* have *diminishing* marginal productivities everywhere and thereby be neoclassical.[10]

A principal implication of the CRS assumption is that it allows a reduction of dimensionality. Considering a neoclassical production function, $Y = F(K, L)$

[10] Proof of claim (iii) is in Appendix A and proofs of claim (iv) and (v) are in Appendix B.

with $L > 0$, we can under CRS write $F(K, L) = LF(K/L, 1) \equiv Lf(k)$, where $k \equiv K/L$ is called the *capital-labor ratio* (sometimes the *capital intensity*) and $f(k)$ is the *production function in intensive form* (sometimes named the per capita production function). Thus output per unit of labor depends only on the capital intensity:

$$y \equiv \frac{Y}{L} = f(k).$$

When the original production function F is neoclassical, under CRS the expression for the marginal productivity of capital simplifies:

$$F_K(K, L) = \frac{\partial Y}{\partial K} = \frac{\partial \left[Lf(k)\right]}{\partial K} = Lf'(k)\frac{\partial k}{\partial K} = f'(k). \tag{2.15}$$

And the marginal productivity of labor can be written

$$\begin{aligned} F_L(K, L) &= \frac{\partial Y}{\partial L} = \frac{\partial \left[Lf(k)\right]}{\partial L} = f(k) + Lf'(k)\frac{\partial k}{\partial L} \\ &= f(k) + Lf'(k)K(-L^{-2}) = f(k) - f'(k)k. \end{aligned} \tag{2.16}$$

A neoclassical CRS production function in intensive form always has a positive first derivative and a negative second derivative, i.e., $f' > 0$ and $f'' < 0$. The property $f' > 0$ follows from (2.15) and (2.2). And the property $f'' < 0$ follows from (2.3) combined with

$$F_{KK}(K, L) = \frac{\partial f'(k)}{\partial K} = f''(k)\frac{\partial k}{\partial K} = f''(k)\frac{1}{L}.$$

For a neoclassical production function with CRS, we also have

$$f(k) - f'(k)k > 0 \text{ for all } k > 0, \tag{2.17}$$

in view of $f(0) \geq 0$ and $f'' < 0$. Moreover,

$$\lim_{k \to 0} \left[f(k) - f'(k)k\right] = f(0). \tag{2.18}$$

Indeed, from the *mean value theorem*[11] we know there exists a number $a \in (0, 1)$ such that for any $k > 0$ we have $f(k) - f(0) = f'(ak)k$. From this follows $f(k) - f'(ak)k = f(0) < f(k) - f'(k)k$, since $f'(ak) > f'(k)$ by $f'' < 0$. In view of $f(0) \geq 0$, this establishes (2.17). And from $f(k) > f(k) - f'(k)k > f(0)$ and continuity of f follows (2.18).

[11] This theorem says that if f is continuous in $[\alpha, \beta]$ and differentiable in (α, β), then there exists at least one point γ in (α, β) such that $f'(\gamma) = (f(\beta) - f(\alpha))/(\beta - \alpha)$.

Under CRS the Inada conditions for MPK can be written

$$\lim_{k\to 0} f'(k) = \infty, \qquad \lim_{k\to\infty} f'(k) = 0. \tag{2.19}$$

In this case standard parlance is just to say that "f satisfies the Inada conditions".

An input which must be positive for positive output to arise is called an *essential input*; an input which is not essential is called an *inessential input*. The second part of (2.19), representing the upper Inada condition for MPK under CRS, has the implication that *labor* is an essential input; but capital need not be, as the production function $f(k) = a + bk/(1+k)$, $a > 0, b > 0$, illustrates. Similarly, under CRS the upper Inada condition for MPL implies that *capital* is an essential input. These claims are proved in Appendix C. Combining these results, when *both* the upper Inada conditions hold and CRS obtain, then both capital and labor are essential inputs.[12]

Fig. 2.3 is drawn to provide an intuitive understanding of a neoclassical CRS production function and at the same time illustrate that the lower Inada conditions are more questionable than the upper Inada conditions. The left panel of Fig. 2.3 shows output per unit of labor for a *CRS neoclassical production function* satisfying the Inada conditions for MPK. The $f(k)$ in the diagram could for instance represent the Cobb-Douglas function in Example 1 with $\beta = 1-\alpha$, i.e., $f(k) = Ak^{\alpha}$. The right panel of Fig. 2.3 shows a non-neoclassical case where only two alternative *Leontief techniques* are available, technique 1: $y = \min(A_1 k, B_1)$, and technique 2: $y = \min(A_2 k, B_2)$. In the exposed case it is assumed that $B_2 > B_1$ and $A_2 < A_1$ (if $A_2 \geq A_1$ at the same time as $B_2 > B_1$, technique 1 would not be efficient, because the same output could be obtained with less input of at least one of the factors by shifting to technique 2). If the available K and L are such that $k \equiv K/L < B_1/A_1$ or $k > B_2/A_2$, some of either L or K, respectively, is idle. If, however, the available K and L are such that $B_1/A_1 < k < B_2/A_2$, it is efficient to *combine* the two techniques and use the fraction μ of K and L in technique 1 and the remainder in technique 2, where $\mu = (B_2/A_2 - k)/(B_2/A_2 - B_1/A_1)$. In this way we get the "labor productivity curve" OPQR (the envelope of the two techniques) in Fig. 2.3. Note that for $k \to 0$, MPK stays equal to $A_1 < \infty$, whereas for all $k > B_2/A_2$, $MPK = 0$.

A similar feature remains true, when we consider *many,* say n, alternative efficient Leontief techniques available. Assuming these techniques cover a considerable range w.r.t. the B/A ratios, we get a labor productivity curve looking more like that of a neoclassical CRS production function. On the one hand, this gives some intuition of what lies behind the assumption of a neoclassical CRS production function. On the other hand, it remains true that for all $k > B_n/A_n$,

[12] Given a Cobb-Douglas production function, both production factors are essential whether we have DRS, CRS, or IRS.

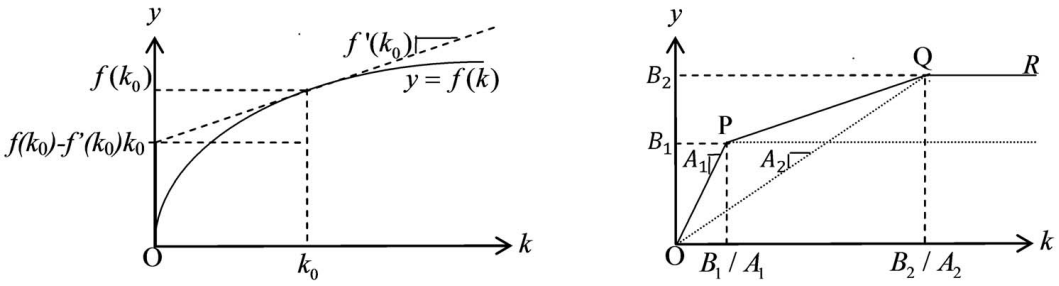

Figure 2.3: Two labor productivity curves based on CRS technologies. Left: neoclassical technology with Inada conditions for MPK satisfied; the graphical representation of MPK and MPL at $k = k_0$ as $f'(k_0)$ and $f(k_0) - f'(k_0)k_0$ are indicated. Right: the line segment PQ makes up an efficient combination of two efficient Leontief techniques.

$MPK = 0$,[13] whereas for $k \to 0$, MPK stays equal to $A_1 < \infty$, thus questioning the lower Inada condition.

The implausibility of the lower Inada conditions is also underlined if we look at their implication in combination with the more reasonable upper Inada conditions. Indeed, the four Inada conditions taken *together* imply, under CRS, that output has no upper bound when either input goes towards infinity for fixed amount of the other input (see Appendix C).

2.2 Technological change

When considering the movement over time of the economy, we shall often take into account the existence of *technological change*. When technological change occurs, the production function becomes time-dependent. Over time the production factors tend to become more productive: more output for given inputs. To put it differently: the isoquants move inward. When this is the case, we say that the technological change displays *technological progress*.

Concepts of neutral technological change

A first step in taking technological change into account is to replace (2.1) by (2.4). Empirical studies often specialize (2.4) by assuming that technological change take a form known as *factor-augmenting* technological change:

$$Y_t = F(A_t K_t, B_t L_t), \tag{2.20}$$

[13] Here we assume the techniques are numbered according to ranking with respect to the size of B.

where F is a (time-independent) neoclassical production function, Y_t, K_t, and L_t are output, capital, and labor input, respectively, at time t, while A_t and B_t are time-dependent "efficiencies" of capital and labor, respectively, reflecting technological change.

In macroeconomics an even more specific form is often assumed, namely the form of *Harrod-neutral technological change.*[14] This amounts to assuming that A_t in (2.20) is a constant (which we can then normalize to one). So only B_t, which is then conveniently denoted T_t, is changing over time, and we have

$$Y_t = F(K_t, T_t L_t). \tag{2.21}$$

The efficiency of labor, T_t, is then said to indicate the *technology level.* Although one can imagine natural disasters implying a fall in T_t, generally T_t tends to rise over time and then we say that (2.21) represents *Harrod-neutral technological progress.* An alternative name often used for this is *labor-augmenting* technological progress. The names "factor-augmenting" and, as here, "labor-augmenting" have become standard and we shall use them when convenient, although they may easily be misunderstood. To say that a change in T_t is labor-augmenting might be understood as meaning that more labor is required to reach a given output level for given capital. In fact, the opposite is the case, namely that T_t has risen so that less labor input is required. The idea is that the technological change affects the output level *as if* the labor input had been increased exactly by the factor by which T was increased, and nothing else had happened. (We might be tempted to say that (2.21) reflects "labor saving" technological change. But also this can be misunderstood. Indeed, keeping L unchanged in response to a rise in T implies that the same output level requires *less capital* and thus the technological change is "capital saving".)

If the function F in (2.21) is homogeneous of degree one (so that the technology exhibits CRS w.r.t. capital and labor), we may write

$$\tilde{y}_t \equiv \frac{Y_t}{T_t L_t} = F(\frac{K_t}{T_t L_t}, 1) = F(\tilde{k}_t, 1) \equiv f(\tilde{k}_t), \qquad f' > 0, f'' < 0.$$

where $\tilde{k}_t \equiv K_t/(T_t L_t) \equiv k_t/T_t$ (habitually called the "effective" capital intensity or, if there is no risk of confusion, just the capital intensity). In rough accordance with a general trend in aggregate productivity data for industrialized countries we often assume that T grows at a constant rate, g, so that in discrete time $T_t = T_0(1+g)^t$ and in continuous time $T_t = T_0 e^{gt}$, where $g > 0$. The popularity in macroeconomics of the hypothesis of labor-augmenting technological progress derives from its consistency with Kaldor's "stylized facts", cf. Chapter 4.

[14] After the English economist Roy F. Harrod, 1900-1978.

There exists two alternative concepts of neutral technological progress. *Hicks-neutral* technological progress is said to occur if technological development is such that the production function can be written in the form

$$Y_t = T_t F(K_t, L_t), \tag{2.22}$$

where, again, F is a (time-independent) neoclassical production function, while T_t is the growing technology level.[15] The assumption of Hicks-neutrality has been used more in microeconomics and partial equilibrium analysis than in macroeconomics. If F has CRS, we can write (2.22) as $Y_t = F(T_t K_t, T_t L_t)$. Comparing with (2.20), we see that in this case Hicks-neutrality is equivalent to $A_t = B_t$ in (2.20), whereby technological change is said to be *equally factor-augmenting.*

Finally, in a symmetric analogy with (2.21), what is known as *capital-augmenting* technological progress is present when

$$Y_t = F(T_t K_t, L_t). \tag{2.23}$$

Here technological change acts as if the capital input were augmented. For some reason this form is sometimes called *Solow-neutral* technological progress.[16] This association of (2.23) to Solow's name is misleading, however. In his famous growth model,[17] Solow assumed Harrod-neutral technological progress. And in another famous contribution, Solow generalized the concept of Harod-neutrality to the case of *embodied* technological change and capital of *different vintages,* see below.

It is easily shown (Exercise 2.5) that the Cobb-Douglas production function (2.8) (with time-independent output elasticities w.r.t. K and L) satisfies all three neutrality criteria at the same time, if it satisfies one of them (which it does if technological change does not affect α and β). It can also be shown that within the class of neoclassical CRS production functions the Cobb-Douglas function is the only one with this property (see Exercise 4.??).

Note that the neutrality concepts do not say anything about the *source* of technological progress, only about the quantitative form in which it materializes. For instance, the occurrence of Harrod-neutrality should not be interpreted as indicating that the technological change emanates specifically from the labor input in some sense. Harrod-neutrality only means that technological innovations predominantly are such that not only do labor and capital in combination become more productive, but this happens to *manifest itself* in the form (2.21), that is, *as if* an improvement in the quality of the labor input had occurred. (Even when improvement in the quality of the labor input is on the agenda, the result may be a reorganization of the production process ending up in a higher B_t along with, or instead of, a higher A_t in the expression (2.20).)

[15] After the English economist and Nobel Prize laureate John R. Hicks, 1904-1989.
[16] After the American economist and Nobel Prize laureate Robert Solow (1924-).
[17] Solow (1956).

Rival versus nonrival goods

When a production function (or more generally a production possibility set) is specified, a given level of technical knowledge is presumed. As this level changes over time, the production function changes. In (2.4) this dependency on the level of knowledge was represented indirectly by the time dependency of the production function. Sometimes it is useful to let the knowledge dependency be explicit by perceiving knowledge as an additional production factor and write, for instance,

$$Y_t = F(X_t, T_t), \tag{2.24}$$

where T_t is now an index of the amount of knowledge, while X_t is a vector of ordinary inputs like raw materials, machines, labor etc. In this context the distinction between rival and nonrival inputs or more generally the distinction between rival and nonrival goods is important. A good is *rival* if its character is such that one agent's use of it inhibits other agents' use of it at the same time. A pencil is thus rival. Many production inputs like raw materials, machines, labor etc. have this property. They are elements of the vector X_t. By contrast, however, technical knowledge is a *nonrival* good. An arbitrary number of factories can simultaneously use the same piece of technical knowledge in the sense of a *list of instructions about how different inputs can be combined to produce a certain output*. An engineering principle or a farmaceutical formula are examples. (Note that the distinction rival-nonrival is different from the distinction excludable-nonexcludable. A good is *excludable* if other agents, firms or households, can be excluded from using it. Other firms can thus be excluded from commercial use of a certain piece of technical knowledge if it is patented. The existence of a patent concerns the legal status of a piece of knowledge and does not interfere with its economic character as a nonrival input.).

What the replication argument really says is that by, conceptually, doubling all the *rival* inputs, we should always be able to double the output, since we just "replicate" what we are already doing. This is then an argument for (at least) CRS w.r.t. the elements of X_t in (2.24). The point is that because of its nonrivalry, we do not need to increase the stock of knowledge. Now let us imagine that the stock of knowledge *is* doubled at the same time as the rival inputs are doubled. Then *more* than a doubling of output should occur. In this sense we may speak of IRS w.r.t. the rival inputs and T taken together.

The perpetual inventory method

Before proceeding, a brief remark about how the capital stock K_t can be measured While data on gross investment, I_t, is typically available in official national income and product accounts, data on K_t usually is not. It has been up to researchers

and research institutions to make their own time-series for capital. One approach to the measurement of K_t is the *perpetual inventory method* which builds upon the accounting relationship

$$K_t = I_{t-1} + (1-\delta)K_{t-1}. \tag{2.25}$$

Assuming a constant capital depreciation rate δ, backward substitution gives

$$K_t = I_{t-1} + (1-\delta)\left[I_{t-2} + (1-\delta)K_{t-2}\right] = \ldots = \sum_{i=1}^{N}(1-\delta)^{i-1}I_{t-i} + (1-\delta)^T K_{t-N}. \tag{2.26}$$

Based on a long time series for I and an estimate of δ, one can insert these observed values in the formula and calculate K_t, starting from a rough conjecture about the initial value K_{t-N}. The result will not be very sensitive to this conjecture since for large N the last term in (2.26) becomes very small.

Embodied vs. disembodied technological progress*

An additional taxonomy of technological change is the following. We say that technological change is *embodied*, if taking advantage of new technical knowledge requires construction of new investment goods. The new technology is incorporated in the design of newly produced equipment, but this equipment will not participate in subsequent technological progress. An example: only the most recent vintage of a computer series incorporates the most recent advance in information technology. Then investment goods produced later (investment goods of a later "vintage") have higher productivity than investment goods produced earlier at the same resource cost. Thus investment becomes an important driving force in productivity increases.

We may formalize embodied technological progress by writing capital accumulation in the following way:

$$K_{t+1} - K_t = Q_t I_t - \delta K_t, \tag{2.27}$$

where I_t is gross investment in period t, i.e., $I_t = Y_t - C_t$, and Q_t measures the "quality" (productivity) of newly produced investment goods. The rising level of technology implies rising Q so that a given level of investment gives rise to a greater and greater addition to the capital stock, K, measured in *efficiency units*. In aggregate models C and I are produced with the same technology, the aggregate production function. From this together with (2.27) follows that Q capital goods can be produced at the same minimum cost as one consumption good. Hence, the equilibrium price, p, of capital goods in terms of the consumption good must equal the inverse of Q, i.e., $p = 1/Q$. The output-capital ratio in value terms is $Y/(pK) = QY/K$.

Note that even if technological change does not directly appear in the production function, that is, even if for instance (2.21) is replaced by $Y_t = F(K_t, L_t)$, the economy may experience a rising standard of living when Q is growing over time.

In contrast, *disembodied technological change* occurs when new technical and organizational knowledge increases the combined productivity of the production factors independently of when they were constructed or educated. If the K_t appearing in (2.21), (2.22), and (2.23) above refers to the total, historically accumulated capital stock as calculated by (2.26), then the evolution of T in these expressions can be seen as representing disembodied technological change. All vintages of the capital equipment benefit from a rise in the technology level T_t. No new investment is needed to benefit.

Based on data for the U.S. 1950-1990, and taking quality improvements into account, Greenwood et al. (1997) estimate that embodied technological progress explains about 60% of the growth in output per man hour. So, empirically, *embodied* technological progress seems to play a dominant role. As this tends not to be fully incorporated in national income accounting at fixed prices, there is a need to adjust the investment levels in (2.26) to better take estimated quality improvements into account. Otherwise the resulting K will not indicate the capital stock measured in efficiency units.

For most issues dealt with in this book the distinction between embodied and disembodied technological progress is not very important. Hence, unless explicitly specified otherwise, technological change is understood to be disembodied.

2.3 The concepts of a representative firm and an aggregate production function

Many macroeconomic models make use of the simplifying notion of a *representative firm.* By this is meant a fictional firm whose production "represents" aggregate production (value added) in a sector or in society as a whole.

Suppose there are n firms in the sector considered or in society as a whole. Let F^i be the production function for firm i so that $Y_i = F^i(K_i, L_i)$, where Y_i, K_i, and L_i are output, capital input, and labor input, respectively, $i = 1, 2, \ldots, n$. Further, let $Y = \Sigma_{i=1}^n Y_i$, $K = \Sigma_{i=1}^n K_i$, and $L = \Sigma_{i=1}^n L_i$. Ignoring technological change, suppose the aggregate variables are related through some function, F^*, such that we can write

$$Y = F^*(K, L),$$

and such that the choices of a single firm facing this production function coincide with the aggregate outcomes, $\Sigma_{i=1}^n Y_i$, $\Sigma_{i=1}^n K_i$, and $\Sigma_{i=1}^n L_i$, in the original econ-

omy. Then $F^*(K, L)$ is called the *aggregate production function* or the production function of the *representative* firm. It is *as if* aggregate production is the result of the behavior of such a single firm.

A simple example where the aggregate production function is well-defined is the following. Suppose that all firms have the *same* production function so that $Y_i = F(K_i, L_i)$, $i = 1, 2, \ldots, n$. If in addition F has CRS, we have

$$Y_i = F(K_i, L_i) = L_i F(k_i, 1) \equiv L_i f(k_i),$$

where $k_i \equiv K_i/L_i$. Hence, facing given factor prices, cost-minimizing firms will choose the same capital intensity $k_i = k$ for all i. From $K_i = kL_i$ then follows $\sum_i K_i = k \sum_i L_i$ so that $k = K/L$. Thence,

$$Y \equiv \sum Y_i = \sum L_i f(k_i) = f(k) \sum L_i = f(k)L = F(k, 1)L = F(K, L).$$

In this (trivial) case the aggregate production function is well-defined and turns out to be exactly the same as the identical CRS production functions of the individual firms. Moreover, given CRS and $k_i = k$ for all i, we have $\partial Y_i/\partial K_i = f'(k_i) = f'(k) = F_K(K, L)$ for all i. So each firm's marginal productivity of capital is the same as the marginal productivity of capital on the basis of the aggregate production function.

Allowing for the existence of *different* production functions at firm level, we may define the aggregate production function as

$$\begin{aligned} F(K, L) &= \max_{(K_1, L_1, \ldots, K_n, L_n) \geq 0} F^1(K_1, L_1) + \cdots + F^n(K_n, L_n) \\ \text{s.t.} \sum_i K_i &\leq K, \quad \sum_i L_i \leq L. \end{aligned}$$

Here it is no longer generally true that $\partial Y_i/\partial K_i$ $(= F_K^i(K_i, L_i) = \partial Y/\partial K$ $(= F_K(K, L)$.

A next step is to allow also for the existence of different output goods, different capital goods, and different types of labor. This makes the issue even more intricate, of course. Yet, if firms are price taking profit maximizers and face nonincreasing returns to scale, we at least know from microeconomics that the aggregate outcome is *as if,* for given prices, the firms jointly maximize aggregate profit on the basis of their combined production technology. The problem is, however, that the conditions needed for this to imply existence of an aggregate production function which is *well-behaved* (in the sense of inheriting simple qualitative properties from its constituent parts) are restrictive.

Nevertheless macroeconomics often treats aggregate output as a single homogeneous good and capital and labor as being two single and homogeneous inputs.

There was in the 1960s a heated debate about the problems involved in this, with particular emphasis on the aggregation of different kinds of equipment into one variable, the capital stock "K". The debate is known as the "Cambridge controversy" because the dispute was between a group of economists from Cambridge University, UK, and a group from Massachusetts Institute of Technology (MIT), which is located in Cambridge, USA. The former group questioned the theoretical robustness of several of the neoclassical tenets, including the proposition that a higher aggregate capital intensity is induced by a lower rate of interest. Starting at the disaggregate level, an association of this sort is not a logical necessity because, with different production functions across the industries, the relative prices of produced inputs tend to change, when the interest rate changes. While acknowledging the possibility of "paradoxical" relationships, the MIT group maintained that in a macroeconomic context they are likely to cause devastating problems only under exceptional circumstances. In the end this is a matter of empirical assessment.[18]

To avoid complexity and because, for many important issues in macroeconomics, there is today no well-tried alternative, this book is about models that use aggregate constructs like "Y", "K", and "L" as simplifying devices, assuming they are, for a broad class of cases, acceptable in a first approximation. Of course there are cases where some disaggregation is pertinent. When for example the role of imperfect competition is in focus, we shall be ready to (modestly) disaggregate the production side of the economy into several product lines, each producing its own differentiated product (cf. Section 2.5.3).

Like the representative firm, the *representative household* and the *aggregate consumption function* are simplifying notions that should be applied only when they do not get in the way of the issue to be studied. The role of budget constraints may make it even more difficult to aggregate over households than over firms. Yet, *if* (and that is a big if) all households have the *same constant* propensity to consume out of income or wealth, aggregation is straightforward and the representative household is a meaningful simplifying concept. On the other hand, if we aim at understanding, say, the *interaction* between lending and borrowing households, perhaps via financial intermediaries, the representative household is not a useful starting point. Similarly, if the theme is conflicts of interests between firm owners and employees, the existence of *different* types of households should be taken into account. Or if we want to assess the welfare costs of business cycle fluctuations, we have to take heterogeneity into account in view of the fact that exposure to unemployment risk tends to be very unevenly distributed.

[18] In his review of the Cambridge controversy Mas-Colell (1989) concluded that: "What the 'paradoxical' comparative statics [of disaggregate capital theory] has taught us is simply that modelling the world as having a single capital good is not *a priori* justified. So be it."

2.4 The neoclassical competitive one-sector setup

Many *long-run* macromodels, including those in the first chapters to follow, share the same abstract setup regarding the firms and the market environment in which they are placed. We give an account here which will serve as a reference point for these later chapters.

The setup is characterized by the following simplifications:

(a) There is only one produced good, an all-purpose good that can be used for consumption as well as investment. Physical capital is just the accumulated amount of what is left of the produced good after consumption. Models using this simplification are called one-sector models. One may think of "corn", a good that can be used for consumption as well as investment in the form of seed to yield corn next period.

(b) A representative firm maximizes profit subject to a neoclassical production function under non-increasing returns to scale.

(c) Capital goods become productive immediately upon purchase or renting (so installation costs and similar features are ignored).

(d) In all markets *perfect competition* rules and so the economic actors are *price takers*, perceiving no constraint on how much they can sell or buy at the going market price. It is understood that market prices are flexible and adjust quickly to levels required for market clearing.

(e) Factor supplies are inelastic.

(f) There is no uncertainty. When a choice of action is made, the consequences are known.

We call such a setup the *neoclassical competitive one-sector setup*. In many respects it is an abstraction. Nevertheless, the outcome under these conditions is of theoretical interest. Think of Galilei's discovery that a falling body falls with a uniform acceleration as long as it is falling through a *perfect vacuum*.

2.4.1 Profit maximization

We consider a single period. Let the representative firm have the neoclassical production function

$$Y = F(K, L), \tag{2.28}$$

where technological change is ignored. Although in this book often CRS will be assumed, we may throw the CRS outcome in relief by starting with a broader view.

From microeconomics we know that equilibrium with perfect competition is compatible with producers operating under the condition of locally *nonincreasing returns* to scale (cf. Fig. 2.2). In standard macroeconomics it is common to accept a lower level of generality and simply assume that F is a *concave* function. This allows us to carry out the analysis *as if* there were non-increasing returns to scale *everywhere* (see Appendix D).[19]

Since F is neoclassical, we have $F_{KK} < 0$ and $F_{LL} < 0$ everywhere. To guarantee concavity it is then necessary and sufficient to add the assumption that

$$D \equiv F_{KK}(K,L)F_{LL}(K,L) - F_{KL}(K,L)^2 \geq 0, \tag{2.29}$$

holds for all (K,L). This is a simple application of a general theorem on concave functions (see Math Tools).

We consider both K and L as variable production factors. Let the factor prices be denoted w_K and w_L, respectively. For the time being we assume the firm rents the machines it uses; then the price, w_K, of capital services is called the *rental price* or the *rental rate*. As *numeraire* (unit of account) we apply the output good. So all prices are measured in terms of the output good which itself has the price 1. Then *profit*, defined as revenue minus costs, is

$$\Pi = F(K,L) - w_K K - w_L L. \tag{2.30}$$

We assume both production inputs are *variable* inputs. Taking the factor prices as given from the factor markets, the firm's problem is to choose (K,L), where $K \geq 0$ and $L \geq 0$, so as to maximize Π. An interior solution will satisfy the first-order conditions

$$\frac{\partial \Pi}{\partial K} = F_K(K,L) - w_K = 0 \text{ or } F_K(K,L) = w_K, \tag{2.31}$$

$$\frac{\partial \Pi}{\partial L} = F_L(K,L) - w_L = 0 \text{ or } F_L(K,L) = w_L. \tag{2.32}$$

Since F is concave, so is the profit function. The first-order conditions are then *sufficient* for (K,L) to be a solution.

It is now convenient to proceed by considering the two cases, DRS and CRS, separately.

[19] By definition, *concavity* means that by applying a weighted average of two factor combinations, (K_1,L_1) and (K_2,L_2), the obtained output is at least as large as the weighted average of the original outputs, Y_1 and Y_2. So, if $0 < \lambda < 1$ and $(K,L) = \lambda(K_1,L_1) + (1-\lambda)(K_2,L_2)$, then $F(K,L) \geq \lambda F(K_1,L_1) + (1-\lambda)F(K_2,L_2)$.

