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 between human beings 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 studies on the major lines of the economics of a society and does so in an intertemporal perspective – evolution over time is in focus.

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 – a link to the economic world outside the window – 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 the concerns of the agents and the technological constraints and social structures (market forms, social conventions, and legal institutions) that condition 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.

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. The economic system as a whole is more than the sum of its parts. This also brings to mind that macroeconomics has to 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 over time.

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, the student should be aware of the existence of differing conceptions and in several matters even conflicting theoretical schools.

1.1.2 The different "runs"

This textbook is about industrialized market economies of today. We study basic concepts, models, and analytical methods of relevance for understanding macroeconomic processes in such economies. 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 at most 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 20 years' interval reflects primarily the change in supply side factors such as the educational level of the labor force, the capital stock, and the technology. Within time horizon also changes in institutions (market structure, government planning and regulation, rules of the game) come into focus.

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, three 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, three-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

[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.

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 *explanation* of changes in technology out of consideration, which is tantamount to regarding these changes as exogenous.[2]

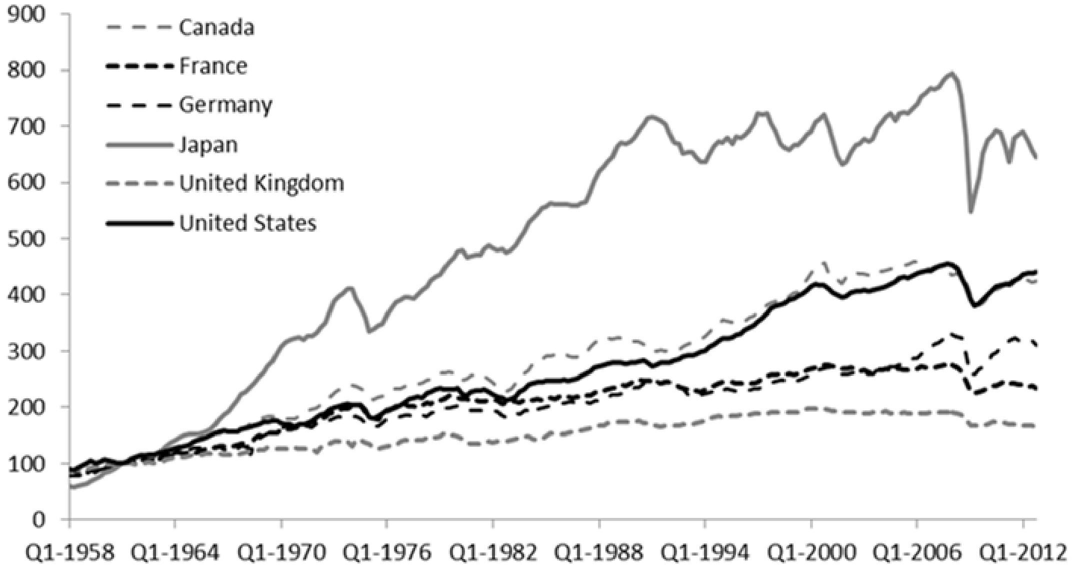

Figure 1.1: Quarterly Industrial Production Index in six major countries (Q1-1958 to Q2-2013; index Q1-1961=100). Source: OECD Industry and Service Statistics. Note: Industrial production includes manufacturing, mining and quarrying, electricity, gas, and water, and construction.

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 Global 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.

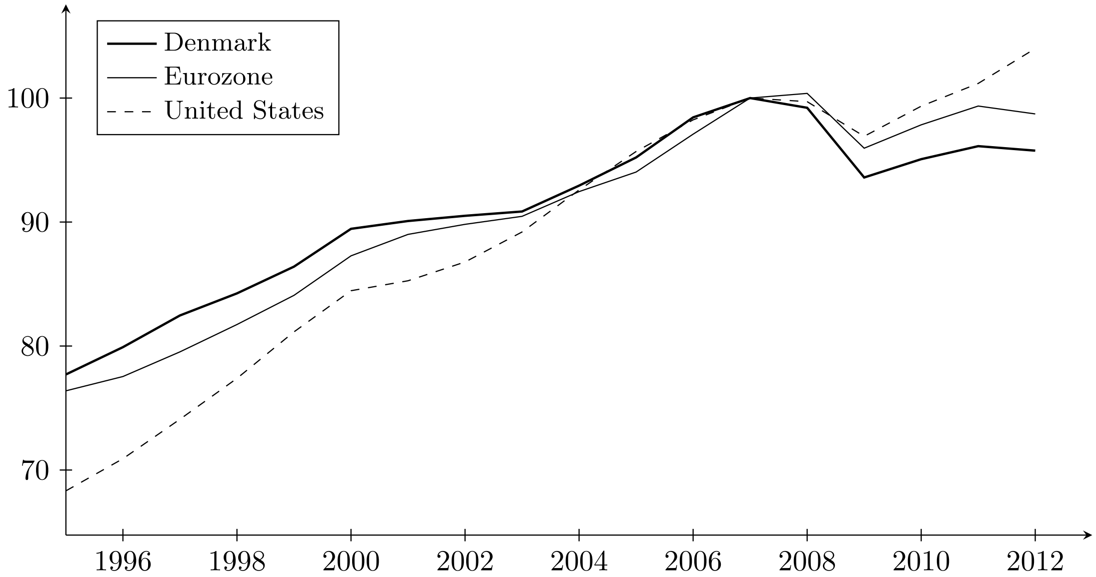

Figure 1.2: Indexed real GDP for Denmark, Eurozone and US, 1995-2012 (2007=100). Source: EcoWin and Statistics Denmark.