The DRS case

Suppose the production function satisfies (2.29) with strict inequality everywhere, i.e.,

$$D > 0.$$

In combination with the neoclassical property of diminishing marginal productivities, this implies that F is *strictly concave* which in turn implies DRS everywhere. The factor demands will now be unique. Indeed, the equations (2.31) and (2.32) define the factor demands K^d and L^d ("d" for demand) as implicit functions of the factor prices:

$$K^d = K(w_K, w_L), \qquad L^d = L(w_K, w_L).$$

An easy way to find the partial derivatives of these functions is to first take the differential[20] of both sides of (2.31) and (2.32), respectively:

$$\begin{aligned} F_{KK}dK^d + F_{KL}dL^d &= dw_K, \\ F_{LK}dK^d + F_{LL}dL^d &= dw_L. \end{aligned}$$

Then we interpret these conditions as a system of two linear equations with two unknowns, the variables dK^d and dL^d. The determinant of the coefficient matrix equals D in (2.29) and is in this case positive everywhere. Using Cramer's rule (see Math Tools), we find

$$\begin{aligned} dK^d &= \frac{F_{LL}dw_K - F_{KL}dw_L}{D}, \\ dL^d &= \frac{F_{KK}dw_L - F_{LK}dw_K}{D}, \end{aligned}$$

so that

$$\frac{\partial K^d}{\partial w_K} = \frac{F_{LL}}{D} < 0, \qquad \frac{\partial K^d}{\partial w_L} = -\frac{F_{KL}}{D} < 0 \text{ if } F_{KL} > 0, \quad (2.33)$$

$$\frac{\partial L^d}{\partial w_K} = -\frac{F_{KL}}{D} < 0 \text{ if } F_{KL} > 0, \ \frac{\partial L^d}{\partial w_L} = \frac{F_{KK}}{D} < 0, \quad (2.34)$$

[20] The *differential* of a differentiable function is a convenient tool for deriving results like (2.33) and (2.34). For a function of one variable, $y = f(x)$, the differential is denoted dy (or df) and is defined as $f'(x)dx$, where dx is some arbitrary real number (interpreted as the change in x). For a differentiable function of two variables, $z = g(x, y)$, the *differential* of the function is denoted dz (or dg) and is defined as $dz = g_x(x, y)dx + g_y(x, y)dy$, where dx and dy are arbitrary real numbers.

in view of $F_{LK} = F_{KL}$.[21]

In contrast to the cases of CRS and IRS, here we cannot be sure that direct complementarity ($F_{KL} > 0$) holds everywhere. In any event, the rule for both factors is that when a factor price increases, the demand for the factor in question decreases and under direct complementarity also the demand for the other factor will decrease. Although there is a substitution effect towards higher demand for the factor whose price has not been increased, this is more than offset by the negative output effect, which is due to the higher marginal costs. This is an implication of perfect competition. In a different market structure output may be determined from the demand side (think of a Keynesian short-run model) and then only the substitution effect will be operative. An increase in one factor price will then *increase* the demand for the other factor.

The CRS case

Under CRS, D in (2.29) takes the value

$$D = 0$$

everywhere, as shown in Appendix B. Then the factor prices no longer determine the factor demands uniquely. But the *relative* factor demand, $k^d \equiv K^d/L^d$, is determined uniquely by the *relative* factor price, w_L/w_K. Indeed, by (2.31) and (2.32),

$$MRS = \frac{F_L(K,L)}{F_K(K,L)} = \frac{f(k) - f'(k)k}{f'(k)} \equiv mrs(k) = \frac{w_L}{w_K}, \tag{2.35}$$

where the second equality comes from (2.15) and (2.16). By straightforward calculation,

$$mrs'(k) = -\frac{f(k)f''(k)}{f'(k)^2} = -\frac{kf''(k)/f'(k)}{\alpha(k)} > 0,$$

where $\alpha(k) \equiv kf'(k)/f(k)$ is the elasticity of f w.r.t. k and the numerator is the elasticity of f' w.r.t. k. For instance, in the Cobb-Douglas case $f(k) = Ak^\alpha$, we get $mrs'(k) = (1-\alpha)/\alpha$. Given w_L/w_K, the last equation in (2.35) gives k^d as an implicit function $k^d = k(w_L/w_K)$, where $k'(w_L/w_K) = 1/mrs'(k) > 0$. The solution is illustrated in Fig. 2.4. Under CRS (indeed, for any homogeneous neoclassical production function) the desired capital-labor ratio is an increasing function of the inverse factor price ratio and independent of the output level.

[21] Applying the full content of the *implicit function theorem* (see Math tools), one could directly have written down the results (2.33) and (2.34) and would not need the procedure outlined here, based on differentials. On the other hand the present procedure is probably more intuitive and easier to remember.

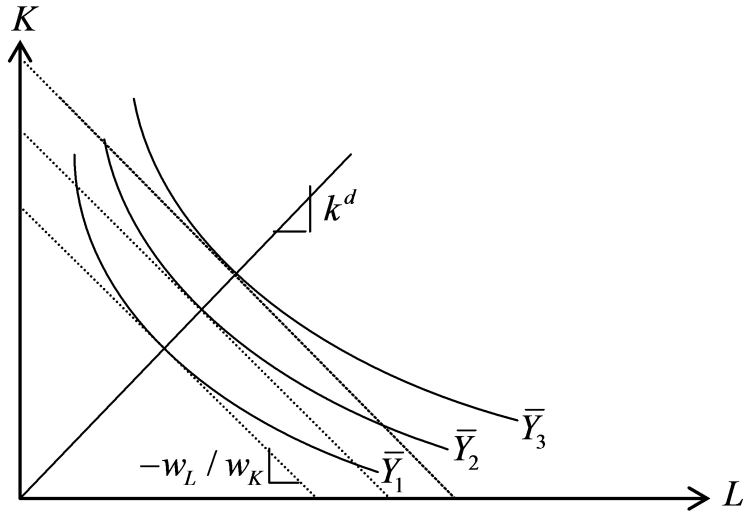

Figure 2.4: Constancy of MRS along rays when the production function is homogeneous of degree h (the cost-minimizing capital intensity is the same at all output levels).

To determine K^d and L^d separately we need to know the level of output. And here we run into the general problem of indeterminacy under perfect competition combined with CRS. Saying that the output level is so as to maximize profit is pointless. Well, if at the going factor prices attainable profit is negative, exit from the market is profit maximizing (or rather loss minimizing), which amounts to $K^d = L^d = 0$. But if the profit is positive, there will be no upper bound to the factor demands. Owing to CRS, doubling the factor inputs will double the profits of a price taking firm. An equilibrium with positive production is only possible if profit is zero. And then the firm is indifferent w.r.t. the level of output. Solving the indeterminacy problem requires a look at the factor markets.

2.4.2 Clearing in factor markets

Considering a closed economy, we denote the available supplies of physical capital and labor K^s and L^s, respectively, and assume these supplies are inelastic. W.r.t. capital this is a "natural" assumption since in a closed economy in the short term the available amount of capital will be *predetermined,* that is, historically determined by the accumulated previous investment in the economy. W.r.t. labor supply it is just a simplifying assumption introduced because the question about possible responses of labor supply to changes in factor prices is a secondary issue in the present context.

The factor markets clear when

$$K^d = K^s, \tag{2.36}$$
$$L^d = L^s. \tag{2.37}$$

Achieving this equilibrium (state of "rest") requires that the factor prices adjust to their equilibrium levels, which are

$$w_K = F_K(K^s, L^s), \tag{2.38}$$
$$w_L = F_L(K^s, L^s), \tag{2.39}$$

by (2.31) and (2.32). This says that in equilibrium the real factor prices are determined by the *marginal productivities of the respective factors at full utilization of the given supplies.* This holds under DRS as well as CRS. So, under non-increasing returns to scale there is, at the macroeconomic level, a unique equilibrium (w_K, w_L, K^d, L^d) given by the above four equilibrium conditions for the factor markets.[22] It is an *equilibrium* in the sense that no agent has an incentive to "deviate".

As to *comparative statics*, since $F_{KK} < 0$, a larger capital supply implies a lower w_K, and since $F_{LL} < 0$, a larger labor supply implies a lower w_L.

The intuitive mechanism behind the *attainment* of equilibrium is that if, for example, for a short moment $w_K < F_K(K^s, L^s)$, then $K^d > K^s$ and so competition between the firms will generate an upward pressure on w_K until equality is obtained. And if for a short moment $w_K > F_K(K^s, L^s)$, then $K^d < K^s$ and so competition between the *suppliers* of capital will generate a downward pressure on w_K until equality is obtained.

Looking more carefully at the matter, however, we see that this intuitive reasoning fits at most the DRS case. In the CRS case we have $F_K(K^s, L^s) = f(k^s)$, where $k^s \equiv K^s/L^s$. Here we can only argue that for instance $w_K < F_K(K^s, L^s)$ implies $k^d > k^s$. And even if this leads to upward pressure on w_K until $k^d = k^s$ is achieved, and even if both factor prices have obtained their equilibrium levels given by (2.38) and (2.39), there is nothing to induce the representative firm (or the many firms in the actual economy taken together) to choose the "right" input *levels* so as to satisfy the clearing conditions (2.36) and (2.37). In this way the indeterminacy under CRS pops up again, this time as a problem endangering stability of the equilibrium.

Stability not guaranteed*

To substantiate the point that the indeterminacy under CRS may endanger stability of competitive equilibrium, let us consider a Walrasian *tâtonnement* ad-

[22] At the microeconomic level, under CRS, industry structure remains indeterminate in that firms are indifferent as to their size.

justment process.[23] We imagine that our period is sub-divided into many short time intervals $(t,\ t+\Delta t)$. In the initial short time interval the factor markets may not be in equilibrium. It is assumed that no capital or labor is hired out of equilibrium. To allow an analysis in continuous time, we let $\Delta t \to 0$. A dot over a variable denotes the time derivative, i.e., $\dot{x}(t) = dx(t)/dt$. The adjustment process assumed is the following:

$$\begin{aligned}
\dot{K}^d(t) &= \lambda_1 \left[F_K(K^d(t), L^d(t)) - w_K(t)\right], \qquad \lambda_1 > 0,\\
\dot{L}^d(t) &= \lambda_2 \left[F_L(K^d(t), L^d(t)) - w_L(t)\right], \qquad \lambda_2 > 0,\\
\dot{w}_K(t) &= K^d(t) - K^s,\\
\dot{w}_L(t) &= L^d(t) - L^s,
\end{aligned}$$

where the initial values, $K^d(0)$, $L^d(0)$, $w_K(0)$, and $w_L(0)$, are given. The parameters λ_1 and λ_2 are constant adjustment speeds. The corresponding adjustment speeds for the factor prices are set equal to one by choice of measurement units of the inputs. Of course, the four endogenous variables should be constrained to be nonnegative, but that is not important for the discussion here. The system has a unique stationary state: $K^d(t) = K^s, L^d(t) = L^s, w_K(t) = K_K(K^s, L^s)$, $w_L(t) = K_L(K^s, L^s)$.

A widespread belief, even in otherwise well-informed circles, seems to be that with such adjustment dynamics, the stationary state is at least *locally asymptotically stable.* By this is meant that there exists a (possibly only small) neighborhood, $\mathcal{N}$, of the stationary state with the property that if the initial state, $(K^d(0),\ L^d(0),\ w_K(0),\ w_L(0))$, belongs to $\mathcal{N}$, then the solution $(K^d(t),\ L^d(t),\ w_K(t),\ w_L(t))$ converges to the stationary state for $t \to \infty$?

Unfortunately, however, this stability property is *not* guaranteed. To bear this out, it is enough to present a counterexample. Let $F(K,L) = K^{\frac{1}{2}}L^{\frac{1}{2}}$, $\lambda_1 = \lambda_2 = K^s = L^s = 1$, and suppose $K^d(0) = L^d(0) > 0$ and $w_K(0) = w_L(0) > 0$. All this symmetry implies that $K^d(t) = L^d(t) = x(t) > 0$ and $w_K(t) = w_L(t) = w(t)$ for all $t \geq 0$. So $F_K(K^d(t), L^d(t)) = 0.5x(t)^{-0.5}x(t)^{0.5} = 0.5$, and similarly $F_L(K^d(t), L^d(t)) = 0.5$ for all $t \geq 0$. Now the system is equivalent to the two-dimensional system,

$$\dot{x}(t) = 0.5 - w(t), \tag{2.40}$$

$$\dot{w}(t) = x(t) - 1. \tag{2.41}$$

Using the theory of coupled linear differential equations, the solution is[24]

$$x(t) = 1 + (x(0) - 1)\cos t - (w(0) - 0.5)\sin t, \tag{2.42}$$

$$w(t) = 0.5 + (w(0) - 0.5)\cos t + (x(0) - 1)\sin t. \tag{2.43}$$

[23] *Tâtonnement* is a French word meaning "groping".

[24] For details, see hints in Exercise 2.6.

The solution exhibits undamped oscillations and never settles down at the stationary state, $(1, 0.5)$, if not being there from the beginning. In fact, the solution curves in the (x, w) plane will be circles around the stationary state. This is so whatever the size of the initial distance, $\sqrt{(x(0)-1)^2+(w(0)-0.5)^2}$, to the stationary point.

The economic mechanism is as follows. Suppose for instance that $x(0) < 1$ and $w(0) < 0.5$. Then to begin with there is excess supply and so w will be falling while, with w below marginal products, x will be increasing. When x reaches its potential equilibrium value, 1, w is at its trough and so induces further increases in the factor demands, thus bringing about a phase where $x > 1$. This excess demand causes w to begin an upturn. When w reaches its potential equilibrium value, 0.5, however, excess demand, $x - 1$, is at its peak and this induces further increases in factor prices, w. This brings about a phase where $w > 0.5$ so that factor prices exceed marginal products, which leads to declining factor demands. But as x comes back to its potential equilibrium value, w is at its peak and drives x further down. Thus excess supply arises which in turn triggers a downturn of w. This continues in never ending oscillations where the overreaction of one variable carries the seed to an overreaction of the other variable soon after and so on.

This possible outcome underlines that the theoretical *existence* of equilibrium is one thing and *stability* of the equilibrium is another. In particular under CRS, where demand *functions* for inputs are absent, the issue of stability can be more intricate than one might at first glance think.

The link between capital costs and the interest rate*

Returning to the description of equilibrium, we shall comment on the relationship between the factor price w_K and the more everyday concept of an interest rate. The factor price w_K is the cost per unit of capital service. It has different names in the literature such as the *rental price,* the *rental rate,* the *unit capital cost,* or the *user cost*. It is related to the interest and depreciation costs that the owner of the capital good in question defrays. In the simple neoclassical setup considered here, it does not matter whether the firm rents the capital it uses or owns it; in the latter case, w_K, is the *imputed* capital cost, i.e., the forgone interest plus depreciation.

As to depreciation it is common in simple macroeconomics to apply the approximation that, due to wear and tear, a constant fraction δ (where $0 \leq \delta \leq 1$) of a given capital stock evaporates per period. If for instance the period length is one year and $\delta = 0.1$, this means that a given machine in the next year has only the fraction 0.9 of its productive capacity in the current year. Otherwise the productive characteristics of a capital good are assumed to be the same whatever its time of birth. Sometimes δ is referred to as the rate of *physical* capital depre-

ciation or the *deterioration rate*. When changes in relative prices can occur, this must be distinguished from the *economic depreciation* of capital which refers to the loss in economic value of a machine after one year.

Let p_{t-1} be the price of a certain type of machine bought at the end of period $t-1$. Let prices be expressed in the same numeraire as that in which the interest rate, r, is measured. And let p_t be the price of the same type of machine one period later. Then the *economic depreciation* in period t is

$$p_{t-1} - (1-\delta)p_t = \delta p_t - (p_t - p_{t-1}).$$

The economic depreciation thus equals the value of the physical wear and tear minus the capital gain (positive or negative) on the machine.

By holding the machine the owner faces an opportunity cost, namely the forgone interest on the value p_{t-1} placed in the machine during period t. If r_t is the interest rate on a loan from the end of period $t-1$ to the end of period t, this interest cost is $r_t p_{t-1}$. The benefit of holding the (new) machine is that it can be rented out to the representative firm and provide the return w_{Kt} at the end of the period. Since there is no uncertainty, in equilibrium we must then have $w_{Kt} = r_t p_{t-1} + \delta p_t - (p_t - p_{t-1})$, or

$$\frac{w_{Kt} - \delta p_t + p_t - p_{t-1}}{p_{t-1}} = r_t. \tag{2.44}$$

This is a *no-arbitrage* condition saying that the rate of return on holding the machine equals the rate of return obtainable in the loan market (no profitable arbitrage opportunities are available).[25]

In the simple setup considered so far, the capital good and the produced good are physically identical and thus have the same price. As the produced good is our numeraire, we have $p_{t-1} = p_t = 1$. This has two implications. *First*, the interest rate, r_t, is a real interest rate so that $1 + r_t$ measures the rate at which future units of output can be traded for current units of output. *Second,* (2.44) simplifies to

$$w_{Kt} - \delta = r_t.$$

Combining this with equation (2.38), we see that in the simple neoclassical setup the equilibrium real interest rate is determined as

$$r_t = F_K(K_t^s, L_t^s) - \delta, \tag{2.45}$$

[25] In continuous time analysis the rental rate, the interest rate, and the price of the machine are considered as differentiable functions of time, $w_K(t)$, $r(t)$, and $p(t)$, respectively. In analogy with (2.44) we then get $w_K(t) = (r(t) + \delta)p(t) - \dot{p}(t)$, where $\dot{p}(t)$ denotes the time derivative of the price $p(t)$.

where K_t^S and L_t^s are predetermined. Under CRS this takes the form $r_t = f'(k_t^s) - \delta$, where $k_t^s \equiv K_t^s / L_t^s$.

We have assumed that the firms rent capital goods from their owners, presumably the households. But as long as there is no uncertainty, no capital adjustment costs, and no taxation, it will have no consequences for the results if instead we assume that the firms own the physical capital they use and finance capital investment by issuing bonds or shares. Then such bonds and shares would constitute financial assets, owned by the households and offering a rate of return r_t as given by (2.45).

2.5 More complex model structures*

The neoclassical setup described above may be useful as a first way of organizing one's thoughts about the production side of the economy. To come closer to a model of how modern economies function, however, many modifications and extensions are needed.

2.5.1 Convex capital installation costs

In the real world the capital goods used by a production firm are usually owned by the firm itself rather than rented for single periods on rental markets. This is because inside the specific plant in which these capital goods are an integrated part, they are generally worth much more than outside. So in practice firms acquire and install fixed capital equipment with a view on maximizing discounted expected profits in the future. The cost associated with this fixed capital investment not only includes the purchase price of new equipment, but also the *installation costs* (the costs of setting up the new fixed equipment in the firm and the associated costs of reorganizing work processes).

Assuming the installation costs are strictly convex in the level of investment, the firm has to solve an *intertemporal* optimization problem. Forward-looking expectations thus become important and this has implications for how equilibrium in the output market is established and how the equilibrium interest rate is determined**.** Indeed, in the simple neoclassical setup above, the interest rate equilibrates the market for capital services. The value of the interest rate is simply tied down by the equilibrium condition (2.39) in this market and what happens in the output market is a trivial consequence of this. But with convex capital installation costs the firm's capital stock is given in the short run and the interest rate(s) become(s) determined elsewhere in the model, as we shall see in chapters 14 and 15.

2.5.2 Long-run vs. short-run production functions

In the discussion of production functions up to now we have been silent about the distinction between "ex ante" and "ex post" substitutability between capital and labor. By ex ante is meant "when plant and machinery are to be decided upon" and by ex post is meant "after the equipment is designed and constructed". In the standard neoclassical competitive setup like in (2.35) there is a presumption that also after the construction and installation of the equipment in the firm, the ratio of the factor inputs can be fully adjusted to a change in the relative factor price. In practice, however, when some machinery has been constructed and installed, its functioning will often require a more or less fixed number of machine operators. What can be varied is just the *degree of utilization* of the machinery. That is, after construction and installation of the machinery, the choice opportunities are no longer described by the neoclassical production function but by a Leontief production function,

$$Y = \min(Au\bar{K}, BL), \qquad A > 0, B > 0, \tag{2.46}$$

where $\bar{K}$ is the size of the installed machinery (a fixed factor in the short run) measured in efficiency units, u is its utilization rate ($0 \leq u \leq 1$), and A and B are given technical coefficients measuring efficiency (cf. Section 2.1.2).

So in the short run the choice variables are u and L. In fact, essentially only u is a choice variable since efficient production trivially requires $L = Au\bar{K}/B$. Under "full capacity utilization" we have $u = 1$ (each machine is used 24 hours per day seven days per week). "Capacity" is given as $A\bar{K}$ per week. Producing efficiently at capacity requires $L = A\bar{K}/B$ and the marginal product by increasing labor input is here nil. But if demand, Y^d, is *less* than capacity, satisfying this demand efficiently requires $L = Y^d/B$ and $u = BL/(A\bar{K}) < 1$. As long as $u < 1$, the marginal productivity of labor is a *constant*, B.

The various efficient input proportions that are possible *ex ante* may be approximately described by a neoclassical CRS production function. Let this function on intensive form be denoted $y = f(k)$. When investment is decided upon and undertaken, there is thus a choice between alternative efficient pairs of the technical coefficients A and B in (2.46). These pairs satisfy

$$f(k) = Ak = B. \tag{2.47}$$

So, for an increasing sequence of k's, $k_1, k_2, \ldots, k_i, \ldots$, the corresponding pairs are $(A_i, B_i) = (f(k_i)/k_i, f(k_i))$, $i = 1, 2, \ldots$.[26] We say that ex ante, depending on the relative factor prices as they are "now" and are expected to evolve in the future,

[26] The points P and Q in the right-hand panel of Fig. 2.3 can be interpreted as constructed this way from the neoclassical production function in the left-hand panel of the figure.

a suitable technique, (A_i, B_i), is chosen from an opportunity set described by the given neoclassical production function. But ex post, i.e., when the equipment corresponding to this technique is installed, the production opportunities are described by a Leontief production function with $(A, B) = (A_i, B_i)$.

In the picturesque language of Phelps (1963), technology is in this case *putty-clay.* Ex ante the technology involves capital which is "putty" in the sense of being in a malleable state which can be transformed into a range of various machinery requiring capital-labor ratios of different magnitude. But once the machinery is constructed, it enters a "hardened" state and becomes "clay". Then factor substitution is no longer possible; the capital-labor ratio at full capacity utilization is fixed at the level $k = B_i/A_i$, as in (2.46). Following the terminology of Johansen (1972), we say that a putty-clay technology involves a "long-run production function" which is neoclassical and a "short-run production function" which is Leontief.

Table 1. Technologies classified according to factor substitutability ex ante and ex post.

	Ex post substitution	
Ex ante substitution	possible	impossible
possible	putty-putty	putty-clay
impossible		clay-clay

In contrast, the standard neoclassical setup assumes the same range of substitutability between capital and labor ex ante and ex post. Then the technology is called *putty-putty.* This term may also be used if ex post there is at least *some* substitutability although less than ex ante. At the opposite pole of putty-putty we may consider a technology which is *clay-clay.* Here neither ex ante nor ex post is factor substitution possible. Table 1 gives an overview of the alternative cases.

The putty-clay case is generally considered the realistic case. As time proceeds, technological progress occurs. To take this into account, we may replace (2.47) and (2.46) by $f(k_t, t) = A_t k_t = B_t$ and $Y_t = \min(A_t u_t \bar{K}_t, B_t L_t)$, respectively. If a new pair of Leontief coefficients, (A_{t_2}, B_{t_2}), efficiency-dominates its predecessor (by satisfying $A_{t_2} \geq A_{t_1}$ and $B_{t_2} \geq B_{t_1}$ with at least one strict equality), it may pay the firm to invest in the new technology at the same time as some old machinery is scrapped. Real wages tend to rise along with technological progress and the scrapping occurs because the revenue from using the old machinery in production no longer covers the associated labor costs.

The clay property ex-post of many technologies is important for short-run analysis. It implies that there may be non-decreasing marginal productivity of

labor up to a certain point. It also implies that in its investment decision the firm will have to take expected future technologies and future factor prices into account. For many issues in long-run analysis the clay property ex-post may be less important, since over time adjustment takes place through new investment.

2.5.3 A simple portrayal of price-making firms

Another modification which is important in short- and medium-run analysis, relates to the assumed market forms. Perfect competition is not a good approximation to market conditions in manufacturing and service industries. To bring perfect competition in the output market in perspective, we give here a brief review of firms' behavior under a form of monopolistic competition that is applied in many short-run models.

Suppose there is a large number of differentiated goods, $i = 1, 2, \ldots, n$, each produced by a separate firm. In the short run n is given. Each firm has monopoly on its own good (supported, say, by a trade mark, patent protection, or simply secrecy regarding the production recipe). The goods are imperfect substitutes to each other and so indirect competition prevails. Each firm is small in relation to the "sum" of competing firms and perceives that these other firms do not respond to its actions.

In the given period let firm i face a given downward-sloping demand curve for its product,

$$Y_i \leq \left(\frac{P_i}{P}\right)^{-\varepsilon} \frac{Y}{n} \equiv \mathcal{D}(P_i), \qquad \varepsilon > 1. \tag{2.48}$$

Here Y_i is the produced quantity and the expression on the right-hand side of the inequality is the demand as a function of the price P_i chosen by the firm.[27] The "general price level" P (a kind of average across the different goods, cf. Chapter 22) and the "general demand level", given by the index Y, matter for the position of the demand curve in the (Y_i, P_i) plan, cf. Fig. 2.5. The price elasticity of demand, ε, is assumed constant and higher than one (otherwise there is no solution to the monopolist's decision problem). Variables that the monopolist perceives as exogenous are implicit in the demand function symbol $\mathcal{D}$. We imagine prices are expressed in terms of money (so they are "nominal" prices, hence denoted by capital letters whereas we generally use small letters for "real" prices).

For simplicity, factor markets are still assumed competitive. Given the nominal factor prices, W_K and W_L, firm i wants to maximize its profit

$$\Pi_i = P_i Y_i - W_K K_i - W_L L_i,$$

[27] We ignore production for inventory holding.

subject to (2.48) and the neoclassical production function $Y_i = F(K_i, L_i)$. For the purpose of simple comparison with the case of perfect competition as described in Section 2.4, we return to the case where both labor and capital are variable inputs in the short run.[28] It is no serious restriction on the problem to assume the monopolist will want to produce the amount demanded so that $Y_i = \mathcal{D}(P_i)$. It is convenient to solve the problem in two steps.

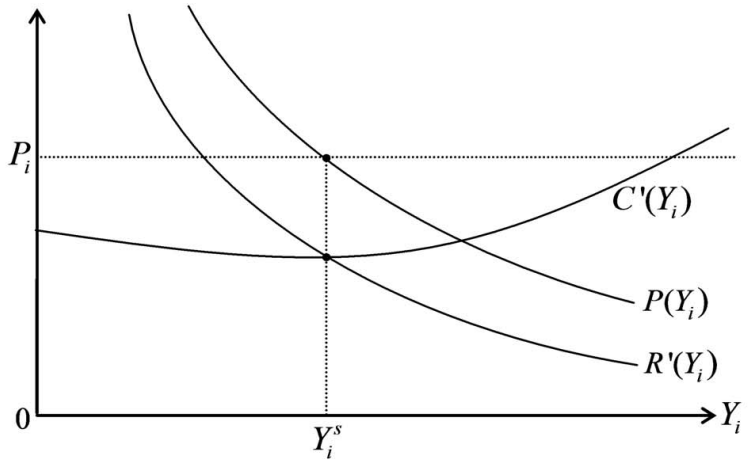

Figure 2.5: Determination of the monopolist price and output.

Step 1. Imagine the monopolist has already chosen the output level Y_i. Then the problem is to minimize cost:

$$\min_{K_i, L_i} \; W_K K_i + W_L L_i \;\text{ s.t. }\; F(K_i, L_i) = Y_i.$$

An interior solution (K_i, L_i) will satisfy the first-order conditions

$$\lambda F_K(K_i, L_i) = W_K, \qquad \lambda F_L(K_i, L_i) = W_L, \tag{2.49}$$

where λ is the Lagrange multiplier. Since F is neoclassical and thereby strictly quasiconcave, the first-order conditions are not only necessary but also sufficient for (K_i, L_i) to be a solution, and (K_i, L_i) will be unique so that we can write these conditional factor demands as functions, $K_i^d = K(W_K, W_L, Y_i)$ and $L_i^d = L(W_K, W_L, Y_i)$. This gives rise to the cost function $\mathcal{C}(Y_i) = W_K K(W_K, W_L, Y_i) + W_L L(W_K, W_L, Y_i)$.

Step 2. Solve

$$\max_{Y_i} \; \Pi(Y_i) = R(Y_i) - \mathcal{C}(Y_i) = \mathcal{P}(Y_i) Y_i - \mathcal{C}(Y_i).$$

[28] Generally, the technology would differ across the different product lines and F should thus be replaced by F^i, but for notational convenience we ignore this.

We have here introduced "total revenue" $R(Y_i) = \mathcal{P}(Y_i)Y_i$, where $\mathcal{P}(Y_i)$ is the inverse demand function defined by $\mathcal{P}(Y_i) \equiv \mathcal{D}^{-1}(Y_i) = [Y_i/(Y/n)]^{-1/\varepsilon} P$ from (2.48). The first-order condition is

$$R'(Y_i) = \mathcal{P}(Y_i) + \mathcal{P}'(Y_i)Y_i = \mathcal{C}'(Y_i), \tag{2.50}$$

where the left-hand side is *marginal revenue* and the right-hand side is *marginal cost*.

A sufficient second-order condition is that $\Pi''(Y_i) = R''(Y_i) - \mathcal{C}''(Y_i) < 0$, i.e., the marginal revenue curve crosses the marginal cost curve from above. In the present case this is surely satisfied if we assume $\mathcal{C}''(Y_i) \geq 0$, which also ensures existence and uniqueness of a solution to (2.50). Substituting this solution, which we denote Y_i^s, cf. Fig. 2.5, into the conditional factor demand functions from Step 1, we find the factor demands, K_i^d and L_i^d. Owing to the downward-sloping demand curves the factor demands are unique whether the technology exhibits DRS, CRS, or IRS. Thus, contrary to the perfect competition case, neither CRS nor IRS pose particular problems.

From the definition $R(Y_i) = P(Y_i)Y_i$ follows

$$R'(Y_i) = P_i\left(1 + \frac{Y_i}{P_i}\mathcal{P}'(Y_i)\right) = P_i\left(1 - \frac{1}{\varepsilon}\right) = P_i\frac{\varepsilon - 1}{\varepsilon}.$$

So the pricing rule is $P_i = (1+\mu)\mathcal{C}'(Y_i)$, where Y_i is the profit maximizing output level and $\mu \equiv \varepsilon/(\varepsilon - 1) - 1 > 0$ is the mark-up on marginal cost. An analytical very convenient feature is thus a *constant* markup.

In parallel with (2.31) and (2.32) the solution to firm i's decision problem is characterized by the *marginal revenue productivity* conditions

$$R'(Y_i^s)F_K(K_i^d, L_i^d) = W_K, \tag{2.51}$$
$$R'(Y_i^s)F_L(K_i^d, L_i^d) = W_L, \tag{2.52}$$

where $Y_i^s = F(K_i^d, L_i^d)$. These conditions follow from (2.49), since the Lagrange multiplier equals marginal cost (see Appendix A), which equals marginal revenue. That is, at profit maximum the marginal revenue products of capital and labor, respectively, equal the corresponding factor prices. Since $P_i > R'(Y_i^s)$, the factor prices are below the value of the marginal productivities. This reflects the market power of the firms.

In macro models a lot of symmetry is often assumed. If there is complete symmetry across product lines and if factor markets clear as in (2.36) and (2.37) with inelastic factor supplies, K^s and L^s, then $K_i^d = K^s/n$ and $L_i^d = L^s/n$. Furthermore, all firms will choose the same price so that $P_i = P$, $i = 1, 2, \ldots, n$.

Then the given factor supplies, together with (2.51) and (2.52), determine the equilibrium *real* factor prices:

$$\begin{aligned} w_K &\equiv \frac{W_K}{P} = \frac{1}{1+\mu} F_K(\frac{K^s}{n}, \frac{L^s}{n}), \\ w_L &\equiv \frac{W_L}{P} = \frac{1}{1+\mu} F_L(\frac{K^s}{n}, \frac{L^s}{n}), \end{aligned}$$

where we have used that $R'(Y_i^s) = P/(1+\mu)$ under these circumstances. As under perfect competition, the real factor prices are proportional to the corresponding marginal productivities, although with a factor of proportionality less than one, namely equal to the inverse of the markup. This observation is sometimes used as a defence for applying the simpler perfect-competition framework for studying certain long-run aspects of the economy. For these aspects, the size of the proportionality factor may be immaterial, at least as long as it is relatively constant over time. Indeed, the constant markups open up for a simple transformation of many of the perfect competition results to monopolistic competition results by inserting the markup factor $1+\mu$ the relevant places in the formulas.

If in the short term only labor is a variable production factor, then (2.51) need not hold. As claimed by Keynesian and New Keynesian thinking, also the prices chosen by the firms may be more or less fixed in the short run because the firms face price adjustment costs ("menu costs") and are reluctant to change prices too often, at least vis-a-vis changes in demand. Then in the short run only the produced quantity will adjust to changes in demand. As long as the output level is within the range where marginal cost is below the price, such adjustments are still beneficial to the firm. As a result, even (2.52) may at most hold "on average" over the business cycle. These matters are dealt with in Part V of this book.

In practice, market power and other market imperfections also play a role in the factor markets, implying that further complicating elements enter the picture. One of the tasks of theoretical and empirical macroeconomics is to clarify the aggregate implications of market imperfections and sort out which market imperfections are quantitatively important in different contexts.

2.5.4 The financing of firms' operations

We have so far talked about aspects related to production and pricing. What about the *financing* of a firm's operations? To acquire not only its fixed capital (structures and machines) but also its raw material and other intermediate inputs, a firm needs *funds* (there are expenses before the proceeds from sale arrive). These funds ultimately come from the accumulated saving of households. In long-run

macromodels to be considered in the next chapters, uncertainty as well as non-neutrality of corporate taxation are ignored; in that context the capital structure (the debt-equity ratio) of firms is indeterminate and irrelevant for production outcomes.[29] In those chapters we shall therefore concentrate on the latter. Later chapters, dealing with short- and medium-run issues, touch upon cases where capital structure and bankruptcy risk matter and financial intermediaries enter the scene.

2.6 Literature notes

As to the question of the empirical validity of the constant returns to scale assumption, Malinvaud (1998) offers an account of the econometric difficulties associated with estimating production functions. Studies by Basu (1996) and Basu and Fernald (1997) suggest returns to scale are about constant or decreasing. Studies by Hall (1990), Caballero and Lyons (1992), Harris and Lau (1992), Antweiler and Treffler (2002), and Harrison (2003) suggest there are quantitatively significant increasing returns, either internal or external. On this background it is not surprising that the case of IRS (at least at industry level), together with market forms different from perfect competition, has in recent years received more attention in macroeconomics and in the theory of economic growth.

Macroeconomists' use of the value-laden term "technological progress" in connection with technological change may seem suspect. But the term should be interpreted as merely a label for certain types of shifts of isoquants in an abstract universe. At a more concrete and disaggregate level analysts of course make use of more refined notions about technological change, recognizing not only benefits of new technologies, but for instance also the risks, including risk of fundamental mistakes (think of the introduction and later abandonment of asbestos in the construction industry). For history of technology see, e.g., Ruttan (2001) and Smil (2003).

When referring to a Cobb-Douglas (or CES) production function some authors implicitly assume that the partial output elasticities w.r.t. inputs time-independent and thereby independent of technological change. For the case where the inputs in question are renewable and nonrenewable natural resources, Growiec and Schumacher (2008) study cases of time-dependency of the partial output elasticities.

When technical change is not "neutral" in one of the senses described, it may be systematically "biased" in alternative "directions". The reader is referred to the specialized literature on economic growth, cf. literature notes to Chapter 1.

[29] In chapter 14 we return to this irrelevance proposition, called the Modigliani-Miller theorem.

Embodied technological progress, sometimes called investment-specific technological progress, is explored in, for instance, Solow (1960), Greenwood et al. (1997), and Groth and Wendner (2014).

Time series for different countries' aggregate and to some extent sectorial capital stocks are available from Penn World Table, ..., EU KLEMS, ..., and the AMECO database, .

The concept of Gorman preferences and conditions ensuring that a representative household is admitted are surveyed in Acemoglu (2009). Another source, also concerning the conditions for the representative firm to be a meaningful notion, is Mas-Colell et al. (1995). For general discussions of the limitations of representative agent approaches, see Kirman (1992) and Gallegati and Kirman (1999). Reviews of the "Cambridge Controversy" are contained in Mas-Colell (1989) and Felipe and Fisher (2003). The last-mentioned authors find the conditions required for the well-behavedness of these constructs so stringent that it is difficult to believe that actual economies are in any sense close to satisfy them. For less distrustful views and constructive approaches to the issues, see for instance Johansen (1972), Malinvaud (1998), Jorgenson et al. (2005), and Jones (2005).

Scarf (1960) provided a series of examples of lack of dynamic stability of an equilibrium price vector in an exchange economy. Mas-Colell et al. (1995) survey the later theoretical development in this field.

The counterexample to guaranteed stability of the neoclassical factor market equilibrium presented towards the end of Section 2.4 is taken from Bliss (1975), where further perspectives are discussed. It may be argued that this kind of stability questions should be studied on the basis of adjustment processes of a less mechanical nature than a Walrasian tâtonnement process. The view would be that trade out of equilibrium should be incorporated in the analysis and agents' behavior out of equilibrium should be founded on some kind of optimization or "satisficing", incorporating adjustment costs and imperfect information. The field is complicated and the theory not settled. Yet it seems fair to say that the studies of adjustment processes out of equilibrium indicate that the equilibrating force of Adam Smith's invisible hand is not without its limits. See Fisher (1983), Osborne and Rubinstein (1990), and Negishi (2008) for reviews and elaborate discussion of these issues.

We introduced the assumption that physical capital depreciation can be described as geometric (in continuous time exponential) evaporation of the capital stock. This formula is popular in macroeconomics, more so because of its simplicity than its realism. An introduction to more general approaches to depreciation is contained in, e.g., Nickell (1978).

2.7 Appendix

A. Strict quasiconcavity

Consider a function $f : \mathcal{A} \to \mathbb{R}$, where $\mathcal{A}$ is a convex set, $\mathcal{A} \subseteq \mathbb{R}^n$.[30] Given a real number a, if $f(x) = a$, the *upper contour set* is defined as $\{x \in \mathcal{A} | \, f(x) \geq a\}$ (the set of input bundles that can produce at least the amount a of output). The function $f(x)$ is called *quasiconcave* if its upper contour sets, for any constant a, are convex sets. If all these sets are strictly convex, $f(x)$ is called *strictly quasiconcave*.

Average and marginal costs To show that (2.14) holds with n production inputs, $n = 1, 2, \ldots$, we derive the cost function of a firm with a neoclassical production function, $Y = F(X_1, X_2, \ldots, X_n)$. Given a vector of strictly positive input prices $\mathbf{w} = (w_1, \ldots, w_n) >> 0$, the firm faces the problem of finding a cost-minimizing way to produce a given positive output level $\bar{Y}$ within the range of F. The problem is

$$\min \sum_{i=1}^{n} w_i X_i \text{ s.t. } F(X_1, \ldots, X_n) = \bar{Y} \text{ and } X_i \geq 0, \; i = 1, 2, \ldots, n.$$

An interior solution, $\mathbf{X}^* = (X_1^*, \ldots, X_n^*)$, to this problem satisfies the first-order conditions $\lambda F_i'(\mathbf{X}^*) = w_i$, where λ is the Lagrange multiplier, $i = 1, \ldots, n$.[31] Since F is neoclassical and thereby strictly quasiconcave in the interior of $\mathbb{R}^n_+$, the first-order conditions are not only necessary but also sufficient for the vector $\mathbf{X}^*$ to be a solution, and $\mathbf{X}^*$ will be unique[32] so that we can write it as a function, $\mathbf{X}^*(\bar{Y}) = (X_1^*(\bar{Y}), \ldots, X_n^*(\bar{Y}))$. This gives rise to the *cost function* $\mathcal{C}(\bar{Y}) = \sum_{i=1}^n w_i X_i^*(\bar{Y})$. So *average cost* is $\mathcal{C}(\bar{Y})/\bar{Y}$. We find *marginal cost* to be

$$\mathcal{C}'(\bar{Y}) = \sum_{i=1}^{n} w_i X_i^{*\prime}(\bar{Y}) = \lambda \sum_{i=1}^{n} F_i'(\mathbf{X}^*) X_i^{*\prime}(\bar{Y}) = \lambda,$$

where the third equality comes from the first-order conditions, and the last equality is due to the constraint $F(\mathbf{X}^*(\bar{Y})) = \bar{Y}$, which, by taking the total derivative on both sides, gives $\sum_{i=1}^n F_i'(\mathbf{X}^*) X_i^{*\prime}(\bar{Y}) = 1$. Consequently, the ratio of average to marginal costs is

$$\frac{\mathcal{C}(\bar{Y})/\bar{Y}}{\mathcal{C}'(\bar{Y})} = \frac{\sum_{i=1}^n w_i X_i^*(\bar{Y})}{\lambda \bar{Y}} = \frac{\sum_{i=1}^n F_i'(\mathbf{X}^*) X_i^*(\bar{Y})}{F(\mathbf{X}^*)},$$

[30] Recall that a set S is said to be *convex* if $x, y \in S$ and $\lambda \in [0, 1]$ implies $\lambda x + (1 - \lambda) y \in S$.

[31] Since in this section we use a bit of vector notation, we exceptionally mark first-order partial derivatives by a prime in order to clearly distinguish from the elements of a vector (so we write F_i' instead of our usual F_i).

[32] See Sydsaeter et al. (2008), pp. 74, 75, and 125.

which in analogy with (2.13) is the elasticity of scale at the point $\mathbf{X}^*$. This proves (2.14).

Sufficient conditions for strict quasiconcavity The claim (iii) in Section 2.1.3 was that a continuously differentiable two-factor production function $F(K, L)$ with CRS, satisfying $F_K > 0, F_L > 0$, and $F_{KK} < 0, F_{LL} < 0$, will automatically also be strictly quasi-concave in the interior of $\mathbb{R}^2$ and thus neoclassical.

To prove this, consider a function of two variables, $z = f(x,\ y)$, that is twice continuously differentiable with $f_1 \equiv \partial z/\partial x > 0$ and $f_2 \equiv \partial z/\partial y > 0$, everywhere. Then the equation $f(x,\ y) = a$, where a is a constant, defines an isoquant, $y = g(x)$, with slope $g'(x) = -f_1(x, y)/f_2(x, y)$. Substitute $g(x)$ for y in this equation and take the derivative w.r.t. x. By straightforward calculation we find

$$g''(x) = -\frac{f_1^2 f_{22} - 2f_1 f_2 f_{21} + f_2^2 f_{11}}{f_2^3} \tag{2.53}$$

If the numerator is negative, then $g''(x) > 0$; that is, the isoquant is strictly convex to the origin. And if this holds for all $(x,\ y)$, then f is strictly quasi-concave in the interior of $\mathbb{R}^2$. A sufficient condition for a negative numerator is that $f_{11} < 0$, $f_{22} < 0$ and $f_{21} \geq 0$. All these conditions, including the last three are satisfied by the given function F. Indeed, F_K, F_L, F_{KK}, and F_{LL} have the required signs. And when F has CRS, F is homogeneous of degree 1 and thereby $F_{KL} > 0$, see Appendix B. Hereby claim (iii) in Section 2.1.3 is proved.

B. Homogeneous production functions

The claim (iv) in Section 2.1.3 was that a two-factor production function with CRS, satisfying $F_K > 0, F_L > 0$, and $F_{KK} < 0, F_{LL} < 0$, has always $F_{KL} > 0$, i.e., there is *direct complementarity* between K and L. This assertion is implied by the following observations on homogeneous functions.

Let $Y = F(K,\ L)$ be a twice continuously differentiable production function with $F_K > 0$ and $F_L > 0$ everywhere. Assume F is homogeneous of degree $h > 0$, that is, for all possible (K, L) and all $\lambda > 0$, $F(\lambda K, \lambda L) = \lambda^h F(K, L)$. According to Euler's theorem (see Math Tools) we then have:

CLAIM 1 For all $(K,\ L)$, where $K > 0$ and $L > 0$,

$$KF_K(K, L) + LF_L(K, L) = hF(K, L). \tag{2.54}$$

Euler's theorem also implies the inverse:

CLAIM 2 If (2.54) is satisfied for all $(K,\ L)$, where $K > 0$ and $L > 0$, then $F(K, L)$ is homogeneous of degree h.

Partial differentiation w.r.t. K and L, respectively, gives, after ordering,

$$KF_{KK} + LF_{LK} = (h-1)F_K \tag{2.55}$$
$$KF_{KL} + LF_{LL} = (h-1)F_L. \tag{2.56}$$

In (2.55) we can substitute $F_{LK} = F_{KL}$ (by Young's theorem). In view of Claim 2 this shows:

CLAIM 3 The marginal products, F_K and F_L, considered as functions of K and L, are homogeneous of degree $h-1$.

We see also that when $h \geq 1$ and K and L are positive, then

$$F_{KK} < 0 \text{ implies } F_{KL} > 0, \tag{2.57}$$
$$F_{LL} < 0 \text{ implies } F_{KL} > 0. \tag{2.58}$$

For $h = 1$ this establishes the direct complementarity result, (iv) in Section 2.1.3, to be proved. A by-product of the derivation is that also when a neoclassical production function is homogeneous of degree $h > 1$ (which implies IRS), does direct complementarity between K and L hold.

Remark. The terminology around complementarity and substitutability may easily lead to confusion. In spite of K and L exhibiting *direct complementarity* when $F_{KL} > 0$, K and L are still *substitutes* in the sense that cost minimization for a given output level implies that a rise in the price of one factor results in higher demand for the other factor.

The claim (v) in Section 2.1.3 was the following. Suppose we face a CRS production function, $Y = F(K, L)$, that has positive marginal products, F_K and F_L, everywhere and isoquants, $K = g(L)$, satisfying the condition $g''(L) > 0$ everywhere (i.e., F is strictly quasi-concave). Then the partial second derivatives must satisfy the neoclassical conditions:

$$F_{KK} < 0, F_{LL} < 0. \tag{2.59}$$

The proof is as follows. The first inequality in (2.59) follows from (2.53) combined with (2.55). Indeed, for $h = 1$, (2.55) and (2.56) imply $F_{KK} = -F_{LK}L/K = -F_{KL}L/K$ and $F_{KL} = -F_{LL}L/K$, i.e., $F_{KK} = F_{LL}(L/K)^2$ (or, in the notation of Appendix A, $f_{22} = f_{11}(x/y)^2$), which combined with (2.53) gives the conclusion $F_{KK} < 0$, when $g'' > 0$. The second inequality in (2.59) can be verified in a similar way.

Note also that for $h = 1$ the equations (2.55) and (2.56) entail

$$KF_{KK} = -LF_{LK} \text{ and } KF_{KL} = -LF_{LL}, \tag{2.60}$$

respectively. By dividing the left- and right-hand sides of the first of these equations with those of the second we conclude that $F_{KK}F_{LL} = F_{KL}^2$ in the CRS case. We see also from (2.60) that, under CRS, the implications in (2.57) and (2.58) can be turned round.

Finally, we asserted in § 2.1.1 that when the neoclassical production function $Y = F(K, L)$ is homogeneous of degree h, then the marginal rate of substitution between the production factors depends only on the factor proportion $k \equiv K/L$. Indeed,

$$MRS_{KL}(K, L) = \frac{F_L(K, L)}{F_K(K, L)} = \frac{L^{h-1}F_L(k, 1)}{L^{h-1}F_K(k, 1)} = \frac{F_L(k, 1)}{F_K(k, 1)} \equiv mrs(k), \qquad (2.61)$$

where $k \equiv K/L$. The result (2.61) follows even if we only assume $F(K, L)$ is *homothetic*. When $F(K, L)$ is homothetic, by definition we can write $F(K, L) \equiv \varphi(G(K, L))$, where G is homogeneous of degree 1 and φ is an increasing function. In view of this, we get

$$MRS_{KL}(K, L) = \frac{\varphi' G_L(K, L)}{\varphi' G_K(K, L)} = \frac{G_L(k, 1)}{G_K(k, 1)},$$

where the last equality is implied by Claim 3 for $h = 1$.

C. The Inada conditions combined with CRS

We consider a neoclassical production function, $Y = F(K, L)$, exhibiting CRS. Defining $k \equiv K/L$, we can then write $Y = LF(k, 1) \equiv Lf(k)$, where $f(0) \geq 0$, $f' > 0$, and $f'' < 0$.

Essential inputs In Section 2.1.2 we claimed that the upper Inada condition for MPL together with CRS implies that without capital there will be no output:

$$F(0, L) = 0 \qquad \text{for any } L > 0.$$

In other words: in this case capital is an essential input. To prove this claim, let $K > 0$ be fixed and let $L \to \infty$. Then $k \to 0$, implying, by (2.16) and (2.18), that $F_L(K, L) = f(k) - f'(k)k \to f(0)$. But from the upper Inada condition for MPL we also have that $L \to \infty$ implies $F_L(K, L) \to 0$. It follows that

$$\text{the upper Inada condition for } MPL \text{ implies } f(0) = 0. \qquad (2.62)$$

Since under CRS, for any $L > 0$, $F(0, L) = LF(0, 1) \equiv Lf(0)$, we have hereby shown our claim.

Similarly, we can show that the upper Inada condition for MPK together with CRS implies that labor is an essential input. Consider the output-capital ratio $x \equiv Y/K$. When F has CRS, we get $x = F(1, \ell) \equiv g(\ell)$, where $\ell \equiv L/K$, $g' > 0$, and $g'' < 0$. Thus, by symmetry with the previous argument, we find that under CRS, the upper Inada condition for MPK implies $g(0) = 0$. Since under CRS $F(K, 0) = KF(1, 0) \equiv Kg(0)$, we conclude that the upper Inada condition for MPK together with CRS implies

$$F(K, 0) = 0 \qquad \text{for any } K > 0,$$

that is, without labor, no output.

Sufficient conditions for output going to infinity when either input goes to infinity Here our first claim is that when F exhibits CRS and satisfies the upper Inada condition for MPL and the lower Inada condition for MPK, then

$$\lim_{L\to\infty} F(K, L) = \infty \qquad \text{for any } K > 0.$$

To prove this, note that Y can be written $Y = Kf(k)/k$, since $K/k = L$. Here,

$$\lim_{k\to 0} f(k) = f(0) = 0,$$

by continuity and (2.62), presupposing the upper Inada condition for MPL. Thus, for any given $K > 0$,

$$\lim_{L\to\infty} F(K, L) = K \lim_{L\to\infty} \frac{f(k)}{k} = K \lim_{k\to 0} \frac{f(k) - f(0)}{k} = K \lim_{k\to 0} f'(k) = \infty,$$

by the lower Inada condition for MPK. This verifies the claim.

Our second claim is symmetric with this and says: when F exhibits CRS and satisfies the upper Inada condition for MPK and the lower Inada condition for MPL, then

$$\lim_{K\to\infty} F(K, L) = \infty \qquad \text{for any } L > 0.$$

The proof is analogue. So, in combination, the four Inada conditions imply, under CRS, that output has no upper bound when either input goes to infinity.

D. Concave neoclassical production functions

Two claims made in Section 2.4 are proved here.

CLAIM 1 When a neoclassical production function $F(K, L)$ is concave, it has non-increasing returns to scale everywhere.

Proof. We consider a concave neoclassical production function, F. Let $\mathbf{x} = (x_1, x_2) = (K, L)$. Then we can write $F(K, L)$ as $F(\mathbf{x})$. By concavity, for all pairs $\mathbf{x}^0, \mathbf{x} \in \mathbb{R}^2_+$, we have $F(\mathbf{x}^0) - F(\mathbf{x}) \leq \sum_{i=1}^2 F_i'(\mathbf{x})(x_i^0 - x_i)$. In particular, for $\mathbf{x}^0 = (0, 0)$, since $F(\mathbf{x}^0) = F(0, 0) = 0$, we have

$$-F(\mathbf{x}) \leq -\sum_{i=1}^{2} F_i'(\mathbf{x})x_i. \tag{2.63}$$

Suppose $\mathbf{x} \in \mathbb{R}^2_{++}$. Then $F(\mathbf{x}) > \mathbf{0}$ in view of F being neoclassical so that $F_K > 0$ and $F_L > 0$. From (2.63) we now find the elasticity of scale to be

$$\sum_{i=1}^{2} F_i'(\mathbf{x})x_i/F(\mathbf{x}) \leq 1. \tag{2.64}$$

In view of (2.13) and (2.12), this implies non-increasing returns to scale everywhere. □

CLAIM 2 When a neoclassical production function $F(K, L)$ is strictly concave, it has decreasing returns to scale everywhere.

Proof. The argument is analogue to that above, but in view of strict concavity the inequalities in (2.63) and (2.64) become strict. This implies that F has DRS everywhere. □

2.8 Exercises

2.1

Part II

LOOKING AT THE LONG RUN

Chapter 3

The basic OLG model: Diamond

There exists two main analytical frameworks for analyzing the basic intertemporal choice, consumption versus saving, and the dynamic long-run implications of this choice: *overlapping generations* models and *representative agent* models. In the first class of models the focus is on (a) the interaction between different generations alive at the same time, and (b) the never-ending entrance of new generations. In the second class of models the household sector is modelled as consisting of a finite number of infinitely-lived agents. One interpretation is that these agents are dynasties where parents take the utility of their descendants fully into account by leaving bequests. This approach, which is also called the Ramsey approach (after the British mathematician and economist Frank Ramsey, 1903-1930), will be described in Chapter 8 (discrete time) and Chapter 10 (continuous time).

In the present chapter we introduce the overlapping generations approach which has shown its usefulness for analysis of questions associated with public debt problems, taxation of capital income, financing of social security (pensions), design of educational systems, non-neutrality of money, and the possibility of speculative bubbles. Our focus will be on the overlapping generations model called Diamond's OLG model[1] after the American economist and Nobel Prize laureate Peter A. Diamond (1940-).

Among the strengths of the model are:

- The *life-cycle* aspect of human behavior is taken into account. Although the economy is infinitely-lived, the individual agents have finite time horizons. During lifetime one's educational level, working capacity, income, and needs change and this is reflected in the individual labor supply and saving behavior. The aggregate implications of the life-cycle behavior of coexisting individual agents at different stages in their life is at the centre of attention.

[1] Diamond (1965).

- The model takes elementary forms of *heterogeneity* in the population into account – there are "old" and there are "young", there are currently-alive people and there are as yet unborn whose preferences are not reflected in current market transactions. Questions relating to the distribution of income and wealth across generations can be studied. For example, how does the investment in capital and environmental protection by current generations affect the conditions for succeeding generations?

3.1 Motives for saving

Before going into the specifics of Diamond's model, let us briefly consider what may motivate people to save:

(a) The *consumption-smoothing motive for saving*. Individuals go through a life cycle where individual income typically has a hump-shaped time pattern; by saving and dissaving the individual attempts to obtain the desired smoothing of consumption across lifetime. This is the essence of the *life-cycle saving hypothesis* put forward by Nobel laureate Franco Modigliani (1918-2003) and associates in the 1950s. This hypothesis states that consumers plan their saving and dissaving in accordance with anticipated variations in income and needs over lifetime. Because needs vary less over lifetime than income, the time profile of saving tends to be hump-shaped with some dissaving early in life (while studying etc.), positive saving during the years of peak earnings and then dissaving after retirement.

(b) The *precautionary motive for saving*. Income as well as needs may vary due to conditions of *uncertainty*: sudden unemployment, illness, or other kinds of bad luck. By saving, the individual can obtain a buffer against such unwelcome events.

Horioka and Watanabe (1997) find that empirically, the saving motives (a) and (b) are of dominant importance (Japanese data). Yet other motives include:

(c) Saving enables the purchase of *durable consumption goods* and owner-occupied housing as well as repayment of debt.

(d) Saving may be motivated by the *desire to leave bequests* to heirs.

(e) Saving may simply be motivated by the fact that financial wealth may lead to *social prestige* and economic or political *power*.

Diamond's OLG model aims at simplicity and concentrates on motive (a). Only one aspect of motive (a) is in fact considered, namely the saving for retirement. People live for two periods only, as "young", working full-time, and as "old", having retired and living by their savings. The Diamond model abstracts from a possible bequest motive.

Now to the details.

3.2 The model framework

The flow of time is divided into successive periods of equal length, taken as the time unit. Given the two-period lifetime of (adult) individuals, the period length is understood to be around, say, 30 years. The main assumptions are:

1. The number of young people in period t, denoted L_t, changes over time according to $L_t = L_0(1+n)^t$, $t = 0, 1, 2, ...$, where n is a constant, $n > -1$. Indivisibility is ignored and so L_t is just considered a positive real number.

2. Only the young work. Each young supplies one unit of labor inelastically. The division of available time between work and leisure is thereby considered as exogenous.

3. Output is homogeneous and can be used for consumption as well as investment in physical capital. Physical capital is the only non-human asset in the economy; it is owned by the old and rented out to the firms. Output is the numeraire (unit of account) used in trading. Money (means of payment) is ignored.[2]

4. The economy is closed (no foreign trade).

5. Firms' technology has constant returns to scale.

6. In each period three markets are open, a market for output, a market for labor services, and a market for capital services. Perfect competition rules in all markets. Uncertainty is absent; when a decision is made, its consequences are known.

7. Agents have perfect foresight.

Assumption 7 entails the following. First, the agents are assumed to have "rational expectations" or, with a better name, "model-consistent expectations".

[2] As to the disregard of money we may imagine that agents have safe electronic accounts in a fictional central bank allowing costless transfers between accounts.

This means that forecasts made by the agents coincide with the forecasts that can be calculated on the basis of the model. Second, as there are no stochastic elements in the model (no uncertainty), the forecasts are point estimates rather than probabilistic forecasts. Thereby the model-consistent expectations take the extreme form of *perfect foresight*: the agents agree in their expectations about the future evolution of the economy and these expectations are point estimates that coincide with the subsequent actual evolution of the economy.

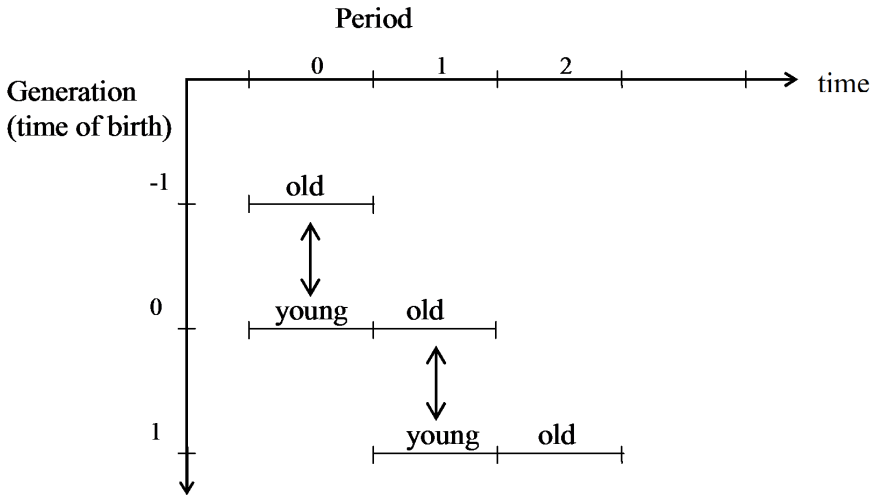

Figure 3.1: The two-period model's time structure.

Of course, this is an unrealistic assumption. The model makes this assumption in order to simplify in a first approach. The results that emerge will be the outcome of economic mechanisms in isolation from expectational errors. In this sense the model constitutes a "pure" case (benchmark case).

The time structure of the model is illustrated in Fig. 3.1. In every period two generations are alive and interact with each other as indicated by the arrows. The young supply labor to the firms, earn a labor income part of which they consume and part of which they save for retirement. The young thereby offset the dissaving by the old and possibly bring about positive net investment in the economy. At the end of the period the savings by the young is converted into direct ownership of new capital goods which constitute the non-consumed part of aggregate output plus capital goods left over from the previous period. In the next period the now old owners of the capital goods rent them out to the firms. We may imagine that the firms are owned by the old, but this ownership is not visible in the equilibrium allocation because pure profits will be nil due to the combination of perfect competition and constant returns to scale.

Let the output good be the numeraire and let $\hat{r}_t$ denote the rental rate for capital in period t; that is, $\hat{r}_t$ is the real price a firm has to pay at the end of period t for the right to use one unit of someone else's physical capital through period t. So the owner of K_t units of physical capital receives a

$$\text{real (net) rate of return on capital } = \frac{\hat{r}_t K_t - \delta K_t}{K_t} = \hat{r}_t - \delta, \tag{3.1}$$

where δ is the rate of physical capital depreciation which is assumed constant, $0 \leq \delta \leq 1$.

Suppose there is also a market for loans, the "credit market". Assume you have lent out one unit of output from the end of period $t-1$ to the end of period t. If the *real interest rate* in the loan market is r_t, then, at the end of period t you should get back $1+r_t$ units of output. In the absence of uncertainty, equilibrium requires that capital and loans give the same rate of return,

$$\hat{r}_t - \delta = r_t. \tag{3.2}$$

This *no-arbitrage* condition indicates how the rental rate for capital and the more everyday concept, the interest rate, would be related in an equilibrium where both the market for capital services and a credit market were active. We shall see, however, that in this model no credit market will be active in an equilibrium. Nevertheless we will follow the tradition and call the right-hand side of (3.2) the *interest rate*.

Table 3.1 provides an overview of the notation. As to our timing convention, notice that any stock variable dated t indicates the amount held at the beginning of period t. That is, the capital stock accumulated by the end of period $t-1$ and available for production in period t is denoted K_t. We therefore write $K_t = (1-\delta)K_{t-1} + I_{t-1}$ and $Y_t = F(K_t, L_t)$, where F is an aggregate production function. In this context it is useful to think of "period t" as running from date t to date $t+1$. So period t is the time interval $[t, t+1)$ on a continuous time axis. Still, all decisions are made at discrete points in time $t = 0, 1, 2, ...$ ("dates"). We imagine that receipts for work and lending as well as payment for the consumption in period t occur at the end of the period. These timing conventions are common in discrete-time growth and business cycle theory;[3] they are convenient because they make switching between discrete and continuous time analysis fairly easy.

[3]In contrast, in *accounting* and *finance* literature, typically K_t would denote the *end-of-period-t* stock that begins to yield its services *next* period.

Table 3.1. List of main variable symbols

Symbol	*Meaning*
L_t	the number of young people in period t
n	generation growth rate
K_t	aggregate capital available in period t
c_{1t}	consumption as young in period t
c_{2t}	consumption as old in period t
w_t	real wage in period t
r_t	real interest rate (from end of per. $t-1$ to end of per. t)
ρ	rate of time preference (impatience)
θ	elasticity of marginal utility
s_t	saving of each young in period t
Y_t	aggregate output in period t
$C_t = c_{1t}L_t + c_{2t}L_{t-1}$	aggregate consumption in period t
$S_t = Y_t - C_t$	aggregate gross saving in period t
$\delta \in [0,1]$	capital depreciation rate
$K_{t+1} - K_t = I_t - \delta K_t$	aggregate net investment in period t

3.3 The saving by the young

Suppose the preferences of the young can be represented by the lifetime utility function specified in (3.3). Given w_t and r_{t+1}, the decision problem of the young in period t then is:

$$\max_{c_{1t},c_{2t+1}} U(c_{1t},c_{2t+1}) = u(c_{1t}) + (1+\rho)^{-1}u(c_{2t+1}) \quad \text{s.t.} \tag{3.3}$$

$$c_{1t} + s_t = w_t \qquad (w_t > 0), \tag{3.4}$$

$$c_{2t+1} = (1+r_{t+1})s_t \qquad (r_{t+1} > -1), \tag{3.5}$$

$$c_{1t} \geq 0, c_{2t+1} \geq 0. \tag{3.6}$$

The interpretation of the variables is given in Table 3.1 above. We may think of the "young" as a household consisting of one adult and $1+n$ children whose consumption is included in c_{1t}. Note that "utility" appears at two levels. There is a *lifetime utility function*, U, and a *period utility function*, u.[4] The latter is assumed to be the same in both periods of life (this has no effects on the qualitative results and simplifies the exposition). The period utility function is assumed continuous and twice continuously differentiable with $u' > 0$ and $u'' < 0$ (positive, but diminishing marginal utility of consumption). Many popular specifications

[4]Other names for these two functions are the *intertemporal utility function* and the *subutility function*, respectively.

of u, e.g., $u(c) = \ln c$, have the property that $\lim_{c\to 0} u(c) = -\infty$; then we *define* $u(0) = -\infty$.

The parameter ρ is called the *rate of time preference.* It acts as a utility discount *rate,* whereas $(1+\rho)^{-1}$ is a utility discount *factor.* Thus ρ indicates the degree of *impatience* w.r.t. the "arrival" of utility. By definition, $\rho > -1$, but $\rho > 0$ is often assumed. When preferences can be represented in this additive way, they are called *time-separable.* In principle, as seen from period t the interest rate appearing in (3.5) should be interpreted as an *expected* real interest rate. But as long as we assume perfect foresight, there is no need to distinguish between actual and expected magnitudes.

Box 3.1. Discount rates and discount factors

By a *discount rate* is meant an interest rate applied in the construction of a discount factor. A *discount factor* is a factor by which future benefits or costs, measured in some unit of account, are converted into present equivalents. The higher the discount rate the lower the discount factor.

One should bear in mind that a discount rate depends on what is to be discounted. In (3.3) the unit of account is "utility" and ρ acts as a *utility discount rate.* In (3.7) the unit of account is the consumption good and r_{t+1} acts as a *consumption discount rate.* If people also work as old, the right-hand side of (3.7) would read $w_t + (1+r_{t+1})^{-1}w_{t+1}$ and thus r_{t+1} would act as an *earnings discount rate.* This will be the same as the consumption discount rate if we think of real income measured in consumption units. But if we think of nominal income, that is, income measured in monetary units, there would be a *nominal earnings discount rate*, namely the *nominal* interest rate, which in an economy with inflation will exceed the consumption discount rate. Unfortunately, confusion of different discount rates is not rare.

In (3.5) the interest rate r_{t+1} acts as a (net) rate of return on saving.[5] An interest rate may also be seen as a discount rate relating to consumption over time. Indeed, by isolating s_t in (3.5) and substituting into (3.4), we may consolidate

[5]While s_t in (3.4) appears as a *flow* (non-consumed income), in (3.5) s_t appears as a *stock* (the accumulated financial wealth at the end of period t). This notation is legitimate because the magnitude of the two is the same when the time unit is the same as the period length.

In real life the gross payoff of individual saving may sometimes be nil (if invested in a project that completely failed). Unless otherwise indicated, it is in this book understood that an interest rate is a number exceeding -1 as indicated in (3.5). Thereby the discount factor $1/(1+r_{t+1})$ is well-defined. In general equilibrium, the condition $1 + r_{t+1} > 0$ is always met in the present model.

the two period budget constraints of the individual into *one* budget constraint,

$$c_{1t} + \frac{1}{1+r_{t+1}} c_{2t+1} = w_t. \tag{3.7}$$

In this *intertemporal budget constraint* the interest rate appears as the discount rate entering the discount factor converting future amounts of consumption into present equivalents, cf. Box 3.1.

Solving the saving problem

To avoid the possibility of corner solutions, we impose the No Fast Assumption

$$\lim_{c\to 0} u'(c) = \infty. \tag{A1}$$

In view of the sizeable period length in the model, this is definitely plausible.

Inserting the two budget constraints into the objective function in (3.3), we get $U(c_{1t}, c_{2t+1}) = u(w_t - s_t) + (1+\rho)^{-1} u((1+r_{t+1})s_t) \equiv \tilde{U}_t(s_t)$, a function of only one decision variable, s_t. According to the non-negativity constraint on consumption in both periods, (3.6), s_t must satisfy $0 \leq s_t \leq w_t$. Maximizing w.r.t. s_t gives the first-order condition

$$\frac{d\tilde{U}_t}{ds_t} = -u'(w_t - s_t) + (1+\rho)^{-1} u'((1+r_{t+1})s_t)(1+r_{t+1}) = 0. \tag{FOC}$$

The second derivative of $\tilde{U}_t$ is

$$\frac{d^2\tilde{U}_t}{ds_t^2} = u''(w_t - s_t) + (1+\rho)^{-1} u''((1+r_{t+1})s_t)(1+r_{t+1})^2 < 0. \tag{SOC}$$

Hence there can at most be one s_t satisfying (FOC). Moreover, for a positive wage income there always exists such an s_t. Indeed:

LEMMA 1 Let $w_t > 0$ and suppose the No Fast Assumption (A1) applies. Then the saving problem of the young has a unique solution $s_t = s(w_t, r_{t+1})$. The solution is interior, i.e., $0 < s_t < w_t$, and s_t satisfies (FOC).

Proof. Assume (A1). For any $s \in (0, w_t)$, $d\tilde{U}_t(s)/ds > -\infty$. Now consider the endpoints $s = 0$ and $s = w_t$. By (FOC) and (A1),

$$\begin{aligned}
\lim_{s\to 0} \frac{d\tilde{U}_t}{ds} &= -u'(w_t) + (1+\rho)^{-1}(1+r_{t+1}) \lim_{s\to 0} u'((1+r_{t+1})s) = \infty, \\
\lim_{s\to w} \frac{d\tilde{U}_t}{ds} &= -\lim_{s\to w_t} u'(w_t - s) + (1+\rho)^{-1}(1+r_{t+1}) u'((1+r_{t+1})w_t) = -\infty.
\end{aligned}$$

By continuity of $\tilde{U}_t$, it follows that there exists an $s_t \in (0, w_t)$ such that at $s = s_t$, $d\tilde{U}_t/ds = 0$; This is an application of the *intermediate value theorem.* It follows that (FOC) holds for this s_t. By (SOC), s_t is unique and can therefore be written as an implicit function, $s(w_t, r_{t+1})$, of the exogenous variables in the problem, w_t and r_{t+1}. □

Inserting the solution for s_t into the two period budget constraints, (3.4) and (3.5), immediately gives the optimal consumption levels, c_{1t} and c_{2t+1}.

The simple optimization method we have used here is called the *substitution method*: by substitution of the constraints into the objective function an unconstrained maximization problem is obtained.[6]

The consumption Euler equation

The first-order condition (FOC) can conveniently be written

$$u'(c_{1t}) = (1+\rho)^{-1}u'(c_{2t+1})(1+r_{t+1}). \tag{3.8}$$

This is known as an *Euler equation,* after the Swiss mathematician L. Euler (1707-1783) who was the first to study dynamic optimization problems. In the present context the condition is called a *consumption Euler equation.*

Intuitively, in an optimal plan the marginal utility cost of saving must equal the marginal utility benefit obtained by saving. The marginal utility cost of saving is the opportunity cost (in terms of current utility) of saving one more unit of account in the current period (approximately). This one unit of account is transferred to the next period with interest so as to result in $1 + r_{t+1}$ units of account in that period. An optimal plan requires that the utility cost equals the utility benefit of having r_{t+1} more units of account in the next period. And this utility benefit is the discounted value of the extra utility that can be obtained next period through the increase in consumption by r_{t+1} units.

It may seem odd to attempt an intuitive interpretation this way, that is, in terms of "utility units". The utility concept is just a convenient mathematical device used to represent the assumed *preferences.* Our interpretation is only meant as an as-if interpretation: as if utility were something concrete. An interpretation in terms of concrete *measurable quantities* goes like this. We rewrite (3.8) as

$$\frac{u'(c_{1t})}{(1+\rho)^{-1}u'(c_{2t+1})} = 1 + r_{t+1}. \tag{3.9}$$

The left-hand side measures the *marginal rate of substitution,* MRS, of consumption as old for consumption as young, evaluated at the point (c_1, c_2). MRS is

[6] Alternatively, one could use the Lagrange method.

defined as the increase in period-$t+1$ consumption needed to compensate for a one-unit marginal decrease in period-t consumption. That is,

$$MRS_{c_2c_1} = -\frac{dc_{2t+1}}{dc_{1t}}|_{U=\bar{U}} = \frac{u'(c_{1t})}{(1+\rho)^{-1}u'(c_{2t+1})}, \tag{3.10}$$

where we have used implicit differentiation in $U(c_{1t}, c_{2t+1}) = \bar{U}$. The right-hand side of (3.9) indicates the marginal rate of transformation, MRT, which is the rate at which saving allows an agent to shift consumption from period t to period $t+1$ via the market. In an optimal plan MRS must equal MRT.

Even though interpretations in terms of "MRS equal to MRT" are more satisfactory, we will often use "as if" interpretations like the one before. They are a convenient short-hand for the more elaborate interpretation.

The Euler equation (3.8) implies that

$$\rho \lessgtr r_{t+1} \text{ causes } u'(c_{1t}) \gtrless u'(c_{2t+1}), \text{ i.e., } c_{1t} \lessgtr c_{2t+1},$$

respectively, in the optimal plan (because $u'' < 0$). That is, absent uncertainty the optimal plan entails either increasing, constant or decreasing consumption over time according to whether the rate of time preference is below, equal to, or above the market interest rate, respectively. For example, when $\rho < r_{t+1}$, the plan is to start with relatively low consumption in order to take advantage of the relatively high rate of return on saving.

Note that there are infinitely many pairs (c_{1t}, c_{2t+1}) satisfying the Euler equation (3.8). Only when requiring the two period budget constraints, (3.4) and (3.5), satisfied, do we get the unique solution s_t and thereby the unique solution for c_{1t} and c_{2t+1}.

Properties of the saving function

The first-order condition (FOC), where the two budget constraints are inserted, determines the saving as an implicit function of the market prices faced by the young decision maker, i.e., $s_t = s(w_t, r_{t+1})$.

The partial derivatives of this function can be found by applying the *implicit function theorem* on (FOC). A practical procedure is the following. We first write $d\tilde{U}_t/ds_t$ as a function, f, of the variables involved, s_t, w_t, and r_{t+1}, i.e.,

$$\frac{d\tilde{U}_t}{ds_t} = -u'(w_t - s_t) + (1+\rho)^{-1}u'((1+r_{t+1})s_t)(1+r_{t+1}) \equiv f(s_t, w_t, r_{t+1}).$$

By (FOC), $f(s_t, w_t, r_{t+1}) = 0$ and so the implicit function theorem (see Math tools) implies

$$\frac{\partial s_t}{\partial w_t} = -\frac{\partial f/\partial w_t}{D} \quad \text{and} \quad \frac{\partial s_t}{\partial r_{t+1}} = -\frac{\partial f/\partial r_{t+1}}{D},$$

where $D \equiv \partial f/\partial s_t \equiv d^2\tilde{U}_t/ds_t^2 < 0$ by (SOC). We find

$$\begin{aligned}\frac{\partial f}{\partial w_t} &= -u''(c_{1t}) > 0,\\ \frac{\partial f}{\partial r_{t+1}} &= (1+\rho)^{-1}\left[u'(c_{2t+1}) + u''(c_{2t+1})s_t(1+r_{t+1})\right].\end{aligned}$$

Consequently, the partial derivatives of the saving function $s_t = s(w_t, r_{t+1})$ are

$$s_w \equiv \frac{\partial s_t}{\partial w_t} = \frac{u''(c_{1t})}{D} > 0 \quad (\text{but } < 1), \tag{3.11}$$

$$s_r \equiv \frac{\partial s_t}{\partial r_{t+1}} = -\frac{(1+\rho)^{-1}[u'(c_{2t+1}) + u''(c_{2t+1})c_{2t+1}]}{D}, \tag{3.12}$$

where in the last expression we have used (3.5).[7]

We see that $0 < s_w < 1$, which implies that $0 < \partial c_{1t}/\partial w_t < 1$ and $0 < \partial c_{2t}/\partial w_t < 1 + r_{t+1}$. The positive sign of these two derivatives indicate that consumption in each of the periods is a *normal* good (which certainly is plausible since we are talking about the total consumption by the individual in each period).[8] The sign of s_r is seen to be ambiguous. This ambiguity reflects that the Slutsky substitution and income effects on consumption as young of a rise in the interest rate are of opposite signs. To understand this, it is useful to keep the intertemporal budget constraint, (3.7), in mind. The *substitution effect* on c_{1t} is negative because the higher interest rate makes future consumption cheaper in terms of current consumption. And the *income effect* on c_{1t} is positive because with a higher interest rate, a given budget can buy more consumption in both periods, cf. (3.7). Generally there would be a third Slutsky effect, a *wealth effect* of a rise in the interest rate. But such an effect is ruled out in this model. This is because there is no labor income in the second period of life. Indeed, as indicated

[7] A perhaps more straightforward procedure, not requiring full memory of the exact content of the implicit function theorem, is based on "implicit differentiation". First, keeping r_{t+1} fixed, one calculates the total derivative w.r.t. w_t on both sides of (FOC). Next, keeping w_t fixed, one calculates the total derivative w.r.t. r_{t+1} on both sides of (FOC).

Yet another possible procedure is based on "total differentiation" in terms of *differentials*. Taking the differential w.r.t. s_t, w_t, and r_{t+1} on both sides of (FOC) gives $-u''(c_{1t})(dw_t - ds_t) + +(1+\rho)^{-1}\cdot\{u''(c_{2t+1})\left[(1+r_{t+1})ds_t + s_t dr_{t+1}\right](1+r_{t+1}) + u'(c_{2t+1})dr_{t+1}\} = 0$. By rearranging we find the ratios ds_t/dw_t and ds_t/dr_{t+1}, which will indicate the value of the partial derivatives (3.11) and (3.12).

[8] Recall, a consumption good is called *normal* for given consumer preferences if the demand for it is an increasing function of the consumer's wealth. Since in this model the consumer is born without any financial wealth, the consumer's wealth at the end of period t is simply the present value of labor earnings through life, which here, evaluated at the beginning of period t, is $w_t/(1+r_t)$ as there is no labor income in the second period of life, cf. (3.7).

by (3.4), the human wealth of a member of generation t, evaluated at the end of period t, is simply w_t, which is independent of r_{t+1}.

Rewriting (3.12) gives

$$s_r = \frac{(1+\rho)^{-1}u'(c_{2t+1})[\theta(c_{2t+1})-1]}{D} \gtreqless 0 \text{ for } \theta(c_{2t+1}) \lesseqgtr 1, \tag{3.13}$$

respectively, where $D<0$, and where $\theta(c_{2t+1})$ is the absolute *elasticity of marginal utility* of consumption in the second period, that is,

$$\theta(c_{2t+1}) \equiv -\frac{c_{2t+1}}{u'(c_{2t+1})}u''(c_{2t+1}) \approx -\frac{\Delta u'(c_{2t+1})/u'(c_{2t+1})}{\Delta c_{2t+1}/c_{2t+1}} > 0,$$

where the approximation is valid for a "small" increase, Δc_{2t+1}, in c_{2t+1}. The inequalities in (3.13) show that when the absolute elasticity of marginal utility is below one, then the substitution effect on consumption as young of an increase in the interest rate dominates the income effect and saving increases. The opposite is true if the elasticity of marginal utility is above one.

The reason that $\theta(c_{2t+1})$ has this role is that $\theta(c_{2t+1})$ reflects how sensitive marginal utility of c_{2t+1} is to a rise in c_{2t+1}. To see the intuition, consider the case where consumption as young, and thus saving, happens to be unaffected by an increase in the interest rate. Even in this case, consumption as old, c_{2t+1}, is automatically increased (in view of the higher income as old through the higher rate of return on the unchanged saving); and the marginal utility of c_{2t+1} is thus decreased in response to a higher interest rate. The point is that this outcome can only be optimal if the elasticity of marginal utility of c_{2t+1} is of "medium" size. A very high absolute elasticity of marginal utility of c_{2t+1} would result in a sharp decline in marginal utility − so sharp that not much would be lost by dampening the automatic rise in c_{2t+1} and instead increase c_{1t}, thus reducing saving. On the other hand, a very low elasticity of marginal utility of c_{2t+1} would result in only a small decline in marginal utility − so small that it is beneficial to take advantage of the higher rate of return and save *more*, thus accepting a first-period utility loss brought about by a lower c_{1t}.

We see from (3.12) that an absolute elasticity of marginal utility equal to exactly one is the case leading to the interest rate being *neutral* vis-a-vis the saving of the young. What is the intuition behind this? Neutrality vis-a-vis the saving of the young of a rise in the interest rate requires that c_{1t} remains unchanged since $c_{1t} = w_t - s_t$. In turn this requires that the marginal utility, $u'(c_{2t+1})$, on the right-hand side of (3.8) falls by the same percentage as $1+r_{t+1}$ rises. At the same time the budget (3.5) as old tells us that c_{2t+1} has to rise by the same percentage as $1+r_{t+1}$ if s_t remains unchanged. Altogether we thus need that $u'(c_{2t+1})$ falls by the same percentage as c_{2t+1} rises. But this requires that the absolute elasticity of $u'(c_{2t+1})$ w.r.t. c_{2t+1} is exactly one.

The elasticity of marginal utility, also called the marginal utility flexibility, will generally depend on the level of consumption, as implicit in the notation $\theta(c_{2t+1})$. There exists a popular special case, however, where the elasticity of marginal utility is constant.

EXAMPLE 1 *The CRRA utility function.* If we impose the requirement that $u(c)$ should have an absolute elasticity of marginal utility of consumption equal to a constant $\theta > 0$, then one can show (see Appendix A) that the utility function must be of the CRRA form:

$$u(c) = \begin{cases} \frac{c^{1-\theta}-1}{1-\theta}, & \text{when } \theta \neq 1, \\ \ln c, & \text{when } \theta = 1. \end{cases} , \tag{3.14}$$

It may seem odd that in the upper case we subtract the constant $1/(1-\theta)$ from $c^{1-\theta}/(1-\theta)$. But adding or subtracting a constant from a utility function does not affect the marginal rate of substitution and consequently not behavior. Notwithstanding that we could do without this constant, its presence in (3.14) has two advantages. One is that in contrast to $c^{1-\theta}/(1-\theta)$, the expression $(c^{1-\theta}-1)/(1-\theta)$ can be interpreted as valid even for $\theta = 1$, namely as identical to $\ln c$. This is because $(c^{1-\theta}-1)/(1-\theta) \to \ln c$ for $\theta \to 1$ (by L'Hôpital's rule for "0/0"). Another advantage is that the kinship between the different members, indexed by θ, of the CRRA family becomes more transparent. Indeed, by defining $u(c)$ as in (3.14), all graphs of $u(c)$ will go through the same point as the log function, namely $(1, 0)$, cf. Fig. 3.2.

The higher is θ, the more "curvature" does the corresponding curve in Fig. 3.2 have. In turn, more "curvature" reflects a higher incentive to smooth consumption across time. The reason is that a large curvature means that the marginal utility will drop sharply if consumption rises and will increase sharply if consumption falls. Consequently, not much utility is lost by lowering consumption when it is relatively high but there is a lot of utility to be gained by raising it when it is relatively low. So the curvature θ indicates the degree of *aversion towards variation in consumption.* Or we may say that θ indicates the strength of the *preference for consumption smoothing.*[9] □

Suppose the period utility is of CRRA form as given in (3.14). (FOC) then yields an explicit solution for the saving of the young:

$$s_t = \frac{1}{1 + (1+\rho)(\frac{1+r_{t+1}}{1+\rho})^{\frac{\theta-1}{\theta}}} w_t. \tag{3.15}$$

[9]The name CRRA is a shorthand for *Constant Relative Risk Aversion* and comes from the theory of behavior under uncertainty. Also in that theory does the CRRA function constitute an important benchmark case. And θ is in that context called the *degree of relative risk aversion.*

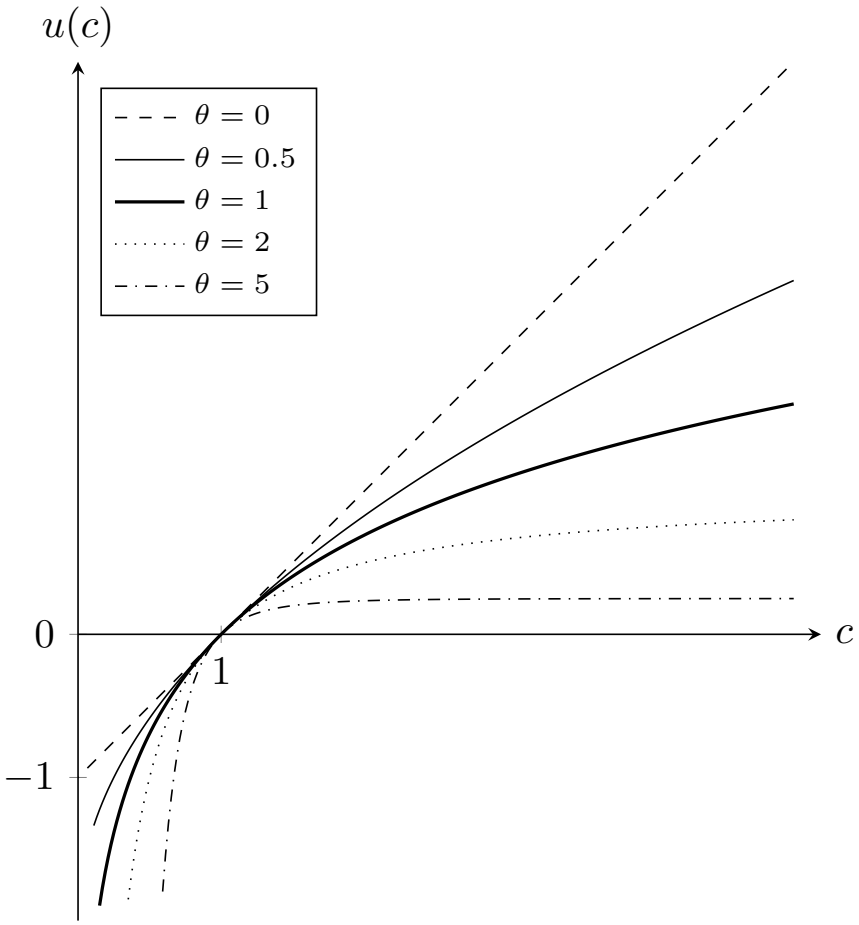

Figure 3.2: The CRRA family of utility functions.

We see that the signs of $\partial s_t/\partial w_t$ and $\partial s_t/\partial r_{t+1}$ shown in (3.11) and (3.13), respectively, are confirmed. Moreover, the saving of the young is in this special case proportional to income with a factor of proportionality that depends on the interest rate (as long as $\theta \neq 1$). But in the general case the saving-income ratio depends also on the income level.

A major part of the attempts at empirically estimating θ suggests that $\theta > 1$. Based on U.S. data, Hall (1988) provides estimates above 5, while Attanasio and Weber (1993) suggest $1.25 \leq \theta \leq 3.33$. For Japanese data Okubo (2011) suggests $2.5 \leq \theta \leq 5.0$. As these studies relate to much shorter time intervals than the implicit time horizon of about 2×30 years in the Diamond model, we should be cautious. But if the estimates *were* valid also to that model, we should expect the income effect on current consumption of an increase in the interest rate to dominate the substitution effect, thus implying $s_r < 0$ *as long as there is no*

wealth effect of a rise in the interest rate.

When the elasticity of marginal utility of consumption is a constant, θ, its inverse, $1/\theta$, equals the *elasticity of intertemporal substitution* in consumption. This concept refers to the willingness to substitute consumption over time when the interest rate changes. Under certain conditions the elasticity of intertemporal substitution reflects the elasticity of the ratio c_{2t+1}/c_{1t} w.r.t. $1 + r_{t+1}$ when we move along a given indifference curve. The next subsection, which can be omitted in a first reading, goes more into detail with the concept.

Digression: The elasticity of intertemporal substitution*

Consider a two-period consumption problem like the one above. Fig. 3.3 depicts a particular indifference curve, $u(c_1) + (1+\rho)^{-1}u(c_2) = \bar{U}$. At a given point, (c_1, c_2), on the curve, the marginal rate of substitution of period-2 consumption for period-1 consumption, MRS, is given by

$$MRS = -\frac{dc_2}{dc_1}\,|_{U=\bar{U}}\ ,$$

that is, MRS at the point (c_1, c_2) is the absolute value of the slope of the tangent to the indifference curve at that point.[10] Under the "normal" assumption of "strictly convex preferences" (as for instance in the Diamond model), MRS is rising along the curve when c_1 decreases (and thereby c_2 increases). Conversely, we can let MRS be the independent variable and consider the corresponding point on the indifference curve, and thereby the ratio c_2/c_1, as a function of MRS. If we raise MRS along the indifference curve, the corresponding value of the ratio c_2/c_1 will also rise.

The *elasticity of intertemporal substitution in consumption* at a given point is defined as the elasticity of the ratio c_2/c_1 w.r.t. the marginal rate of substitution of c_2 for c_1, when we move along the indifference curve through the point (c_1, c_2). Letting the elasticity w.r.t. x of a differentiable function $f(x)$ be denoted $\mathrm{E}\ell_x f(x)$, the elasticity of intertemporal substitution in consumption can be written

$$\mathrm{E}\ell_{MRS}\frac{c_2}{c_1} = \frac{MRS}{c_2/c_1}\frac{d\,(c_2/c_1)}{dMRS}\,|_{U=\bar{U}} \approx \frac{\frac{\Delta(c_2/c_1)}{c_2/c_1}}{\frac{\Delta MRS}{MRS}},$$

where the approximation is valid for a "small" increase, ΔMRS, in MRS.

A more concrete understanding is obtained when we take into account that in the consumer's optimal plan, MRS equals the ratio of the discounted prices

[10] When the meaning is clear from the context, to save notation we just write MRS instead of the more precise $MRS_{c_2c_1}$.

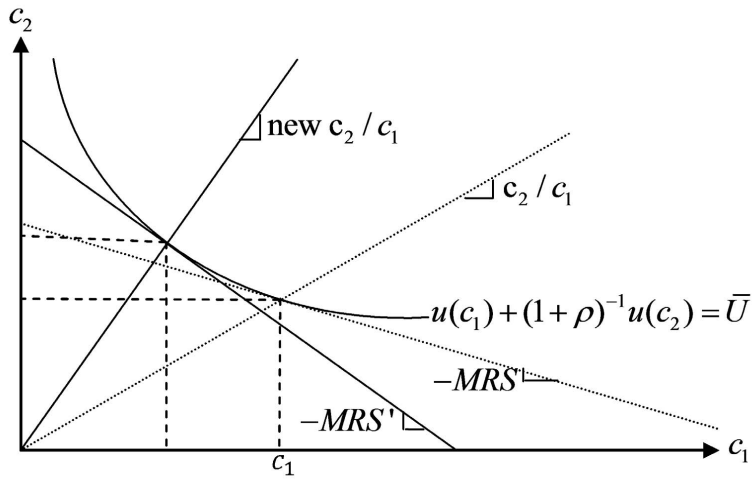

Figure 3.3: Substitution of period 2-consumption for period 1-consumption as MRS increases to MRS'.

of good 1 and good 2, that is, the ratio $1/(1/(1+r))$ given in (3.7). Indeed, from (3.10) and (3.9), omitting the time indices, we have

$$MRS = -\frac{dc_2}{dc_1}|_{U=\bar{U}} = \frac{u'(c_1)}{(1+\rho)^{-1}u'(c_2)} = 1 + r \equiv R. \tag{3.16}$$

Letting $\sigma(c_1, c_2)$ denote the elasticity of intertemporal substitution, evaluated at the point (c_1, c_2), we then have

$$\sigma(c_1, c_2) = \frac{R}{c_2/c_1}\frac{d\,(c_2/c_1)}{dR}|_{U=\bar{U}} \approx \frac{\frac{\Delta(c_2/c_1)}{c_2/c_1}}{\frac{\Delta R}{R}}. \tag{3.17}$$

Consequently, the elasticity of intertemporal substitution can here be interpreted as the approximate percentage increase in the consumption ratio, c_2/c_1, triggered by a one percentage increase in the inverse price ratio, holding the utility level unchanged.[11]

Given $u(c)$, we let $\theta(c)$ be the absolute elasticity of marginal utility of consumption, i.e., $\theta(c) \equiv -cu''(c)/u'(c)$. As shown in Appendix B, we then find the elasticity of intertemporal substitution to be

$$\sigma(c_1, c_2) = \frac{c_2 + Rc_1}{c_2\theta(c_1) + Rc_1\theta(c_2)}. \tag{3.18}$$

[11] This characterization is equivalent to saying that the elasticity of substitution between two consumption goods indicates the approximate percentage *decrease* in the ratio of the chosen quantities of the goods (when moving along a given indifference curve) induced by a one-percentage *increase* in the *corresponding* price ratio.

We see that if $u(c)$ belongs to the CRRA class and thereby $\theta(c_1) = \theta(c_2) = \theta$, then $\sigma(c_1, c_2) = 1/\theta$. In this case (as well as whenever $c_1 = c_2$) the elasticity of marginal utility and the elasticity of intertemporal substitution are simply the inverse of each other.

3.4 Production

Output is homogeneous and can be used for consumption as well as investment in physical capital. The capital stock is thereby just accumulated non-consumed output. We may imagine a "corn economy" where output is corn, part of which is eaten (flour) while the remainder is accumulated as capital (seed corn).

The specification of technology and production conditions follows the simple competitive one-sector setup discussed in Chapter 2. Although the Diamond model is a long-run model, we shall in this chapter for simplicity ignore technological change.

The representative firm

There is a representative firm with a neoclassical production function and constant returns to scale (CRS). Omitting the time argument t when not needed for clarity, we have

$$Y = F(K, L) = LF(k, 1) \equiv Lf(k), \qquad f' > 0, f'' < 0, \tag{3.19}$$

where Y is output (GNP) per period, K is capital input, L is labor input, and $k \equiv K/L$ is the capital-labor ratio. The derived function, f, is the production function in intensive form. Capital installation and other adjustment costs are ignored. Hence profit is $\Pi \equiv F(K, L) - \hat{r}K - wL$. The firm maximizes Π under perfect competition. This gives, first, $\partial\Pi/\partial K = F_K(K, L) - \hat{r} = 0$, that is,

$$F_K(K, L) = \frac{\partial [Lf(k)]}{\partial K} = f'(k) = \hat{r}. \tag{3.20}$$

Second, $\partial\Pi/\partial L = F_L(K, L) - w = 0$, that is,

$$F_L(K, L) = \frac{\partial [Lf(k)]}{\partial L} = f(k) - kf'(k) = w. \tag{3.21}$$

The interpretation is that the firm will in every period use capital up to the point where the marginal productivity of capital equals the rental rate given from the market. Similarly, the firm will employ labor up to the point where the marginal productivity of labor equals the wage rate given from the market.

In view of $f'' < 0$, a $k > 0$ satisfying (3.20) is unique. Let us call it the *desired capital-labor ratio*. Owing to CRS, however, at this stage the separate factor inputs, K and L, are indeterminate; only their ratio, k, is determinate.[12] We will now see how the equilibrium conditions for the factor markets select the factor prices and the level of factor inputs consistent with equilibrium.

Factor prices in equilibrium

Let the aggregate demand for capital services and labor services be denoted K^d and L^d, respectively. Clearing in factor markets in period t implies

$$K_t{}^d = K_t, \tag{3.22}$$

$$L_t{}^d = L_t = L_0(1+n)^t, \tag{3.23}$$

where K_t is the aggregate supply of capital services and L_t the aggregate supply of labor services. As was called attention to in Chapter 1, unless otherwise specified it is understood that the rate of utilization of each production factor is constant over time and normalized to one. So the quantity K_t will at one and the same time measure both the capital input, a flow, and the available capital stock. Similarly, the quantity L_t will at one and the same time measure both the labor input, a flow, and the size of the labor force as a stock (= the number of young people).

The aggregate input demands, K^d and L^d, are linked through the desired capital-labor ratio, k^d. In equilibrium we have $K_t^d/L_t^d = k_t{}^d = K_t/L_t \equiv k_t$, by (3.22) and (3.23). The k in (3.20) and (3.21) can thereby be identified with the ratio of the stock supplies, $k_t \equiv K_t/L_t > 0$, which is a predetermined variable. Interpreted this way, (3.20) and (3.21) *determine* the equilibrium factor prices $\hat{r}_t$ and w_t in each period. In view of the no-arbitrage condition (3.2), the real interest rate satisfies $r_t = \hat{r}_t - \delta$, where δ is the capital depreciation rate, $0 \leq \delta \leq 1$, and so in equilibrium we end up with

$$r_t = f'(k_t) - \delta \equiv r(k_t) \qquad (r'(k_t) = f''(k_t) < 0), \tag{3.24}$$

$$w_t = f(k_t) - k_t f'(k_t) \equiv w(k_t) \qquad (w'(k_t) = -k_t f''(k_t) > 0), \tag{3.25}$$

where causality is from the right to the left in the two equations. In line with our general perception of perfect competition, cf. Section 2.4 of Chapter 2, it is understood that the factor prices, $\hat{r}_t$ and w_t, adjust quickly to the market-clearing levels.

[12] It might seem that k is overdetermined because we have two equations, (3.20) and (3.21), but only one unknown. This reminds us that for *arbitrary* factor prices, $\hat{r}$ and w, there will *not* exist a k satisfying both (3.20) and (3.21). But in equilibrium the factor prices faced by the firm are not arbitrary. They are equilibrium prices, i.e., they are adjusted so that (3.20) and (3.21) become consistent.

Technical Remark. In these formulas it is understood that $L > 0$, but we may allow $K = 0$, i.e., $k = 0$. In case $f'(0)$ is not immediately well-defined, we interpret $f'(0)$ as $\lim_{k\to 0^+} f'(k)$ if this limit exists. If it does not, it must be because we are in a situation where $\lim_{k\to 0^+} f'(k) = \infty$, since $f''(k) < 0$ (an example is the Cobb-Douglas function, $f(k) = Ak^\alpha$, $0 < \alpha < 1$, where $\lim_{k\to 0^+} f'(k) = \lim_{k\to 0^+} A\alpha k^{\alpha-1} = +\infty$). In this situation we simply include $+\infty$ in the range of $r(k)$ and define $r(0) \cdot 0 \equiv \lim_{k\to 0^+}(f'(k) - \delta)k = 0$, where the last equality comes from the general property of a neoclassical CRS production function that $\lim_{k\to 0^+} kf'(k) = 0$, cf. (2.18) of Chapter 2. Letting $r(0) \cdot 0 = 0$ also fits well with intuition since, when $k = 0$, nobody receives capital income anyway. Note that since $\delta \in [0, 1]$, $r(k) > -1$ for all $k \geq 0$. What about $w(0)$? We interpret $w(0)$ as $\lim_{k\to 0} w(k)$. From (2.18) of Chapter 2 we have that $\lim_{k\to 0^+} w(k) = f(0) \equiv F(0, 1) \geq 0$. If capital is essential, $F(0, 1) = 0$. Otherwise, $F(0, 1) > 0$. Finally, since $w' > 0$, we have, for $k > 0$, $w(k) > 0$ as also noted in Chapter 2. □

To fix ideas we have assumed that households (here the old) own the physical capital and rent it out to the firms. In view of perfect competition and constant returns to scale, pure profit is nil in equilibrium. As long as the model ignores uncertainty and capital installation costs, the results will be unaffected if instead we let the firms themselves own the physical capital and finance capital investment by issuing bonds and shares. These bonds and shares would then be accumulated by the households and constitute their financial wealth instead of the capital goods themselves. The equilibrium rate of return, r_t, would be the same.

3.5 The dynamic path of the economy

As in other fields of economics, it is important to distinguish between the set of technically feasible allocations and an allocation brought about, within this set, by a specific economic institution (the rules of the game). The economic institution assumed by the Diamond model is the private-ownership perfect-competition market institution.

We shall in the next subsections introduce three different concepts concerning allocations over time in this economy. The three concepts are: *technically feasible paths*, *temporary equilibrium*, and *equilibrium path*. These concepts are mutually related in the sense that there is a whole *set* of technically feasible paths, *within which* there may exist a unique equilibrium path, which in turn is a sequence of states that have certain properties, including the temporary equilibrium property.

3.5.1 Technically feasible paths

When we speak of technically feasible paths, the focus is merely upon what is feasible from the point of view of the given technology as such and available initial resources. That is, we disregard the agents' preferences, their choices given the constraints, their interactions in markets, the market forces etc.

The technology is represented by (3.19) and there are two exogenous resources, the labor force, $L_t = L_0(1+n)^t$, and the initial capital stock, K_0. From national income accounting aggregate consumption can be written $C_t \equiv Y_t - S_t = F(K_t, L_t) - S_t$, where S_t denotes aggregate gross saving, and where we have inserted (3.19). In a closed economy aggregate gross saving equals (ex post) aggregate gross investment, $K_{t+1} - K_t + \delta K_t$. So

$$C_t = F(K_t, L_t) - (K_{t+1} - K_t + \delta K_t). \tag{3.26}$$

Let c_t denote aggregate consumption per unit of labor in period t, i.e.,

$$c_t \equiv \frac{C_t}{L_t} = \frac{c_{1t}L_t + c_{2t}L_{t-1}}{L_t} = c_{1t} + \frac{c_{2t}}{1+n}.$$

Combining this with (3.26) and using the definitions of k and $f(k)$, we obtain the dynamic resource constraint of the economy:

$$c_{1t} + \frac{c_{2t}}{1+n} = f(k_t) + (1-\delta)k_t - (1+n)k_{t+1}. \tag{3.27}$$

DEFINITION 1 Let $\bar{k}_0 \geq 0$ be the historically given initial ratio of available capital and labor. The path $\{(k_t, c_{1t}, c_{2t})\}_{t=0}^{\infty}$ is called *technically feasible* if it has $k_0 = \bar{k}_0$ and for all $t = 0, 1, 2, \ldots$, (3.27) has $k_t \geq 0$, $c_{1t} \geq 0$, and $c_{2t} \geq 0$.

The next subsections consider how, for given household preferences, the private-ownership market institution with profit-maximizing firms under perfect competition generates a *selection* within the set of technically feasible paths. A member of this selection (which may but need not have just one member) is called an *equilibrium path.* It constitutes a sequence of states with certain properties, one of which is the temporary equilibrium property.

3.5.2 A temporary equilibrium

Standing in a given period, it is natural to think of next period's interest rate as an *expected* interest rate that provisionally can deviate from the ex post realized one. We let r_{t+1}^e denote the expected real interest rate of period $t+1$ as seen from period t.

Essentially, by a temporary equilibrium in period t is meant a state where for a *given* r^e_{t+1}, all markets clear in the period. There are three markets, namely two factor markets and a market for produced goods. We have already described the two factor markets. In the market for produced goods the representative firm supplies the amount $Y^s_t = F(K^d_t, L^d_t)$ in period t. The demand side in this market has two components, consumption, C_t, and *gross* investment, I_t. Equilibrium in the goods market requires that demand equals supply, i.e.,

$$C_t + I_t = c_{1t}L_t + c_{2t}L_{t-1} + I_t = Y^s_t = F(K^d_t, L^d_t), \tag{3.28}$$

where consumption by the young and old, c_{1t} and c_{2t}, respectively, were determined in Section 3.

By definition, aggregate gross investment equals aggregate net investment, I^N_t, plus capital depreciation, i.e.,

$$I_t = I^N_t + \delta K_t \equiv I^N_{1t} + I^N_{2t} + \delta K_t \equiv S^N_{1t} + S^N_{2t} + \delta K_t = s_t L_t + (-K_t) + \delta K_t. \tag{3.29}$$

The first equality follows from the definition of net investment and the assumption that capital depreciation equals δK_t. Next comes an identity reflecting that aggregate net investment is the sum of net investment by the young and net investment by the old. In turn, saving in this model is directly an act of acquiring capital goods. So the net investment by the young, I^N_{1t}, and the old, I^N_{2t}, are identical to their net saving, S^N_{1t} and S^N_{2t}, respectively. As we have shown, the net saving by the young in the model equals $s_t L_t$. And the net saving by the old is negative and equals $-K_t$. Indeed, because they have no bequest motive, the old consume all they have and leave nothing as bequests. Hence, the young in any period enter the period with no non-human wealth. Consequently, any non-human wealth existing at the beginning of a period must belong to the old in that period and be the result of their saving as young in the previous period. As K_t constitutes the aggregate non-human wealth in our closed economy at the beginning of period t, we therefore have

$$s_{t-1}L_{t-1} = K_t. \tag{3.30}$$

Recalling that the net saving of any group is by definition the same as the increase in its non-human wealth, the net saving of the old in period t is $-K_t$. Aggregate net saving in the economy is thus $s_t L_t + (-K_t)$, and (3.29).is thereby explained.

DEFINITION 2 For a given period t with capital stock $K_t \geq 0$ and labor supply $L_t > 0$, let the expected real interest rate be given as $r^e_{t+1} > -1$. With $k_t \equiv K_t/L_t$, a *temporary equilibrium* in period t is a state $(k_t, c_{1t}, c_{2t}, w_t, r_t)$ of the economy such that (3.22), (3.23), (3.28), and (3.29) hold (i.e., all markets clear) for $c_{1t} = w_t - s_t$ and $c_{2t} = (k_t + r_t k_t)(1 + n)$, where $s_t = s(w_t, r^e_{t+1})$, as defined in

Lemma 1, while $w_t = w(k_t) > 0$ and $r_t = r(k_t)$, as defined in (3.25) and (3.24), respectively.

The reason for the requirement $w_t > 0$ in the definition is that if $w_t = 0$, people would have nothing to live on as young and nothing to save from for retirement. The system would not be economically viable in this case. With regard to the equation for c_{2t} in the definition, note that (3.30) gives $s_{t-1} = K_t/L_{t-1} = (K_t/L_t)(L_t/L_{t-1}) = k_t(1+n)$, which is the wealth of each old at the beginning of period t. Substituting into $c_{2t} = (1+r_t)s_{t-1}$, we get $c_{2t} = (1+r_t)k_t(1+n)$, which can also be written $c_{2t} = (k_t + r_t k_t)(1+n)$. This last way of writing c_{2t} has the advantage of being applicable even if $k_t = 0$, cf. Technical Remark in Section 3.4. The remaining conditions for a temporary equilibrium are self-explanatory.

PROPOSITION 1 Suppose the No Fast Assumption (A1) applies. Consider a given period t with a given $k_t \geq 0$. Then for any $r^e_{t+1} > -1$,
(i) if $k_t > 0$, there exists a temporary equilibrium, $(k_t, c_{1t}, c_{2t}, w_t, r_t)$, and c_{1t} and c_{2t} are positive;
(ii) if $k_t = 0$, a temporary equilibrium exists if and only if capital is not essential; in that case, $w_t = w(k_t) = w(0) = f(0) > 0$ and c_{1t} and s_t are positive (while $c_{2t} = 0$);
(iii) whenever a temporary equilibrium exists, it is unique.

Proof. We begin with (iii). That there is at most one temporary equilibrium is immediately obvious since w_t and r_t are functions of the given k_t : $w_t = w(k_t)$ and $r_t = r(k_t)$. And given w_t, r_t, and r^e_{t+1}, c_{1t} and c_{2t} are uniquely determined.

(i) Let $k_t > 0$. Then, by (3.25), $w(k_t) > 0$. We claim that the state $(k_t, c_{1t}, c_{2t}, w_t, r_t)$, with $w_t = w(k_t)$, $r_t = r(k_t)$, $c_{1t} = w(k_t) - s(w(k_t), r^e_{t+1})$, and $c_{2t} = (1+r(k_t))k_t(1+n)$, is a temporary equilibrium. Indeed, Section 3.4 showed that the factor prices $w_t = w(k_t)$ and $r_t = r(k_t)$ are consistent with clearing in the factor markets in period t. Given that these markets clear (by price adjustment), it follows by Walras' law (see Appendix C) that also the third market, the goods market, clears in period t. So all criteria in Definition 2 are satisfied. That $c_{1t} > 0$ follows from $w(k_t) > 0$ and the No Fast Assumption (A1), in view of Lemma 1. That $c_{2t} > 0$ follows from $c_{2t} = (1 + r(k_t))k_t(1+n)$ when $k_t > 0$, since $r(k_t) > -1$ always.

(ii) Let $k_t = 0$. Suppose $f(0) > 0$. Then, by Technical Remark in Section 3.4, $w_t = w(0) = f(0) > 0$ and $c_{1t} = w_t - s(w_t, r^e_{t+1})$ is well-defined, positive, and less than w_t, in view of Lemma 1; so $s_t = s(w_t, r^e_{t+1}) > 0$. The old in period 0 will starve since $c_{2t} = (0+0)(1+n)$, in view of $r(0) \cdot 0 = 0$, cf. Technical Remark in Section 3.4. Even though this is a bad situation for the old, it is consistent with the criteria in Definition 2. On the other hand, if $f(0) = 0$, we get $w_t = f(0) = 0$, which violates one of the criteria in Definition 2. □

Point (ii) of the proposition says that a temporary equilibrium *may* exist even in a period where $k = 0$. The old in this period will starve and not survive. But if capital is not essential, the young get positive labor income out of which they will save a part for their old age and be able to maintain life also next period which will be endowed with positive capital. Then, by our assumptions the economy is viable forever.[13]

Generally, the term "equilibrium" is used to denote a state of "rest", possibly only "temporary rest". The temporary equilibrium in the present model is an example of a state of "temporary rest" in the following sense: (a) the agents optimize, given their expectations and the constraints they face; and (b) the aggregate demands and supplies in the given period are mutually consistent, i.e., markets clear. The qualification "temporary" is motivated by two features. First, in the next period circumstances may be different, among other things as a consequence of the currently chosen actions. Second, the given expectations may turn out wrong.

3.5.3 An equilibrium path

The concept of an equilibrium path, also called an intertemporal equilibrium, requires more conditions satisfied. The concept refers to a sequence of temporary equilibria such that *expectations* of the agents are *fulfilled* in every period:

DEFINITION 3 An *equilibrium path* is a technically feasible path $\{(k_t, c_{1t}, c_{2t})\}_{t=0}^{\infty}$ such that for $t = 0, 1, 2, \ldots$, the state $(k_t, c_{1t}, c_{2t}, w_t, r_t)$ is a temporary equilibrium with $r_{t+1}^e = r\,(k_{t+1})$.

To characterize such a path, we forward (3.30) one period and rearrange so as to get

$$K_{t+1} = s_t L_t. \tag{3.31}$$

Since $K_{t+1} \equiv k_{t+1}L_{t+1} = k_{t+1}L_t(1+n)$, this can be written

$$k_{t+1} = \frac{s\,(w\,(k_t)\,, r\,(k_{t+1}))}{1+n}, \tag{3.32}$$

using that $s_t = s(w_t, r_{t+1}^e)$, $w_t = w(k_t)$, and $r_{t+1}^e = r_{t+1} = r\,(k_{t+1})$ in a sequence of temporary equilibria with fulfilled expectations. Equation (3.32) is a first-order difference equation, known as the *fundamental difference equation* or the *law of motion* of the Diamond model.

PROPOSITION 2 Suppose the No Fast Assumption (A1) applies. Then,

[13] For simplicity, the model ignores that in practice a certain minimum per capita consumption level (the subsistence minimum) is needed for viability.

(i) for any $k_0 > 0$ there exists at least one equilibrium path;
(ii) if $k_0 = 0$, an equilibrium path exists if and only if $f(0) > 0$ (i.e., capital not essential);
(iii) in any case, an equilibrium path has a positive real wage in all periods and positive capital in all periods except possibly the first;
(iv) an equilibrium path satisfies the first-order difference equation (3.32).

Proof. (i) and (ii): see Appendix D. (iii) For a given t, let $k_t \geq 0$. Then, since an equilibrium path is a sequence of temporary equilibria, we have $w_t = w(k_t) > 0$ and $s_t = s(w(k_t), r^e_{t+1})$, where $r^e_{t+1} = r(k_{t+1})$. Hence, by Lemma 1, $s(w(k_t), r^e_{t+1}) > 0$, which implies $k_{t+1} > 0$, in view of (3.32). This shows that only for $t = 0$ is $k_t = 0$ possible along an equilibrium path. (iv) This was shown in the text above. □

The formal proofs of point (i) and (ii) of the proposition are placed in appendix because they are quite technical. But the graphs in the ensuing figures 3.4-3.7 provide an intuitive verification. The "only if" part of point (ii) reflects the not very surprising fact that *if* capital were an essential production factor, no capital "now" would imply no income "now", hence no saving and investment and thus no capital in the next period and so on. On the other hand, the "if" part of point (ii) says that when capital is not essential, an equilibrium path can set off even from an initial period with no capital. Then point (iii) adds that an equilibrium path will have positive capital in all subsequent periods. Finally, as to point (iv), note that the fundamental difference equation, (3.32), rests on equation (3.31). Recall from the previous subsection that the economic logic behind this key equation is that since capital is the only non-human asset in the economy and the young are born without any inheritance, the aggregate capital stock at the beginning of period $t+1$ *must* be owned by the old generation in that period. It must thereby equal the aggregate saving these people had in the previous period where they were young.

The transition diagram

To be able to further characterize equilibrium paths, we construct a transition diagram in the (k_t, k_{t+1}) plane. The *transition curve* is defined as the set of points (k_t, k_{t+1}) satisfying (3.32). Its form and position depends on the households' preferences and the firms' technology. Fig. 3.4 shows one *possible*, but far from necessary configuration of this curve. A complicating circumstance is that the equation (3.32) has k_{t+1} on both sides. Sometimes we are able to solve the equation explicitly for k_{t+1} as a function of k_t, but sometimes we can do so only implicitly. What is even worse is that there are cases where k_{t+1} is not unique for a given k_t. We will proceed step by step.

First, what can we say about the *slope* of the transition curve? In general a point on the transition curve has the property that at least in a small neighborhood of this point the equation (3.32) will define k_{t+1} as an implicit function of k_t.[14] Taking the total derivative w.r.t. k_t on both sides of (3.32), we get

$$\frac{dk_{t+1}}{dk_t} = \frac{1}{1+n}\left(s_w w'(k_t) + s_r r'(k_{t+1})\frac{dk_{t+1}}{dk_t}\right). \tag{3.33}$$

By ordering, the slope of the transition curve within this small neighborhood can be written

$$\frac{dk_{t+1}}{dk_t} = \frac{s_w\left(w(k_t), r(k_{t+1})\right) w'(k_t)}{1+n-s_r\left(w(k_t), r(k_{t+1})\right) r'(k_{t+1})}, \tag{3.34}$$

when $s_r(w(k_t), r(k_{t+1}))r'(k_{t+1}) \neq 1+n$. Since $s_w > 0$ and $w'(k_t) = -k_t\, f''(k_t) > 0$, the numerator in (3.34) is always positive and we have

$$\frac{dk_{t+1}}{dk_t} \gtrless 0 \text{ for } s_r(w(k_t), r(k_{t+1})) \gtrless \frac{1+n}{r'(k_{t+1})},$$

respectively (recall that $r'(k_{t+1}) = f''(k_{t+1}) < 0$).

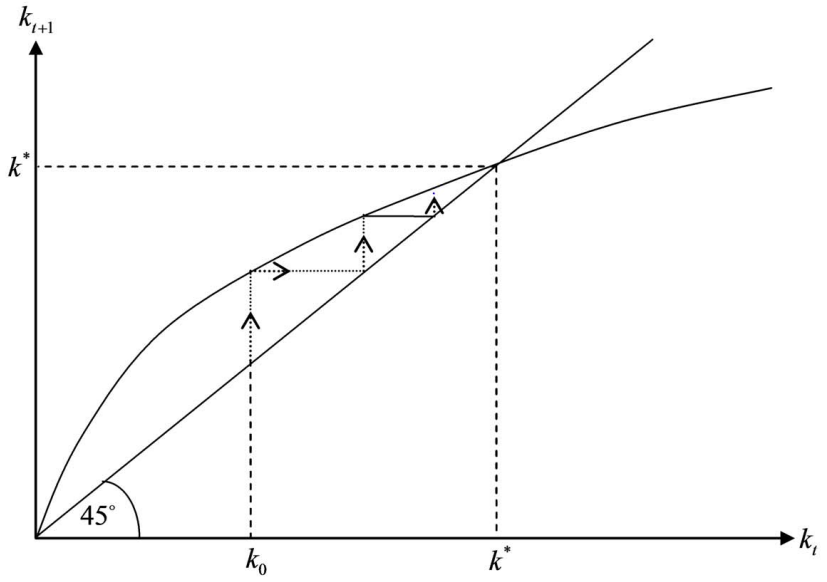

Figure 3.4: Transition curve and the resulting dynamics in the log-utility Cobb-Douglas case.

It follows that the transition curve is universally upward-sloping if and only if $s_r(w(k_t), r(k_{t+1})) > (1+n)/r'(k_{t+1})$ everywhere along the transition curve. The

[14] An exception occurs if the denominator in (3.34) below vanishes.

intuition behind this becomes visible by rewriting (3.34) in terms of small changes in k_t and k_{t+1}. Since $\Delta k_{t+1}/\Delta k_t \approx dk_{t+1}/dk_t$ for Δk_t "small", (3.34) implies

$$[1+n-s_r(\cdot)\,r'(k_{t+1})]\,\Delta k_{t+1} \approx s_w(\cdot)\;\; w'(k_t)\Delta k_t. \qquad (*)$$

Let $\Delta k_t > 0$. This rise in k_t will always raise wage income and, via the resulting rise in s_t, raise k_{t+1}, everything else equal. Everything else is *not* equal, however, since a rise in k_{t+1} implies a fall in the rate of interest. There are four cases to consider:

Case 1: $s_r(\cdot) = 0$. Then there is no feedback effect from the fall in the rate of interest. So the tendency to a rise in k_{t+1} is neither offset nor fortified.

Case 2: $s_r(\cdot) > 0$. Then the tendency to a rise in k_{t+1} will be partly offset through the *dampening* effect on saving resulting from the fall in the interest rate. This negative feedback can not fully offset the tendency to a rise in k_{t+1}. The reason is that the negative feedback on the saving of the young will only be there *if* the interest rate falls in the first place. We cannot in a period have both a *fall* in the interest rate triggering lower saving *and* a *rise* in the interest rate (via a lower k_{t+1}) *because* of the lower saving. So a *sufficient* condition for a universally upward-sloping transition curve is that the saving of the young is a non-decreasing function of the interest rate.

Case 3: $(1+n)/r'(k_{t+1}) < s_r(\cdot) < 0$. Then the tendency to a rise in k_{t+1} will be fortified through the *stimulating* effect on saving resulting from the fall in the interest rate.

Case 4: $s_r(\cdot) < (1+n)/r'(k_{t+1}) < 0$. Then the expression in brackets on the left-hand side of (*) is negative and requires therefore that $\Delta k_{t+1} < 0$ in order to comply with the positive right-hand side. This is a situation of multiple temporary equilibria, a situation where self-fulfilling expectations operate. We shall explore this case in the next sub-section.

Another feature of the transition curve is the following:

LEMMA 2 *(the transition curve is nowhere flat)* For all $k_t > 0$, $dk_{t+1}/dk_t \neq 0$.

Proof. Since $s_w > 0$ and $w'(k_t) > 0$ always, the numerator in (3.34) is always positive. □

The implication is that no part of the transition curve can be horizontal.[15]

When the transition curve crosses the 45° degree line for some $k_t > 0$, as in the example in Fig. 3.4, we have a steady state at this k_t. Formally:

DEFINITION 4 An equilibrium path $\{(k_t, c_{1t}, c_{2t})\}_{t=0}^{\infty}$ is in a *steady state* with capital-labor ratio $k^* > 0$ if the fundamental difference equation, (3.32), is satisfied with k_t as well as k_{t+1} replaced by k^*.

[15] This would not necessarily hold if the utility function were not time-separable.

This exemplifies the notion of a steady state as a stationary point in a dynamic process. Some economists use the term "dynamic equilibrium" instead of "steady state". As in this book the term "equilibrium" refers to situations where the constraints and decided actions of the market participants are mutually compatible, an economy can be in "equilibrium" without being in a steady state. A steady state is seen as a *special* sequence of temporary equilibria with fulfilled expectations, namely one with the property that the dynamic variable, here k, entering the fundamental difference equation does not change over time.

EXAMPLE 2 (the log utility Cobb-Douglas case) Let $u(c) = \ln c$ and $Y = AK^{\alpha}L^{1-\alpha}$, where $A > 0$ and $0 < \alpha < 1$. Since $u(c) = \ln c$ is the case $\theta = 1$ in Example 1, by (3.15) we have $s_r = 0$. Indeed, with logarithmic utility the substitution and income effects on s_t offset each other; and, as discussed above, in the Diamond model there can be no wealth effect of a rise in r_{t+1}. Further, the equation (3.32) reduces to a transition *function,*

$$k_{t+1} = \frac{(1-\alpha)Ak_t^{\alpha}}{(1+n)(2+\rho)}. \tag{3.35}$$

The associated transition curve is shown in Fig. 3.4 and there is for $k_0 > 0$ both a unique equilibrium path and a unique steady state with capital-labor ratio

$$k^* = \left(\frac{(1-\alpha)A}{(2+\rho)(1+n)}\right)^{1/(1-\alpha)} > 0.$$

At $k_t = k^*$ the slope of the transition curve is necessarily less than one. The dynamics therefore lead to convergence to the steady state as illustrated in the figure.[16] In the steady state the interest rate is $r^* = f'(k^*) - \delta = \alpha(1+n)(2+\rho)/(1-\alpha) - \delta$. Note that a higher n results in a lower k^*, hence a higher r^*. □

Because the Cobb-Douglas production function implies that capital is essential, (3.35) implies $k_{t+1} = 0$ if $k_t = 0$. The state $k_{t+1} = k_t = 0$ is thus a stationary point of the difference equation (3.35) considered in isolation. This state is not, however, an equilibrium path as defined above (not a steady state of an *economic* system since there is no production). We may call it a *trivial* steady state in contrast to the economically viable steady state $k_{t+1} = k_t = k^* > 0$ which is then called a *non-trivial* steady state.

Theoretically, there may be more than one (non-trivial) steady state. Nonexistence of a steady state is also possible. But before considering these possibilities, the next subsection (which may be skipped in a first reading) addresses an even more defiant feature which is that for a given k_0 there may exist more than one equilibrium path.

[16] A formal proof can be based on the mean value theorem.

The possibility of multiple equilibrium paths*

It turns out that a *backward-bending* transition curve like that in Fig. 3.5 is possible within the model. Not only are there two steady states but for $k_t \in (\underline{k}, \overline{k})$ there are *three temporary equilibria* with self-fulfilling expectations. That is, for a given k_t in this interval, there are three different values of k_{t+1} that are consistent with self-fulfilling expectations. Exercise 3.3 at the end of the chapter documents this possibility by way of a numerical example.

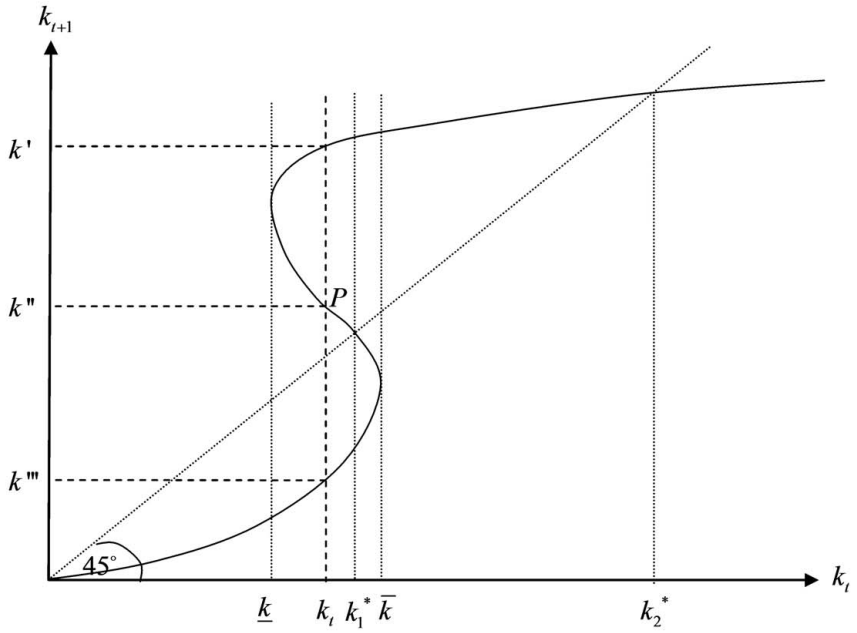

Figure 3.5: A backward-bending transition curve leads to multiple temporary equilibria with self-fulfilling expectations.

The theoretical possibility of multiple equilibria with self-fulfilling expectations requires that there is at least one interval on the horizontal axis where a section of the transition curve has negative slope. Let us see if we can get an intuitive understanding of why in this situation multiple equilibria can arise. Consider the specific configuration in Fig. 3.5 where k', k'', and k''' are the possible values for the capital-labor ratio next period when $k_t \in (\underline{k}, \overline{k})$. In a neighborhood of the point P associated with the intermediate value, k'', the slope of the transition curve is negative. As we saw above, this requires not only that in this neighborhood $s_r(w_t, r(k_{t+1})) < 0$, but that the stricter condition $s_r(w_t, r(k_{t+1}))$ $< (1+n)/f''(k'')$ holds (we take w_t as given since k_t is given and $w_t = w(k_t)$). That the point P with coordinates (k_t, k'') is on the transition curve indicates that given $w_t = w(k_t)$ and an expected interest rate $r^e_{t+1} = r(k'')$, the induced saving

by the young, $s(w_t, r(k''))$, will be such that $k_{t+1} = k''$, that is, the expectation is fulfilled. The fact that also the point (k_t, k'), where $k' > k''$, is on transition curve indicates that also a lower interest rate, $r(k')$, can be self-fulfilling. By this is meant that *if* an interest rate at the level $r(k')$ is expected, then this expectation induces *more* saving by the young, just enough more to make $k_{t+1} = k' > k''$, thus confirming the expectation of the lower interest rate level $r(k')$. What makes this possible is exactly the negative dependency of s_t on r^e_{t+1}. The fact that also the point (k_t, k'''), where $k''' < k''$, is on the transition curve can be similarly interpreted. It is also $s_r < 0$ that makes it possible that *less* saving by the young than at P can be induced by an expected *higher* interest rate, $r(k''')$, than at P.

These ambiguities point to a *serious problem with the assumption of perfect foresight.* The model presupposes that all the young *agree* in their expectations. Only then will one of the three mentioned temporary equilibria appear. But the model is silent about how the needed coordination of expectations is brought about, and if it is, why this coordination ends up in one rather than another of the three possible equilibria with self-fulfilling expectations. Each single young is isolated in the market and will not know what the others will expect. The market mechanism in the model provides no coordination of expectations. As it stands, the model cannot determine how the economy will evolve in this situation.

This is of course a weakness. Yet the encountered phenomenon itself − that multiple self-fulfilling equilibrium paths are theoretically possible − is certainly of interest and plays an important role in certain business cycle theories of booms and busts.

For now we plainly want to circumvent non-uniqueness. There are at least two ways to rule out the possibility of multiple equilibrium paths. One simple approach is to discard the assumption of perfect foresight. Instead, some kind of adaptive expectations may be assumed, for example in the form of *myopic foresight,* also called *static expectations.* This means that the expectation formed by the agents in the current period about the value of a variable next period is that it will stay the same as in the current period. So here the assumption would be that the young have the expectation $r^e_{t+1} = r_t$. Then, given $k_0 > 0$, a *unique* sequence of temporary equilibria $\{(k_t, c_{1t}, c_{2t}, w_t, r_t)\}_{t=0}^{\infty}$ is generated by the model. *Oscillations* in the sense of repetitive movements up and down of k_t are possible. Even *chaotic* trajectories are possible (see Exercise 3.6).

Outside steady state the agents will experience that their expectations are systematically wrong. And the assumption of myopic foresight rules out that learning occurs. This may be too simplistic, although it *can* be argued that human beings to a certain extent have a psychological disposition to myopic foresight.

Another approach to the indeterminacy problem in the Diamond model is

motivated by the presumption that the possibility of multiple equilibria is basically due to the rough time structure of the model. Each period in the model corresponds to half of an adult person's lifetime. Moreover, in the first period of life there is no capital income, in the second there is no labor income. This coarse notion of time may artificially generate a multiplicity of equilibria or, with myopic foresight, oscillations. An expanded model where people live many periods may "smooth" the responses of the system to the events impinging on it. Indeed, with working life stretching over more than one period, wealth effects of changes in the interest rate arise, thereby reducing the likelihood of a backward-bending transition curve.

Anyway, in a first approach the analyst may want to stay with a rough time structure because of its analytical convenience and then make the best of it by imposing conditions on the utility function, the production function, and/or parameter values so as to rule out multiple equilibria. Following this approach we stay with the assumption of perfect foresight, but assume that circumstances are such that multiple temporary equilibria with self-fulfilling expectations do not arise.

Conditions for uniqueness of the equilibrium path

Sufficient for the equilibrium path to be unique is that preferences and technology in combination are such that the slope of the transition curve is everywhere positive. Hence we impose the Positive Slope Assumption that

$$s_r(w(k_t), r(k_{t+1})) > \frac{1+n}{f''(k_{t+1})} \tag{A2}$$

everywhere along an equilibrium path. This condition is of course always satisfied when $s_r \geq 0$ (reflecting an elasticity of marginal utility of consumption not above one) and *can* be satisfied even if $s_r < 0$ (as long as s_r is small in absolute value). Essentially, it is an assumption that the income effect on consumption as young of a rise in the interest rate does not dominate the substitution effect "too much".

Unfortunately, a condition like (A2) is not in itself very informative. This is because it is expressed in terms of an *endogenous* variable, k_{t+1}, for given k_t. A model assumption should preferably be stated in terms of what is *given,* also called the "primitives" of the model, that is, the exogenous elements which in this model comprise the assumed preferences, demography, technology, and the market form. We can state sufficient conditions, however, in terms of the "primitives", such that (A2) is ensured. Here we state two such sufficient conditions, both involving a CRRA period utility function with parameter θ as defined in (3.14):

(a) If $0 < \theta \leq 1$, then (A2) holds for all $k_t > 0$ along an equilibrium path.

(b) If the production function is of CES-type,[17] i.e., $f(k) = A(\alpha k^{\gamma} + 1 - \alpha)^{1/\gamma}$, $A > 0$, $0 < \alpha < 1$, $-\infty < \gamma < 1$, then (A2) holds along an equilibrium path even for $\theta > 1$, if the elasticity of substitution between capital and labor, $1/(1-\gamma)$, is not too small, i.e., if

$$\frac{1}{1-\gamma} > \frac{1 - 1/\theta}{1 + (1+\rho)^{-1/\theta}(1 + f'(k) - \delta)^{(1-\theta)/\theta}} \tag{3.36}$$

for all $k > 0$. In turn, sufficient for this is that $(1-\gamma)^{-1} > 1 - \theta^{-1}$.

That (a) is sufficient for (A2) is immediately visible in (3.15). The sufficiency of (b) is proved in Appendix D. The elasticity of substitution between capital and labor is a concept analogue to the elasticity of intertemporal substitution in consumption. It is a measure of the sensitivity of the chosen $k = K/L$ with respect to the relative factor price. The next chapter goes more into detail with the concept and shows, among other things, that the Cobb-Douglas production function corresponds to $\gamma = 0$. So the Cobb-Douglas production function will satisfy the inequality $(1-\gamma)^{-1} > 1 - \theta^{-1}$ (since $\theta > 0$), hence also the inequality (3.36).

With these or other sufficient conditions in the back of our mind we shall now proceed imposing the Positive Slope Assumption (A2). To summarize:

PROPOSITION 3 *(uniqueness of an equilibrium path)* Suppose the No Fast and Positive Slope assumptions, (A1) and (A2), apply. Then, if $k_0 > 0$, there exists a unique equilibrium path.
(i) if $k_0 > 0$, there exists a unique equilibrium path;
(ii) if $k_0 = 0$, an equilibrium path exists if and only if $f(0) > 0$ (i.e., capital not essential).

When the conditions of Proposition 3 hold, the fundamental difference equation, (3.32), of the model defines k_{t+1} as an implicit function of k_t,

$$k_{t+1} = \varphi(k_t),$$

for all $k_t > 0$, where $\varphi(k_t)$ is called a *transition function.* The derivative of this implicit function is given by (3.34) with k_{t+1} on the right-hand side replaced by $\varphi(k_t)$, i.e.,

$$\varphi'(k_t) = \frac{s_w\left(w\left(k_t\right), r\left(\varphi(k_t)\right)\right) w'(k_t)}{1 + n - s_r\left(w\left(k_t\right), r\left(\varphi(k_t)\right)\right) r'(\varphi(k_t))} > 0. \tag{3.37}$$

The positivity for all $k_t > 0$ is due to (A2). Example 2 above leads to a transition function.

[17] CES stands for Constant Elasticity of Substitution. CES production functions are considered in detail in Chapter 4.

Having determined the evolution of k_t, we have in fact determined the evolution of "everything" in the economy: the factor prices $w(k_t)$ and $r(k_t)$, the saving of the young $s_t = s(w(k_t), r(k_{t+1}))$, and the consumption by both the young and the old. The mechanism behind the evolution of the economy is the Walrasian (or Classical) mechanism where prices, here w_t and r_t, always adjust so as to generate market clearing as if there were a Walrasian auctioneer and where expectations always adjust so as to be model consistent.

Existence and stability of a steady state?

Possibly the equilibrium path converges to a steady state. To address this issue, we examine the possible configurations of the transition curve in more detail. In addition to being positively sloped everywhere, the transition curve will always, for $k_t > 0$, be situated strictly below the solid curve, $k_{t+1} = w(k_t)/(1+n)$, shown in Fig. 3.6. In turn, the latter curve is always, for $k_t > 0$, strictly below the stippled curve, $k_{t+1} = f(k_t)/(1+n)$, in the figure. To be precise:

LEMMA 3 *(ceiling and roof)* Suppose the No Fast Assumption (A1) applies. Along an equilibrium path, whenever $k_t > 0$,

$$0 < k_{t+1} < \frac{w(k_t)}{1+n} < \frac{f(k_t)}{1+n}, \qquad t = 0, 1, \ldots.$$

Proof. From (iii) of Proposition 2, an equilibrium path has $w_t = w(k_t) > 0$ and $k_{t+1} > 0$ for $t = 0, 1, 2, \ldots$. Thus,

$$0 < k_{t+1} = \frac{s_t}{1+n} < \frac{w_t}{1+n} = \frac{w(k_t)}{1+n} = \frac{f(k_t) - f'(k_t)k_t}{1+n} < \frac{f(k_t)}{1+n},$$

where the first equality comes from (3.32), the second inequality from Lemma 1 in Section 3.3, and the last inequality from the fact that $f'(k_t)k_t > 0$ when $k_t > 0$. □

We will call the graph $(k_t, w(k_t)/(1+n))$ in Fig. 3.6 a *ceiling*. It acts as a ceiling on k_{t+1} simply because the saving of the young cannot exceed the income of the young, $w(k_t)$. And we will call the graph $(k_t, f(k_t)/(1+n))$ a *roof,* because "everything of interest" occurs below it. The roof can be drawn directly on the basis of the production function $f(k_t)$.

To characterize the position of the roof relative to the 45° line, we consider the lower Inada condition, $\lim_{k\to 0} f'(k) = \infty$.

LEMMA 4 The roof, $\mathcal{R}(k) \equiv f(k)/(1+n)$, has positive slope everywhere, crosses the 45° line for at most one $k > 0$ and can only do that from above. A necessary and sufficient condition for the *roof* to be above the 45° line for small k is that either $\lim_{k\to 0} f'(k) > 1 + n$ or $f(0) > 0$ (capital not essential).

Proof. Since $f' > 0$, the roof has positive slope. Since $f'' < 0$, it can only cross the 45° line once and only from above. If and only if $\lim_{k\to 0} f'(k) > 1 + n$, then for small k_t, the roof is steeper than the 45° line. Obviously, if $f(0) > 0$, then close to the origin, the roof will be above the 45° line. □

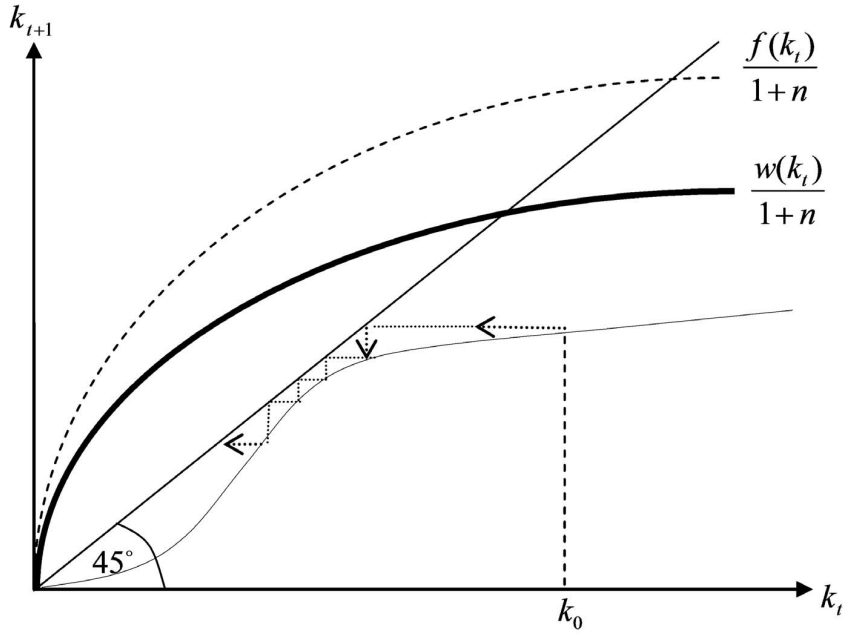

Figure 3.6: A case where both the roof and the ceiling cross the 45° line, but the transition curve does not (no steady state exists).

LEMMA 5 Given $w(k) = f(k) - f'(k)k$ for all $k \geq 0$, where $f(k)$ satisfies $f(0) \geq 0$, $f' > 0$, $f'' < 0$, the following holds:
(i) $\lim_{k\to\infty} w(k)/k = 0$;
(ii) the ceiling, $\mathcal{C}(k) \equiv w(k)/(1 + n)$, is positive and has positive slope for all $k > 0$; moreover, there exists $\bar{k} > 0$ such that $\mathcal{C}(k) < k$ for all $k > \bar{k}$.

Proof. (i) In view of $f(0) \geq 0$ combined with $f'' < 0$, we have $w(k) > 0$ for all $k > 0$. Hence, $\lim_{k\to\infty} w(k)/k \geq 0$ if this limit exists. Consider an arbitrary $k_1 > 0$. We have $f'(k_1) > 0$. For all $k > k_1$, it holds that $0 < f'(k) < f'(k_1)$, in view of $f' > 0$ and $f'' < 0$, respectively. Hence, $\lim_{k\to\infty} f'(k)$ exists and

$$0 \leq \lim_{k\to\infty} f'(k) < f'(k_1). \tag{3.38}$$

We have

$$\lim_{k\to\infty} \frac{w(k)}{k} = \lim_{k\to\infty} \frac{f(k)}{k} - \lim_{k\to\infty} f'(k). \tag{3.39}$$

There are two cases to consider. *Case 1:* $f(k)$ has an upper bound. Then, $\lim_{k\to\infty} f(k)/k = 0$ so that $\lim_{k\to\infty} w(k)/k = -\lim_{k\to\infty} f'(k) = 0$, by (3.39) and (3.38), as $w(k)/k > 0$ for all $k > 0$. *Case 2:* $\lim_{k\to\infty} f(k) = \infty$. Then, by L'Hôpital's rule for "∞/∞", $\lim_{k\to\infty}(f(k)/k) = \lim_{k\to\infty} f'(k)$ so that (3.39) implies $\lim_{k\to\infty} w(k)/k = 0$.

(ii) As $n > -1$ and $w(k) > 0$ for all $k > 0$, $\mathcal{C}(k) > 0$ for all $k > 0$. From $w'(k) = -kf''(k) > 0$ follows $\mathcal{C}'(k) = -kf''(k)/(1+n) > 0$ for all $k > 0$; that is, the ceiling has positive slope everywhere. For $k > 0$, the inequality $\mathcal{C}(k) < k$ is equivalent to $w(k)/k < 1+n$. By (i) follows that for all $\varepsilon > 0$, there exists $k_\varepsilon > 0$ such that $w(k)/k < \varepsilon$ for all $k > k_\varepsilon$. Now, letting $\varepsilon = 1+n$ and $\bar{k} = k_\varepsilon$ proves that there exists $\bar{k} > 0$ such that $w(k)/k < 1+n$ for all $k > \bar{k}$. □

While the roof can be above the 45° line for all $k_t > 0$, the ceiling can not. Indeed, (ii) of the lemma implies that if for small k_t the ceiling is above the 45° line, the ceiling will necessarily cross the 45° line at least once for larger k_t.

In view of the ceiling being always an upper bound on k_{t+1}, what is the point of introducing also the roof? The point is that the roof is a more straightforward construct since it is directly given by the production function and is always strictly concave. The ceiling is generally a more complex construct. It can have convex sections and for instance cross the 45° line at more than one point if at all. .

A necessary condition for existence of a (non-trivial) steady state is that the roof is above the 45^0 line for small k_t. But this is not sufficient for also the transition curve to be above the 45^0 line for small k_t. Fig. 3.6 illustrates this. Here the transition curve is in fact everywhere below the 45^0 line. In this case no steady state exists and the dynamics imply convergence towards the "catastrophic" point $(0,0)$. Given the rate of population growth, the saving of the young is not sufficient to avoid famine in the long run. This will for example happen if the technology implies so low productivity that even if all income of the young were saved, we would have $k_{t+1} < k_t$ for all $k_t > 0$, cf. Exercise 3.2. The Malthusian mechanism will be at work and bring down n (outside the model). This exemplifies that even a trivial steady state (the point (0,0)) may be of interest in so far as it may be the point the economy is heading to without ever reaching it.

To help existence of a steady state we will impose the condition that either capital is not essential or preferences and technology fit together in such a way that the slope of the transition curve is larger than one for small k_t. That is, we assume that either

$$\begin{array}{lrcll} \text{(i)} & f(0) & > & 0 & \text{or} \\ \text{(ii)} & \lim\limits_{k\to 0} \varphi'(k) & > & 1, & \end{array} \tag{A3}$$

where $\varphi'(k)$ is implicitly given in (3.37). Whether condition (i) of (A3) holds in a given situation can be directly checked from the production function. If it does

not, we should check condition (ii). But this condition is less amenable because the transition function φ is not one of the "primitives" of the model. There exist cases, though, where we can find an explicit transition function and try out whether (ii) holds (like in Example 2 above). But generally we can not. Then we have to resort to *sufficient* conditions for (ii) of (A3), expressed in terms of the "primitives". For example, if the period utility function belongs to the CRRA class and the production function is Cobb-Douglas at least for small k, then (ii) of (A3) holds (see Appendix E). Anyway, as (i) and (ii) of (A3) can be interpreted as reflecting two different kinds of "early steepness" of the transition curve, we shall call (A3) the Early Steepness Assumption.[18]

PROPOSITION 4 *(existence and stability of a steady state)* Assume that the No Fast Assumption (A1) and the Positive Slope assumption (A2) apply as well as the Early Steepness Assumption (A3). Then there exists at least one steady state $k^* > 0$ that is locally asymptotically stable. Oscillations do not occur.

Proof. By (A1), Lemma 3 applies. From Proposition 2 we know that if (i) of (A3) holds, then $k_{t+1} = s_t/(1+n) > 0$ even for $k_t = 0$. Alternatively, (ii) of (A3) is enough to ensure that the transition curve lies above the 45° line for small k_t. By Lemma 4 the roof then does the same. According to (ii) of Lemma 5, for large k_t the ceiling is below the 45° line. Being below the ceiling, cf. Lemma 3, the transition curve must therefore cross the 45° line at least once. Let k^* denote the smallest k_t at which it crosses. Then $k^* > 0$ is a steady state with the property $0 < \varphi'(k^*) < 1$. By graphical inspection we see that this steady state is asymptotically stable. For oscillations to come about there must exist a steady state, k^{**}, with $\varphi'(k^{**}) < 0$, but this is impossible in view of (A2). □

From Proposition 4 we conclude that, given k_0, the assumptions (A1) - (A3) ensure existence and uniqueness of an equilibrium path; moreover, the equilibrium path converges towards *some* steady state. Thus with these assumptions, for any $k_0 > 0$, sooner or later the system settles down at some steady state $k^* > 0$. For the factor prices we therefore have

$$\begin{aligned} r_t &= f'(k_t) - \delta \to f'(k^*) - \delta \equiv r^*, \quad \text{and} \\ w_t &= f(k_t) - k_t f'(k_t) \to f(k^*) - k^* f'(k^*) \equiv w^*, \end{aligned}$$

for $t \to \infty$. But there may be more than one steady state and therefore only *local* stability is guaranteed. This can be shown by examples, where the utility function, the production function, and parameters are specified in accordance with the assumptions (A1) - (A3) (see Exercise 3.5 and ...).

[18] In (i) of (A3), the "steepness" is rather a "hop" at $k = 0$ if we imagine k approaching nil from below.

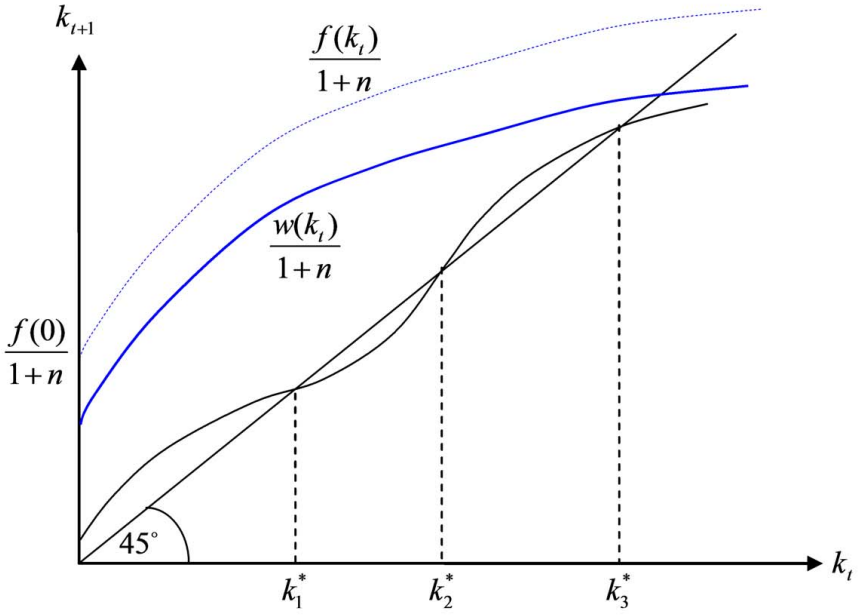

Figure 3.7: A case of multiple steady states (and capital being not essential).

Fig. 3.7 illustrates such a case (with $f(0) > 0$ so that capital is not essential). Moving West-East in the figure, the first steady state, k_1^*, is stable, the second, k_2^*, unstable, and the third, k_3^*, stable. In which of the two stable steady states the economy ends up depends on the initial capital-labor ratio, k_0. The lower steady state, k_1^*, is known as a *poverty trap*. If $0 < k_0 < k_2^*$, the economy is caught in the trap and converges to the low steady state. But with high enough k_0 ($k_0 > k_2^*$), perhaps obtained by foreign aid, the economy avoids the trap and converges to the high steady state. Looking back at Fig. 3.6, we can interpret that figure's scenario as exhibiting an *inescapable* poverty trap.

It turns out that CRRA utility combined with a Cobb-Douglas production function ensures both that (A1) - (A3) hold and that a *unique* (non-trivial) steady state exists. So in this case *global* asymptotic stability of the steady state is ensured.[19] Example 2 and Fig. 3.4 above display a special case of this, the case $\theta = 1$.

This is of course a convenient case for the analyst. A Diamond model satisfying assumptions (A1) - (A3) *and* featuring a unique steady state is called a *well-behaved* Diamond model.

We end this section with the question: Is it possible that aggregate consumption, along an equilibrium path, for some periods exceeds aggregate income? We

[19] See last section of Appendix E.

shall see that this is indeed possible in the model if K_0 (wealth of the old in the initial period) is large enough. Indeed, from national accounting we have:

$$\begin{aligned} C_{10} + C_{20} &= F(K_0, L_0) - I_0 > F(K_0, L_0) \Leftrightarrow I_0 < 0 \\ &\Leftrightarrow K_1 < (1-\delta)K_0 \Leftrightarrow K_0 - K_1 > \delta K_0. \end{aligned}$$

So aggregate consumption in period 0 being greater than aggregate income is equivalent to a fall in the capital stock from period 0 to period 1 greater than the capital depreciation in period 0. Consider the log utility Cobb-Douglas case in Fig. 3.4 and suppose $\delta < 1$ and $L_t = L_0 = 1$, i.e., $n = 0$. Then $k_t = K_t$ for all t and by (3.35), $K_{t+1} = \frac{(1-\alpha)A}{2+\rho} K_t^\alpha$. Thus $K_1 < (1-\delta)K_0$ for

$$K_0 > \left(\frac{(1-\alpha)A}{(2+\rho)(1-\delta)} \right)^{1/(1-\alpha)}.$$

As initial K is arbitrary, this situation is possible. When it occurs, it reflects that the financial wealth of the old is so large that their consumption (recall they consume all their financial wealth as well as the interest on this wealth) exceeds what is left of current aggregate production after subtracting the amount consumed by the young. So aggregate gross investment in the economy will be negative. Of course this is only feasible if capital goods can be "eaten" or at least be immediately (without further resources) converted into consumption goods. As it stands, the model has implicitly assumed this to be the case. And this is in line with the general setup since the output good is homogeneous and can either be consumed or piled up as capital.

We now turn to efficiency problems.

3.6 The golden rule and dynamic inefficiency

An economy described by the Diamond model has the property that even though there is perfect competition and no externalities, the outcome brought about by the market mechanism may not be Pareto optimal.[20] Indeed, the economy may *overaccumulate* forever and thus suffer from a distinctive form of production inefficiency.

A key element in understanding the concept of overaccumulation is the concept of a *golden-rule capital-labor ratio*. Overaccumulation occurs when aggregate

[20] Recall that a *Pareto optimal* path is a technically feasible path with the property that no other technically feasible path will make at least one individual better off without making someone else worse off. A technically feasible path which is not Pareto optimal is called *Pareto inferior*.

saving maintains a capital-labor ratio above the golden-rule value forever. Let us consider these concepts in detail.

In the present section generally the period length is arbitrary except when we relate to the Diamond model and the period length therefore is half of adult lifetime.

The golden-rule capital-labor ratio

The golden rule is a principle that relates to technically feasible paths. The principle does not depend on the market form.

Consider the economy-wide resource constraint $C_t = Y_t - S_t = F(K_t, L_t) - (K_{t+1} - K_t + \delta K_t)$, where we assume that F is neoclassical with CRS. Accordingly, aggregate consumption per unit of labor can be written

$$c_t \equiv \frac{C_t}{L_t} = \frac{F(K_t, L_t) - (K_{t+1} - K_t + \delta K_t)}{L_t} = f(k_t) + (1-\delta)k_t - (1+n)k_{t+1}, \tag{3.40}$$

where k is the capital-labor ratio K/L. Note that C_t will generally be greater than the workers' consumption. One should simply think of C_t as the flow of produced consumption goods in the economy and c_t as this flow divided by aggregate employment, including the labor that in period t produces investment goods. How the consumption goods are distributed to different members of society is not our concern here.

DEFINITION 5 By the *golden-rule capital-labor ratio*, k_{GR}, is meant that value of the capital-labor ratio k, which results in the highest possible sustainable level of consumption per unit of labor.

Sustainability requires replicability forever. We therefore consider a steady state. In a steady state $k_{t+1} = k_t = k$ so that (3.40) simplifies to

$$c = f(k) - (\delta + n)k \equiv c(k). \tag{3.41}$$

Maximizing gives the first-order condition

$$c'(k) = f'(k) - (\delta + n) = 0. \tag{3.42}$$

In view of $c''(k) = f''(k) < 0$, the condition (3.42) is both necessary and sufficient for an interior maximum. Let us assume that $\delta + n > 0$ and that f satisfies the condition

$$\lim_{k\to\infty} f'(k) < \delta + n < \lim_{k\to 0} f'(k).$$

Then (3.42) has a solution in k, and it is unique because $c''(k) < 0$. The solution is called k_{GR} so that

$$f'(k_{GR}) - \delta = n.$$

That is:

PROPOSITION 5 *(the golden rule)* The highest sustainable consumption level per unit of labor in society is obtained when in steady state the net marginal productivity of capital equals the growth rate of the economy.

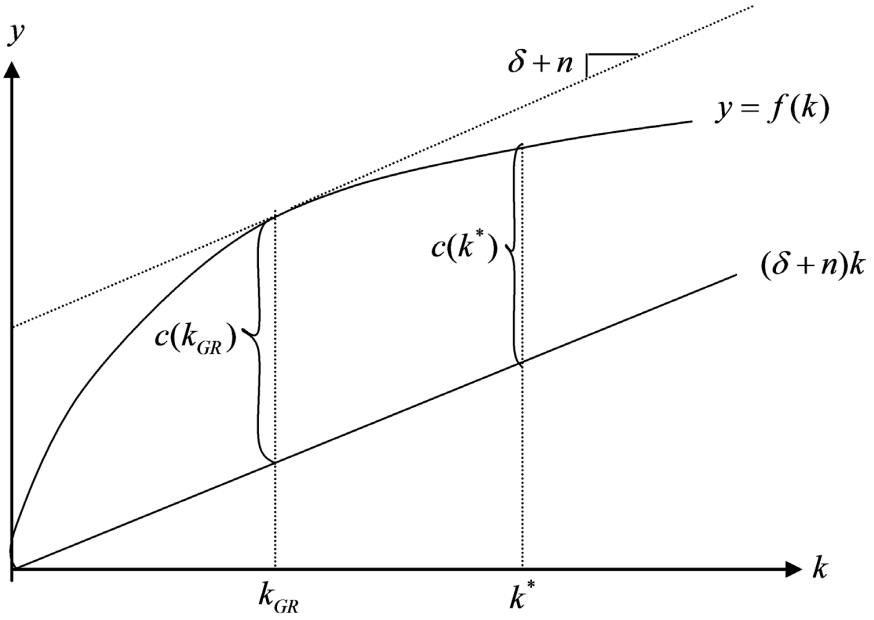

Figure 3.8: A steady state with overaccumulation.

It follows that if a society aims at the highest sustainable level of consumption and initially has $k_0 < k_{GR}$, it should increase its capital-labor ratio up to the point where the extra output obtainable by a further small increase is exactly offset by the extra gross investment needed to maintain the capital-labor ratio at that level. The intuition is visible from (3.41). The golden-rule capital-labor ratio, k_{GR}, strikes the right balance in the trade-off between high output per unit of labor and a not too high investment requirement. Although a steady state with $k > k_{GR}$ would imply higher output per unit of labor, it would also imply that a large part of that output is set aside for investment (namely the amount $(\delta + n)k$ per unit of labor) to counterbalance capital depreciation and growth in the labor force; without this investment the high capital-labor ratio k^* would not be maintained. With $k > k_{GR}$ this feature would dominate the first effect so that consumption per unit of labor ends up low. Fig. 3.8 illustrates.

The name golden rule hints at the golden rule from the Bible: "Do unto others as you would have them to do unto you." We imagine that God asks the newly born generation: "What capital-labor ratio would you prefer to be presented

with, given that you must hand over the same capital-labor ratio to the next generation?" The appropriate answer is: the golden-rule capital-labor ratio.

The possibility of overaccumulation in a competitive market economy

The equilibrium path in the Diamond model with perfect competition implies an interest rate $r^* = f'(k^*) - \delta$ in a steady state. As an implication,

$$r^* \gtreqless n \Leftrightarrow f'(k^*) - \delta \gtreqless n \Leftrightarrow k^* \lesseqgtr k_{GR}, \text{ respectively,}$$

in view of $f'' < 0$. Hence, a long-run interest rate below the growth rate of the economy indicates that $k^* > k_{GR}$. This amounts to a Pareto-inferior state of affairs. Indeed, everyone can be made better off if by a coordinated reduction of saving and investment, k is reduced. A formal demonstration of this is given in connection with Proposition 6 in the next subsection. Here we give an account in more intuitive terms.

Consider Fig. 3.8. Let k be gradually reduced to the level k_{GR} by refraining from investment in period t_0 and forward until this level is reached. When this happens, let k be maintained at the level k_{GR} forever by providing for the needed investment per young, $(\delta + n)k_{GR}$. Then there would be higher aggregate consumption in period t_0 and every future period. Both the immediate reduction of saving and a resulting lower capital-labor ratio to be maintained contribute to this result. There is thus scope for both young and old to consume more in every future period.

In the Diamond model a simple policy implementing such a Pareto improvement in the case where $k^* > k_{GR}$ (i.e., $r^* < n$) is to incur a lump-sum tax on the young, the revenue of which is immediately transferred lump sum to the old, hence, fully consumed. Suppose this amounts to a transfer of one good from each young to the old. Since there are $1 + n$ young people for each old person, every old receives in this way $1 + n$ goods in the same period. Let this transfer be repeated every future period. By decreasing their saving by one unit, the young can maintain unchanged consumption in their youth, and when becoming old, they receive $1 + n$ goods from the next period's young and so on. In effect, the "return" on the tax payment by the young is $1 + n$ next period. This is more than the $1 + r^*$ that could be obtained via the market through own saving.[21]

[21] In this model with no utility of leisure, a tax on wage income, or a mandatory pay-as-you-go pension contribution (see Chapter 5) would act like a lump-sum tax on the young.

The described tax-transfers policy will affect the equilibrium interest rate negatively. By choosing an appropriate size of the tax this policy, combined with competitive markets, will under certain conditions (see Chapter 5.1) bring the economy to the golden-rule steady state where overaccumulation has ceased and $r^* = n$.

A proof that $k^* > k_{GR}$ is indeed theoretically possible in the Diamond model can be based on the log utility-Cobb-Douglas case from Example 2 in Section 3.5.3. As indicated by the formula for r^* in that example, the outcome $r^* < n$, which is equivalent to $k^* > k_{GR}$, can always be obtained by making the parameter $\alpha \in (0,1)$ in the Cobb-Douglas function small enough. The intuition is that a small α implies a high $1-\alpha$, that is, a high wage income $wL = (1-\alpha)K^{\alpha}L^{-\alpha} \cdot L$ $= (1-\alpha)Y$; this leads to high saving by the young, since $s_w > 0$. The result is a high k_{t+1} which generates a high real wage also next period and may in this manner be sustained forever.

An intuitive understanding of the fact that the perfectly competitive market mechanism may thus lead to overaccumulation, can be based on the following argument. Assume, first, that $s_r < 0$. In this case, if the young in period t expects the rate of return on their saving to end up small (less than n), the decided saving will be large in order to provide for consumption after retirement. But the aggregate result of this behavior is a high k_{t+1} and therefore a low $f'(k_{t+1})$. In this way the expectation of a low r_{t+1} is confirmed by the actual events. The young persons each do the best they can as atomistic individuals, taking the market conditions as given. Yet the aggregate outcome is an equilibrium with overaccumulation, hence a Pareto-inferior outcome.

Looking at the issue more closely, we see that $s_r < 0$ is not crucial for this outcome. Suppose $s_r = 0$ (the log utility case) and that in the current period, k_t is, for some historical reason, at least temporarily considerably above k_{GR}. Thus, current wages are high, hence, s_t is relatively high (there is in this case no offsetting effect on s_t from the relatively low expected r_{t+1}). Again, the aggregate result is a high k_{t+1} and thus the expectation is confirmed. Consequently, the situation in the next period is the same and so on. By continuity, even if $s_r > 0$, the argument goes through as long as s_r is not too large.

Dynamic inefficiency and the double infinity

Another name for the overaccumulation phenomenon is *dynamic inefficiency.*

DEFINITION 6 A technically feasible path $\{(c_t, k_t)\}_{t=0}^{\infty}$ with the property that there does not exist another technically feasible path with higher c_t in some periods without smaller c_t in other periods is called *dynamically efficient.* A technically feasible path $\{(c_t, k_t)\}_{t=0}^{\infty}$ which is not dynamically efficient is called *dynamically inefficient.*

PROPOSITION 6 A technically feasible path $\{(c_t, k_t)\}_{t=0}^{\infty}$ with the property that for $t \to \infty$, $k_t \to k^* > k_{GR}$, is dynamically inefficient.

Proof. Let $k^* > k_{GR}$. Then there exists an $\varepsilon > 0$ such that $k \in (k^* - 2\varepsilon, k^* + 2\varepsilon)$

implies $f'(k) - \delta < n$ since $f'' < 0$. By concavity of f,

$$f(k) - f(k-\varepsilon) \leq f'(k-\varepsilon)\varepsilon. \tag{3.43}$$

Consider a technically feasible path $\{(c_t, k_t)\}_{t=0}^{\infty}$ with $k_t \to k^*$ for $t \to \infty$ (the reference path). Then there exists a t_0 such that for $t \geq t_0$, $k_t \in (k^* - \varepsilon, k^* + \varepsilon)$, $f'(k_t) - \delta < n$ and $f'(k_t - \varepsilon) - \delta < n$. Consider an alternative feasible path $\left\{(\hat{c}_t, \hat{k}_t)\right\}_{t=0}^{\infty}$, where a) for $t = t_0$ consumption is increased relative to the reference path such that $\hat{k}_{t_0+1} = k_{t_0} - \varepsilon$; and b) for all $t > t_0$, consumption is such that $\hat{k}_{t+1} = k_t - \varepsilon$. We now show that after period t_0, $\hat{c}_t > c_t$. Indeed, for all $t > t_0$, by (3.40),

$$\begin{aligned}
\hat{c}_t &= f(\hat{k}_t) + (1-\delta)\hat{k}_t - (1+n)\hat{k}_{t+1} \\
&= f(k_t - \varepsilon) + (1-\delta)(k_t - \varepsilon) - (1+n)(k_{t+1} - \varepsilon) \\
&\geq f(k_t) - f'(k_t - \varepsilon)\varepsilon + (1-\delta)(k_t - \varepsilon) - (1+n)(k_{t+1} - \varepsilon) \quad \text{(by (3.43))} \\
&> f(k_t) - (\delta + n)\varepsilon + (1-\delta)k_t - (1+n)k_{t+1} + (\delta + n)\varepsilon \\
&= f(k_t) + (1-\delta)k_t - (1+n)k_{t+1} = c_t,
\end{aligned}$$

by (3.40). □

Moreover, it can be shown[22] that:

PROPOSITION 7 A technically feasible path $\{(c_t, k_t)\}_{t=0}^{\infty}$ such that for $t \to \infty$, $k_t \to k^* \leq k_{GR}$, is dynamically efficient.

Accordingly, a steady state with $k^* < k_{GR}$ is never dynamically inefficient. This is because increasing k from this level always has its price in terms of a decrease in *current* consumption; and at the same time decreasing k from this level always has its price in terms of lost *future* consumption. But a steady state with $k^* > k_{GR}$ is always dynamically inefficient. Intuitively, staying forever with $k = k^* > k_{GR}$, implies that society *never* enjoys its great capacity for producing consumption goods.

The fact that $k^* > k_{GR}$, and therefore dynamic inefficiency, cannot be ruled out might seem to contradict the First Welfare Theorem from the microeconomic theory of general equilibrium. This is the theorem saying that under certain conditions (essentially that increasing returns to scale are absent are absent, markets are competitive, no goods are of public good character, and there are no externalities, then market equilibria are Pareto optimal. In fact, however, the First Welfare Theorem also presupposes a finite number of periods or, if the number of periods is infinite, then a finite number of agents. In contrast, in the OLG model

[22] See Cass (1972).

there is a *double infinity*: an infinite number of periods *and* agents. Hence, the First Welfare Theorem breaks down. Indeed, the case $r^* < n$, i.e., $k^* > k_{GR}$, can arise under *laissez-faire*. Then, as we have seen, everyone can be made better off by a coordinated intervention by some social arrangement (a government for instance) such that k is reduced.

The essence of the matter is that the double infinity opens up for technically feasible reallocations which are definitely beneficial when $r^* < n$ and which a central authority can accomplish but the market can not. That *nobody* need loose by the described kind of redistribution is due to the double infinity: the economy goes on forever and there is no last generation. Nonetheless, some kind of centralized *coordination* is required to accomplish a solution.

There is an analogy in "Gamow's bed problem": There are an infinite number of inns along the road, each with one bed. On a certain rainy night all innkeepers have committed their beds. A late guest comes to the first inn and asks for a bed. "Sorry, full up!" But the minister of welfare hears about it and suggests that each incumbent guest move down the road one inn.[23]

Whether the theoretical possibility of overaccumulation should be a matter of practical concern is an empirical question about the relative size of rates of return and economic growth. To answer the question meaningfully, we need an extension of the criterion for overaccumulation so that the presence of technological progress and rising per capita consumption in the long run can be taken into account. This is one of the topics of the next chapter. At any rate, we can already here reveal that there exists no indication that overaccumulation has ever been an actual problem in industrialized market economies.

A final remark before concluding. Proposition 5 about the golden rule can be generalized to the case where instead of one there are n different capital goods in the economy. Essentially the generalization says that assuming CRS-neoclassical production functions with n different capital goods as inputs, one consumption good, no technological change, and perfectly competitive markets, a steady state in which per-unit-of labor consumption is maximized has interest rate equal to the growth rate of the labor force when technological progress is ignored (see, e.g., Mas-Colell, 1989).

3.7 Concluding remarks

(Unfinished)

In several respects the conclusions we get from OLG models are different than those from representative agent models to be studied later. In OLG models the

[23] George Gamow (1904-1968) was a Russian physicist. The problem is also known as *Hilbert's hotel problem*, after the German mathematician David Hilbert (1862-1943).

aggregate quantities are the outcome of the interplay of finite-lived agents at different stages in their life cycle. The turnover in the population plays a crucial role. In this way the OLG approach lays bare the possibility of coordination failure on a grand scale. In contrast, in a representative agent model, aggregate quantities are just a multiple of the actions of the representative household.

Regarding analytical tractability, the complexity implied by having in every period two different coexisting generations is in some respects more than compensated by the fact that the finite time horizon of the households make the *dynamics* of the model *one-dimensional*: we end up with a first-order difference equation in the capital-labor ratio, k_t, in the economy. In contrast, the dynamics of the basic representative agent model (Chapter 8 and 10) is two-dimensional (owing to the assumed infinite horizon of the households considered as dynasties).

Miscellaneous notes:

OLG gives theoretical insights concerning macroeconomic implications of life cycle behavior, allows heterogeneity, provides training in seeing the economy as consisting of a heterogeneous population where the *distribution* of agent characteristics matters for the aggregate outcome.

Farmer (1993), p. 125, notes that OLG models are difficult to apply and for this reason much empirical work in applied general equilibrium theory has regrettably instead taken the representative agent approach.

Outlook: Rational speculative bubbles in general equilibrium, cf. Chapter ?.

3.8 Literature notes

1. The Nobel Laureate Paul A. Samuelson (1915-2009) is one of the pioneers of OLG models. Building on the French economist and Nobel laureate Maurice Allais (1911-2010), Samuelson's famous article, Samuelson (1958), was concerned with a missing market problem. Imagine a two-period OLG economy where, as in the Diamond model, only the young have an income (which in turn is, by Samuelson, assumed exogenous). Contrary to the Diamond model, however, there is neither capital nor other stores of value. Then, in the laissez-faire market economy the old have to starve. This is clearly a Pareto-inferior allocation; if each member of the young generation hands over to the old generation one unit of account, and this is repeated every period, everyone will be better off. Since for every old there are $1 + n$ young, the implied rate of return would be n, the population growth rate. Such transfers do not arise under laissez-faire. A kind of social contract is required. As Samuelson pointed out, a government could in period 0 issue paper notes, "money", and transfer them to the members of the old generation who would then use them to buy goods from the young. Provided the young believed the notes to be valuable in the next period, they would accept

them in exchange for some of their goods in order to use them in the next period for buying from the new young generation and so on. We have here an example of how a social institution can solve a coordination problem. This gives a flavour of Samuelson's contribution although in his original article he assumed three periods of life.

2. Diamond (1965) extended Samuelson's contribution by adding capital accumulation. Because of its antecedents Diamonds OLG model is sometimes called the Samuelson-Diamond model or the Allais-Samuelson-Diamond model. In our exposition we have drawn upon clarifications by Galor and Ryder (1989) and de la Croix and Michel (2002). The last mentioned contribution is an extensive exploration of discrete-time OLG models and their applications.

3. The *life-cycle saving hypothesis* was put forward by Franco Modigliani (1918-2003) and associates in the 1950s. See for example Modigliani and Brumberg (1954). Numerous extensions of the framework, relating to the motives (b) - (e) in the list of Section 3.1, see for instance de la Croix and Michel (2002).

4. A review of the empirics of life-cycle behavior and attempts at refining life-cycle models are given in Browning and Crossley (2001).

5. Regarding the dynamic efficiency issue, both the propositions 6 and 7 were shown in a stronger form by the American economist David Cass (1937-2008). Cass established the *general* necessary and sufficient condition for a feasible path $\{(c_t, k_t)\}_{t=0}^{\infty}$ to be dynamically efficient (Cass 1972). Our propositions 6 and 7 are more restrictive in that they are limited to paths that converge. Partly intuitive expositions of the deeper aspects of the theory are given by Shell (1971) and Burmeister (1980).

6. Diamond has also contributed to other fields of economics, including search theory for labor markets. In 2010 Diamond, together with Dale Mortensen and Christopher Pissarides, was awarded the Nobel price in economics.

From here very incomplete:

The two-period structure of Diamonds OLG model leaves little room for considering, e.g., education and dissaving in the early years of life. This kind of issues is taken up in three-period extensions of the Diamond model, see ...

Multiple equilibria, self-fulfilling expectations, optimism and pessimism..

Dynamic inefficiency, see also Burmeister (1980).

Bewley 1977, 1980.

Two-sector OLG: Galor (1992). Galor's book??

On the golden rule in a general setup, see Mas-Colell (1989).

3.9 Appendix

A. On the CRRA utility function

Derivation of the CRRA function Consider a utility function $u(c)$, defined for all $c > 0$ and satisfying $u'(c) > 0$, $u''(c) < 0$. Let the absolute value of the elasticity of marginal utility be denoted $\theta(c)$, that is, $\theta(c) \equiv -cu''(c)/u'(c) > 0$. We claim that if $\theta(c)$ is a positive constant, θ, then up to a positive linear transformation $u(c)$ must be of the form

$$u(c) = \begin{cases} \frac{c^{1-\theta}}{1-\theta}, & \text{when } \theta \neq 1, \\ \ln c, & \text{when } \theta = 1, \end{cases} \qquad (*)$$

i.e., of CRRA form.

Proof. Suppose $\theta(c) = \theta > 0$. Then, $u''(c)/u'(c) = -\theta/c$. By integration, $\ln u'(c) = -\theta \ln c + A$, where A is an arbitrary constant. Take the antilogarithm function on both sides to get $u'(c) = e^A e^{-\theta \ln c} = e^A c^{-\theta}$. By integration we get

$$u(c) = \begin{cases} e^A \frac{c^{1-\theta}}{1-\theta} + B, & \text{when } \theta \neq 1, \\ e^A \ln c + B, & \text{when } \theta = 1, \end{cases}$$

where B is an arbitrary constant. This proves the claim. Letting $A = B = 0$, we get (*). □

When we want to make the kinship between the members of the CRRA family transparent, we maintain $A = 0$ and for $\theta = 1$ also $B = 0$, whereas for $\theta \neq 1$ we set $B = -1/(1-\theta)$. In this way we achieve that all members of the CRRA family will be represented by curves going through the same point as the log function, namely the point $(1, 0)$, cf. Fig. 3.2. And adding or subtracting a constant does not affect marginal rates of substitution and consequently not behavior.

The domain of the CRRA function We want to extend the domain to include $c = 0$. If $\theta \geq 1$, the CRRA function, whether in the form $u(c) = (c^{1-\theta} - 1)/(1-\theta)$ or in the form (*), is defined only for $c > 0$, not for $c = 0$. This is because for $c \to 0$ we get $u(c) \to -\infty$. In this case we simply define $u(0) = -\infty$. This will create no problems since the CRRA function anyway has the property that $u'(c) \to \infty$, when $c \to 0$ (whether θ is larger or smaller than one). The marginal utility thus becomes very large as c becomes very small, that is, the No Fast Assumption is satisfied. This will ensure that the chosen c is strictly positive whenever there is a positive budget. So throughout this book we define the domain of the CRRA function to be $[0, \infty)$.

The range of the CRRA function Considering the CRRA function $u(c) \equiv \left(c^{1-\theta}-1\right)(1-\theta)^{-1}$ for $c \in [0,\infty)$, we have:

$$\begin{aligned} \text{for } 0 &< \theta < 1, \text{ the range of } u(c) \text{ is } \left[-(1-\theta)^{-1}, \infty\right), \\ \text{for } \theta &= 1, \text{ the range of } u(c) \text{ is } [-\infty, \infty), \\ \text{for } \theta &> 1, \text{ the range of } u(c) \text{ is } [-\infty, -(1-\theta)^{-1}). \end{aligned}$$

Thus, in the latter case $u(c)$ is bounded from above and so allows asymptotic "saturation" to occur.

B. Deriving the elasticity of intertemporal substitution in consumption

Referring to Section 3.3, we here show that the definition of $\sigma(c_1, c_2)$ in (3.17) gives the result (3.18). Let $x \equiv c_2/c_1$ and $\beta \equiv (1+\rho)^{-1}$. Then the first-order condition (3.16) and the equation describing the considered indifference curve constitute a system of two equations

$$\begin{aligned} u'(c_1) &= \beta u'(xc_1)R, \\ u(c_1) + \beta u(xc_1) &= \bar{U}. \end{aligned}$$

For a fixed utility level $U = \bar{U}$ these equations define c_1 and x as implicit functions of R, $c_1 = c(R)$ and $x = x(R)$. We calculate the total derivative w.r.t. R in both equations and get, after ordering,

$$\begin{aligned} & \left[u''(c_1) - \beta R u''(xc_1)x\right] c'(R) - \beta R u''(xc_1)c_1 x'(R) \\ &= \beta u'(xc_1), \qquad (3.44) \\ \left[u'(c_1) + \beta u'(xc_1)x\right] c'(R) &= -\beta u'(xc_1)c_1 x'(R). \qquad (3.45) \end{aligned}$$

Substituting $c'(R)$ from (3.45) into (3.44) and ordering now yields

$$-\left[x\frac{c_1 u''(c_1)}{u'(c_1)} + R\frac{xc_1 u''(xc_1)}{u'(xc_1)}\right]\frac{R}{x}x'(R) = x + R.$$

Since $-cu''(c)/u'(c) \equiv \theta(c)$, this can be written

$$\frac{R}{x}x'(R) = \frac{x+R}{x\theta(c_1) + R\theta(xc_1)}.$$

Finally, in view of $xc_1 = c_2$ and the definition of $\sigma(c_1, c_2)$, this gives (3.18).

C. Walras' law

In the proof of Proposition 1 we referred to Walras' law. Here is how Walras' law works in each period in a model like this. We consider period t, but for simplicity we skip the time index t on the variables. There are three markets, a market for capital services, a market for labor services, and a market for output goods. Suppose a "Walrasian auctioneer" calls out the price vector $(\hat{r}, w, 1)$, where $\hat{r} > 0$ and $w > 0$, and asks all agents, i.e., the young, the old, and the representative firm, to declare their supplies and demands.

The supplies of capital and labor are by assumption inelastic and equal to K units of capital services and L units of labor services. But the demand for capital and labor services depends on the announced $\hat{r}$ and w. Let the potential pure profit of the representative firm be denoted Π. If $\hat{r}$ and w are so that $\Pi < 0$, the firm declares $K^d = 0$ and $L^d = 0$. If on the other hand at the announced $\hat{r}$ and w, $\Pi = 0$ (as when $\hat{r} = r(k) + \delta$ and $w = w(k)$), the desired capital-labor ratio is given as $k^d = f'^{-1}(\hat{r})$ from (3.20), but the firm is indifferent w.r.t. the absolute level of the factor inputs. In this situation the auctioneer tells the firm to declare $L^d = L$ (recall L is the given labor supply) and $K^d = k^d L^d$ which is certainly acceptable for the firm. Finally, if $\Pi > 0$, the firm is tempted to declare infinite factor demands, but to avoid that, the auctioneer imposes the rule that the maximum allowed demands for capital and labor are $2K$ and $2L$, respectively. Within these constraints the factor demands will be uniquely determined by $\hat{r}$ and w and we have

$$\Pi = \Pi(\hat{r}, w, 1) = F(K^d, L^d) - \hat{r}K^d - wL^d. \tag{3.46}$$

The owners of both the capital stock K and the representative firm must be those who saved in the previous period, namely the currently old. These elderly will together declare the consumption $c_2 L_{-1} = (1 + \hat{r} - \delta)K + \Pi$ and the net investment $-K$ (which amounts to disinvestment). The young will declare the consumption $c_1 L = wL - s(w, r^e_{+1})L$ and the net investment $sL = s(w, r^e_{+1})L$. So aggregate declared consumption will be $C = (1 + \hat{r} - \delta)K + \Pi + wL - s(w, r^e_{+1})L$ and aggregate net investment $I - \delta K = s(w, r^e_{+1})L - K$. It follows that $C + I = wL + \hat{r}K + \Pi$. The aggregate declared supply of output is $Y^s = F(K^d, L^d)$. The values of excess demands in the three markets now add to

$$\begin{aligned} Z(\hat{r}, w, 1) &\equiv w(L^d - L) + \hat{r}(K^d - K) + C + I - Y^s \\ &= wL^d - wL + \hat{r}K^d - \hat{r}K + wL + \hat{r}K + \Pi - F(K^d, L^d) \\ &= wL^d + \hat{r}K^d + \Pi - F(K^d, L^d) = 0, \end{aligned}$$

by (3.46).

This is a manifestation of Walras' law for each period: *whatever the announced price vector for the period is, the aggregate value of excess demands in the period is zero.* The reason is the following. When each household satisfies its budget constraint and each firm pays out its ex ante profit,[24] then the economy as a whole has to satisfy an aggregate budget constraint for the period considered.

The budget constraints, demands, and supplies operating in this thought experiment (and in Walras' law in general) are the *Walrasian* budget constraints, demands, and supplies. Outside equilibrium these are somewhat artificial constructs. A Walrasian budget constraint is based on the assumption that the desired actions can be realized. This assumption will be wrong unless $\hat{r}$ and w are already at their equilibrium levels. But the assumption that desired actions can be realized is never falsified because the thought experiment does not allow trades to take place outside Walrasian equilibrium. Similarly, the Walrasian consumption demand by the worker is rather hypothetical outside equilibrium. This demand is based on the income the worker *would* get if fully employed at the announced real wage, not on the actual employment (or unemployment) at that real wage.

These ambiguities notwithstanding, the important message of Walras' law goes through, namely that when two of the three markets clear (in the sense of the Walrasian excess demands being nil), so does the third.

D. Proof of (i) and (ii) of Proposition 2

For convenience we repeat the fundamental difference equation characterizing an equilibrium path:

$$k_{t+1} = \frac{s\left(w\left(k_t\right), r\left(k_{t+1}\right)\right)}{1+n},$$

where $w(k) \equiv f(k) - f'(k)k > 0$ for all $k > 0$ and $r(k) \equiv f'(k) - \delta > -1$ for all $k \geq 0$. The key to the proof of Proposition 2 about existence of an equilibrium path is the following lemma.

LEMMA D1 Suppose the No Fast Assumption (A1) applies and let $w > 0$ and $n > -1$ be given. Then the equation

$$\frac{s\left(w, r\left(k\right)\right)}{k} = 1 + n. \tag{3.47}$$

has at least one solution $k > 0$.

[24] By ex ante profit is meant the hypothetical profit calculated on the basis of firms' desired supply evaluated at the announced price vector, $(\hat{r}, w, 1)$.

Proof. Note that $1+n>0$. From Lemma 1 in Section 3.3 follows that for all possible values of $r(k)$, $0<s(w,r(k))<w$. Hence, for any $k>0$,

$$0<\frac{s\left(w,r\left(k\right)\right)}{k}<\frac{w}{k}.$$

Letting $k\to\infty$ we then have $s\left(w,r\left(k\right)\right)/k\to 0$ since $s\left(w,r\left(k\right)\right)/k$ is squeezed between 0 and 0 (as indicated in the two graphs in Fig. 3.9).

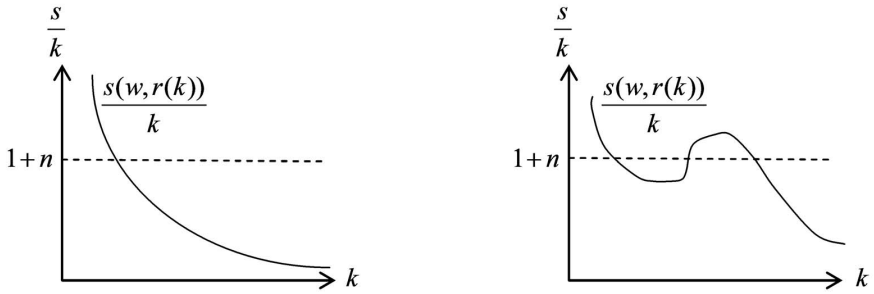

Figure 3.9: Existence of a solution to equation (3.47).

Next we consider $k\to 0$. There are two cases.
Case 1: $\lim_{k\to 0}s\left(w,r\left(k\right)\right)>0$.[25] Then obviously $\lim_{k\to 0}s\left(w,r\left(k\right)\right)/k=\infty$.
Case 2: $\lim_{k\to 0}s\left(w,r\left(k\right)\right)=0$.[26] In this case we have

$$\lim_{k\to 0}r\left(k\right)=\infty. \tag{3.48}$$

Indeed, since $f'(k)$ rises monotonically as $k\to 0$, the only alternative would be that $\lim_{k\to 0}r\left(k\right)$ exists and is $<\infty$; then, by Lemma 1 in Section 3.3, we would be in case 1 rather than case 2. By the second-period budget constraint, with $r=r(k)$, consumption as old is $c_2=s\left(w,r\left(k\right)\right)\left(1+r(k)\right)\equiv c(w,k)>0$ so that

$$\frac{s\left(w,r\left(k\right)\right)}{k}=\frac{c(w,k)}{\left[1+r(k)\right]k}.$$

The right-hand side of this equation goes to ∞ for $k\to 0$ since $\lim_{k\to 0}\left[1+r(k)\right]k=0$ by Technical Remark in Section 3.4 and $\lim_{k\to 0}c(w,k)=\infty$; this latter fact follows from the first-order condition (3.8), which can be written

$$0\leq u'(c(w,k))=(1+\rho)\frac{u'(w-s(w,r(k))}{1+r(k)}\leq(1+\rho)\frac{u'(w)}{1+r(k)}.$$

[25] If the limit does not exist, the proof applies to the *limit inferior* of $s\left(w,r\left(k\right)\right)$ for $k\to 0$. The limit inferior for $i\to\infty$ of a sequence $\{x_i\}_{i=0}^{\infty}$ is defined as $\lim_{i\to\infty}\inf\left\{x_j|\,j=i,i+1,\dots\right\}$, where inf of a set $S_i=\{x_j|\,j=i,i+1,\dots\}$ is defined as the greatest lower bound for S_i.

[26] If the limit does not exist, the proof applies to the *limit inferior* of $s\left(w,r\left(k\right)\right)$ for $k\to 0$.

Taking limits on both sides gives

$$\lim_{k\to 0} u'(c(w,k)) = (1+\rho)\lim_{k\to 0}\frac{u'(w-s\left(w,r\left(k\right)\right))}{1+r(k)} = (1+\rho)\lim_{k\to 0}\frac{u'(w)}{1+r(k)} = 0,$$

where the second equality comes from the fact that we are in case 2 and the third comes from (3.48). But since $u'(c) > 0$ and $u''(c) < 0$ for all $c > 0$, $\lim_{k\to 0} u'(c(w,k)) = 0$ requires $\lim_{k\to 0} c(w,k) = \infty$, as was to be shown.

In both Case 1 and Case 2 we thus have that $k \to 0$ implies $s\left(w,r\left(k\right)\right)/k \to \infty$. Since $s\left(w,r\left(k\right)\right)/k$ is a continuous function of k, there must be at least one $k > 0$ such that (3.47) holds (as illustrated by the two graphs in Fig. 3.14). □

Now, to prove (i) of Proposition 2, consider an arbitrary $k_t > 0$. We have $w(k_t) > 0$. In (3.47), let $w = w(k_t)$. By Lemma C1, (3.47) has a solution $k > 0$. Set $k_{t+1} = k$. Starting with $t = 0$, from a given $k_0 > 0$ we thus find a $k_1 > 0$ and letting $t = 1$, from the now given k_1 we find a k_2 and so on. The resulting infinite sequence $\{k_t\}_{t=0}^{\infty}$ is an equilibrium path. In this way we have proved existence of an equilibrium path if $k_0 > 0$. Thereby (i) of Proposition 2 is proved.

But what if $k_0 = 0$? Then, if $f(0) = 0$, no temporary equilibrium is possible in period 0, in view of (ii) of Proposition 1; hence there can be no equilibrium path. Suppose $f(0) > 0$. Then $w(k_0) = w(0) = f(0) > 0$, as explained in Technical Remark in Section 3.4. Let w in equation (3.47) be equal to $f(0)$. By Lemma C1 this equation has a solution $k > 0$. Set $k_1 = k$. Letting period 1 be the new initial period, we are back in the case with initial capital positive. This proves (ii) of Proposition 2.

E. Sufficient conditions for certain properties of the transition curve

Positive slope everywhere For convenience we repeat here the condition (3.36):

$$\frac{1}{1-\gamma} > \frac{1-\sigma}{1+(1+\rho)^{-\sigma}(1+f'(k)-\delta)^{\sigma-1}}, \qquad (*)$$

where we have substituted $\sigma \equiv 1/\theta$. In Section 3.5.3 we claimed that in the CRRA-CES case this condition is sufficient for the transition curve to be positively sloped everywhere. We here prove the claim.

Consider an arbitrary $k_t > 0$ and let $w \equiv w(k_t) > 0$. Knowing that $w'(k_t) > 0$ for all $k_t > 0$, we can regard k_{t+1} as directly linked to w. With k representing k_{t+1}, k must satisfy the equation $k = s(w, r(k))/(1+n)$. A sufficient condition for this equation to implicitly define k as an increasing function of w is also a sufficient condition for the transition curve to be positively sloped for all $k_t > 0$.

When $u(c)$ belongs to the CRRA class, by (3.15) with $\sigma \equiv 1/\theta$, we have $s(w, r(k)) = \left[1+(1+\rho)^{\sigma}(1+r(k))^{1-\sigma}\right]^{-1} w$. The equation $k = s(w, r(k))/(1+n)$

then implies

$$\frac{w}{1+n} = k\left[1+(1+\rho)^{\sigma}R(k)^{1-\sigma}\right] \equiv h(k), \tag{3.49}$$

where $R(k) \equiv 1 + r(k) \equiv 1 + f'(k) - \delta > 0$ for all $k > 0$. It remains to provide a sufficient condition for obtaining $h'(k) > 0$ for all $k > 0$. We have

$$h'(k) = 1 + (1+\rho)^{\sigma}R(k)^{1-\sigma}\left[1-(1-\sigma)\eta(k)\right], \tag{3.50}$$

since $\eta(k) \equiv -kR'(k)/R(k) > 0$, the sign being due to $R'(k) = f''(k) < 0$. So $h'(k) > 0$ if and only if $1-(1-\sigma)\eta(k) > -(1+\rho)^{-\sigma}R(k)^{\sigma-1}$, a condition equivalent to

$$\frac{1}{\eta(k)} > \frac{1-\sigma}{1+(1+\rho)^{-\sigma}R(k)^{\sigma-1}}. \tag{3.51}$$

To make this condition more concrete, consider the CES production function

$$f(k) = A(\alpha k^{\gamma} + 1 - \alpha), \qquad A > 0, 0 < \alpha < 1, \gamma < 1. \tag{3.52}$$

Then $f'(k) = \alpha A^{\gamma}(f(k)/k)^{1-\gamma}$ and defining $\pi(k) \equiv f'(k)k/f(k)$ we find

$$\eta(k) = (1-\gamma)\frac{(1-\pi(k))f'(k)}{1-\delta+f'(k)} \leq (1-\gamma)(1-\pi(k)) < 1-\gamma, \tag{3.53}$$

where the first inequality is due to $0 \leq \delta \leq 1$ and the second to $0 < \pi(k) < 1$, which is an implication of strict concavity of f combined with $f(0) \geq 0$. Thus, $\eta(k)^{-1} > (1-\gamma)^{-1}$ so that if (*) holds for all $k > 0$, then so does (3.51), i.e., $h'(k) > 0$ for all $k > 0$. We have hereby shown that (*) is sufficient for the transition curve to be positively sloped everywhere.

Transition curve steep for k small Here we specialize further and consider the CRRA-Cobb-Douglas case: $u(c) = (c^{1-\theta}-1)/(1-\theta)$, $\theta > 0$, and $f(k) = Ak^{\alpha}$, $A > 0$, $0 < \alpha < 1$. In the prelude to Proposition 4 in Section 3.5 it was claimed that if this combined utility and technology condition holds at least for small k, then (ii) of (A3) is satisfied. We now show this.

Letting $\gamma \to 0$ in (3.52) gives the Cobb-Douglas function $f(k) = Ak^{\alpha}$ (this is proved in the appendix to Chapter 4). With $\gamma = 0$, clearly $(1-\gamma)^{-1} = 1 > 1-\sigma$, where $\sigma \equiv \theta^{-1} > 0$. This inequality implies that (*) above holds and so the transition curve is positively sloped everywhere. As an implication there is a transition function, φ, such that $k_{t+1} = \varphi(k_t)$, $\varphi'(k_t) > 0$. Moreover, since $f(0) = 0$, we have, by Lemma 5, $\lim_{k_t \to 0}\varphi(k_t) = 0$.

Given the imposed CRRA utility, the fundamental difference equation of the model is

$$k_{t+1} = \frac{w(k_t)}{(1+n)\left[1+(1+\rho)^{\sigma}R(k_{t+1})^{1-\sigma}\right]} \tag{3.54}$$

or, equivalently,

$$h(k_{t+1}) = \frac{w(k_t)}{1+n},$$

where $h(k_{t+t})$ is defined as in (3.49). By implicit differentiation we find $h'(k_{t+1})\varphi'(k_t) = w'(k_t)/(1+n)$, i.e.,

$$\varphi'(k_t) = \frac{w'(k_t)}{(1+n)h'(k_{t+1})} > 0.$$

If $k^* > 0$ is a steady-state value of k_t, (3.54) implies

$$1 + (1+\rho)^\sigma R(k^*)^{1-\sigma} = \frac{w(k^*)}{(1+n)k^*}, \tag{3.55}$$

and the slope of the transition curve at the steady state will be

$$\varphi'(k^*) = \frac{w'(k^*)}{(1+n)h'(k^*)} > 0. \tag{3.56}$$

If we can show that such a $k^* > 0$ exists, is unique, and implies $\varphi'(k^*) < 1$, then the transition curve crosses the 45° line from above, and so (ii) of (A3) follows in view of $\lim_{k_t \to 0} = 0$.

Defining $x(k) \equiv f(k)/k = Ak^{\alpha-1}$, where $x'(k) = (\alpha-1)Ak^{\alpha-2} < 0$, and using that $f(k) = Ak^\alpha$, we have $R(k) = 1 + \alpha x(k) - \delta$ and $w(k)/k = (1-\alpha)x(k)$. Hence, (3.55) can be written

$$1 + (1+\rho)^\sigma (1 + \alpha x^* - \delta)^{1-\sigma} = \frac{1-\alpha}{1+n} x^*, \tag{3.57}$$

where $x^* = x(k^*)$. It is easy to show graphically that this equation has a unique solution $x^* > 0$ whether $\sigma < 1$, $\sigma = 1$, or $\sigma > 1$. Then $k^* = (x^*/A)^{1/(\alpha-1)} > 0$ is also unique.

By (3.50) and (3.57),

$$\begin{aligned} h'(k^*) &= 1 + (\frac{1-\alpha}{1+n}x^* - 1)\,[1 - (1-\sigma)\eta(k^*)] > 1 + (\frac{1-\alpha}{1+n}x^* - 1)(1 - \eta(k^*)) \\ &\geq 1 + (\frac{1-\alpha}{1+n}x^* - 1)\alpha, \end{aligned}$$

where the first inequality is due to $\sigma > 0$ and the second to the fact that $\eta(k) \leq 1 - \alpha$ in view of (3.53) with $\gamma = 0$ and $\pi(k) = \alpha$. Substituting this together with $w'(k^*) = (1-\alpha)\alpha x^*$ into (3.56) gives

$$0 < \varphi'(k^*) < \frac{\alpha x^*}{1 + n + \alpha x^*} < 1, \tag{3.58}$$

as was to be shown.

The CRRA-Cobb-Douglas case is well-behaved For the case of CRRA utility and Cobb-Douglas technology with CRS, existence and uniqueness of a steady state has just been proved. Asymptotic stability follows from (3.58). So the CRRA-Cobb-Douglas case is well-behaved.

3.10 Exercises

3.1 The dynamic accounting relation for a closed economy is

$$K_{t+1} = K_t + S^N \qquad (*)$$

where K_t is the aggregate capital stock and S_t^N is aggregate net saving. In the Diamond model, let S_{1t} be aggregate net saving of the young in period t and S_{2t} aggregate net saving of the old in the same period. On the basis of (*) give a direct proof that the link between two successive periods takes the form $k_{t+1} = s_t/(1+n)$, where s_t is the saving of each young, n is the population growth rate, and k_{t+1} is the capital/labor ratio at the beginning of period $t+1$. *Hint:* by definition, the increase in financial wealth is the same as net saving (ignoring gifts).

3.2 Suppose the production function in Diamond's OLG model is $Y = A(\alpha K^\gamma + (1-\alpha)L^\gamma)^{1/\gamma}$, $A > 0, 0 < \alpha < 1, \gamma < 0$, and $A\alpha^{1/\gamma} < 1+n$. a) Given $k \equiv K/L$, find the equilibrium real wage, $w(k)$. b) Show that $w(k) < (1+n)k$ for all $k > 0$. *Hint:* consider the roof. c) Comment on the implication for the long-run evolution of the economy. *Hint:* consider the ceiling.

3.3 *(multiple temporary equilibria with self-fulfilling expectations)* Fig. 3.10 shows the transition curve for a Diamond OLG model with $u(c) = c^{1-\theta}/(1-\theta)$, $\theta = 8$, $\rho = 0.4$, $n = 0.2$, $\delta = 0.6$, $f(k) = A(bk^p + 1 - b)^{1/p}$, $A = 7$, $b = 0.33$, $p = -0.4$.

a) Let $t = 0$. For a given k_0 slightly below 1, how many temporary equilibria with self-fulfilling expectations are there?

b) Suppose the young in period 0 expect the real interest rate on their saving to be relatively low. Describe by words the resulting equilibrium path in this case. Comment (what is the economic intuition behind the path?).

c) In the first sentence under b), replace "low" by "high". How is the answer to b) affected? What kind of difficulty arises?

3.4 *(plotting the transition curve by MATLAB)* This exercise requires computation on a computer. You may use *MATLAB OLG program.*[27]

[27] Made by Marc P. B. Klemp and available at the address:

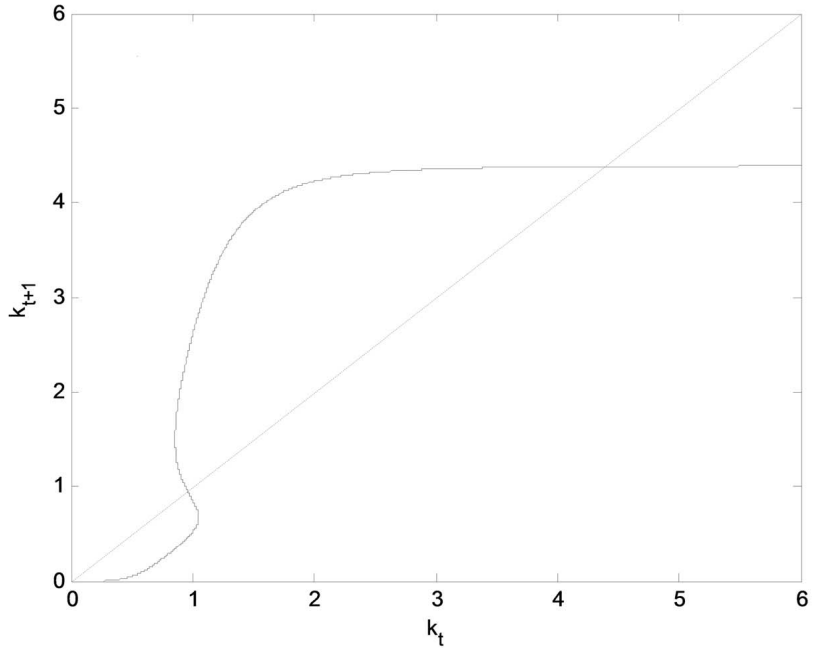

Figure 3.10: Transition curve for Diamond's OLG model in the case described in Exercise 3.3.

a) Enter the model specification from Exercise 3.3 and plot the transition curve.

b) Plot examples for two other values of the substitution parameter: $p = -1.0$ and $p = 0.5$. Comment.

c) Find the approximate largest lower bound for p such that higher values of p eliminates multiple equilibria.

d) In continuation of c), what is the corresponding elasticity of factor substitution, ψ? *Hint:* as shown in §4.4, the formula is $\psi = 1/(1-p)$.

e) The empirical evidence for industrialized countries suggests that $0.4 < \psi < 1.0$. Is your ψ from d) empirically realistic? Comment.

3.5 *(one stable and one unstable steady state)* Consider the following Diamond model: $u(c) = \ln c$, $\rho = 2.3$, $n = 2.097$, $\delta = 1.0$, $f(k) = A(bk^p + 1 - b)^{1/p}$, $A = 20$, $b = 0.5$, $p = -1.0$.

http://www.econ.ku.dk/okocg/Computation/main.htm.

a) Plot the transition curve of the model. *Hint:* you may use either a program like *MATLAB OLG Program* (available on the course website) or first a little algebra and then Excel (or similar simple software).

b) Comment on the result you get. Will there exist a poverty trap? Why or why not?

c) At the stable steady state calculate numerically the output-capital ratio, the aggregate saving-income ratio, the real interest rate, and the capital income share of gross national income.

d) Briefly discuss how your results in c) comply with your knowledge of corresponding empirical magnitudes in industrialized Western countries?

e) There is one feature which this model, as a long-run model, ought to incorporate, but does not. Extend the model, taking this feature into account, and write down the fundamental difference equation for the extended model in algebraic form.

f) Plot the new transition curve. *Hint:* given the model specification, this should be straightforward if you use Excel (or similar); and if you use MATLAB OLG Program, note that by a simple "trick" you can transform your new model into the "old" form.

g) The current version of the MATLAB OLG Program is not adapted to this question. So at least here you need another approach, for instance based on a little algebra and then Excel (or similar simple software). Given $k_0 = 10$, calculate numerically the time path of k_t and plot the *time profile* of k_t, i.e., the graph (t, k_t) in the tk-plane. Next, do the same for $k_0 = 1$. Comment.

3.6 *(dynamics under myopic foresight)*
(incomplete) Show the possibility of a chaotic trajectory.

3.7 Given the period utility function is CRRA, derive the saving function of the young in Diamond's OLG model. *Hint:* substitute the period budget constraints into the Euler equation.

3.8 *Short questions* a) A steady-state capital-labor ratio can be in the "dynamically efficient" region or in the "dynamically inefficient" region. How are the two mentioned regions defined? b) Give a simple characterization of the two regions. c) The First Welfare Theorem states that, given certain conditions, any competitive equilibrium ($\equiv$ Walrasian equilibrium) is Pareto optimal. Give a list of circumstances that each tend to obstruct Pareto optimality of a competitive equilibrium.

3.9 Consider a Diamond OLG model for a closed economy. Let the utility discount rate be denoted ρ and let the period utility function be specified as $u(c) = \ln c$.

a) Derive the saving function of the young. Comment.

b) Let the aggregate production function be a neoclassical production function with CRS and ignore technological progress. Let L_t denote the number of young in period t. Derive the fundamental difference equation of the model.

From now, assume that the production function is $Y = \alpha L + \beta KL/(K + L)$, where $\alpha > 0$ and $\beta > 0$ (as in Problem 2.4).

c) Draw a transition diagram illustrating the dynamics of the economy. Make sure that you draw the diagram so as to exhibit consistency with the production function.

d) Given the above information, can we be sure that there exists a unique and globally asymptotically stable steady state? Why or why not?

e) Suppose the economy is in a steady state up to and including period $t_0 > 0$. Then, at the shift from period t_0 to period $t_0 + 1$, a negative technology shock occurs such that the technology level in period $t_0 + 1$ is below that of period t_0. Illustrate by a transition diagram the evolution of the economy from period t_0 onward. Comment.

f) Let $k \equiv K/L$. In the $(t, \ln k)$ plane, draw a graph of $\ln k_t$ such that the qualitative features of the time path of $\ln k$ before and after the shock, including the long run, are exhibited.

g) How, if at all, is the real interest rate in the long run affected by the shock?

h) How, if at all, is the real wage in the long run affected by the shock?

i) How, if at all, is the labor income share of national income in the long run affected by the shock?

j) Explain by words the economic intuition behind your results in h) and i).

3.10