1.2 Elements of macroeconomic analysis

1.2.1 Model elements

Basic categories

- Agents: We use simple descriptions of the economic agents (decision makers): 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.

- Households face *budget constraints*, and firms face *technological constraints,* in macroeconomics.typically described as *production functions.*

- Resources: *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, semi-manufacture, and energy (often lumped together as *intermediate goods). Natural resources* include land and other non-reproducible means of production. *Human capital* is the stock of productive skills embodied in an individual.

- Goods, labor, and assets *markets.*

- Market forms and other *rules* regulating the economic interactions.

Types of variables

Endogenous variable = variable whose value is determined within the particular model considered. *Exogenous* variable = variable whose value the particular model considered takes as given.

Stock = a variable measured as a quantity at a given point in time. *Flow.*= a variable measured as a quantity per time unit.

State variable = 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 *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 and is thus a jump variable.

A state variable may alternatively be called a *predetermined* variable. And a jump variable may alternatively be called a *non-predetermined variable* or a *control* variable.

Types of 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

Static versus dynamic models

A *static model* is a model where time does not enter or at least where all variables refer to the same point in time. A *dynamic* model is a model that establishes a link from the state of the economic system (including its recent history) to the subsequent state. A dynamic model thus allows a derivation of the evolution over time of the endogenous variables.

Macroeconomics is about studies processes in real time and the emphasis is thus on dynamic models. Occasionally we consider *quasi-static models.* The modifier "quasi-" is meant to indicate that although the model concentrates on a single period, it considers some variables as inherited from *the past* and some variables that involve expectations about the future. What we call *temporary equilibrium models* belong to this category. Their role is to serve as a prelude to a more elaborate dynamic model dealing with a sequence of states.

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?

A study of *dynamic effects of a parameter shift in real time* is a variety of dynamic analysis. Comparative analysis is a different thing; in *comparative dynamics* we compare solutions to a dynamic model under alternative values of the parameters and exogenous variables; in *comparative statics* we compare solutions to a static model under alternative values of the parameters and exogenous variables.

In dynamic modeling and analysis we have a choice between framing the model in period terms or in continuous time. *Period analysis*, also called discrete time analysis, is the method we generally apply up to Chapter 9, where a transition to *continuous-time analysis* is undertaken.

Partial equilibrium analysis versus general equilibrium analysis

We say that a given single market is in *partial equilibrium* at a given point in time if for 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 takes 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 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 the outcome in a single market 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 in the literature also 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 derivation of the saving function of the household is by some authors included under the heading partial equilibrium analysis for the reason that the real wage and real interest rate appearing as arguments in the derived saving function are arbitrary. In this book, however, we shall call the analysis of a single decision maker's problem *partial analysis*, not partial equilibrium analysis. The motivation 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 From input to 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 *reproducible* durable means of production, the other being *labor,* often 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) may nevertheless raise several conceptual issues.

The time dimension of input and output

A key issue is: how are the variables entering (1.1) *denominated,* that is, what is the *dimension* of the variables? Or in what units are the variables measured? It is most satisfactory, from a theoretical as well as 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. It is less recognized, however, 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.* We thus 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:

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

where the dimension of the variables is indicated in brackets. Note that to be correct, an equation should have not only the same quantity on both sides, but also the same dimension. Both conditions are satisfied by (1.2), since $K \times 1 = K$ and (machines × (machine-yrs/yr)/machine) = (machine-yrs/yr). Sometimes we consider equations where a bare number, also known as a *dimensionless quantity*, appears on both sides. Such quantities may arise as the the product or ratio of two quantities that are not dimensionless. For instance, the fraction of income saved is a bare number since both saving and income are measured in the same units, say, euros per year, whereby the dimensions cancel out. In such cases the variable in question is said to have dimension *one.*[3]

Considering the labor input, suppose similarly that the stock of laborers is L men and that on average they deliver one *man-year* (say h hours) per year. Then the total labor input is L man-years per year:

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

Now, a reason that stocks and flows may be confused is that often the same symbol, K, appearing in the production function as a capital *input flow*, also,

[3] It is like in physics. Length, time, and speed are measured in dimensional units, such as metre, second and metre/second whereas the alcohol percentage in a beverage is a bare number.

within the same model, appears as the capital *stock* in an accumulation equation like

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

In this equation, I_t is gross investment in period t, and δ is the rate of physical capital depreciation due to wear and tear ($0 \leq \delta \leq 1$). So the symbol K_t must represent the capital *stock* at the beginning of period t. 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 L_t measures a stock (number of persons) whereas in (1.1) and (1.3) L measures a flow that depends on the average rate of utilization of the 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 should ensure that our notation *is* consistent. This requires normalization of the utilization rates for capital and labor in the production function so as 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 and their dimensions are the same.

As an illustration of the importance of being aware of the distinction between stock and flows, let

$$\begin{aligned} Y &= GDP \text{ per year, and} \\ P &= \text{average size of population over the year.} \end{aligned}$$

Then income per year per capita can be decomposed the following way:

$$\begin{aligned} \frac{GDP}{P} &\equiv \frac{\text{value added/yr}}{\text{\#people}} = \frac{\text{value added/yr}}{\text{\#hours of work/yr}} \\ &\quad \times \frac{\text{\#hours of work/yr}}{\text{\#employed workers}} \times \frac{\text{\#employed workers}}{\text{\#workers}} \times \frac{\text{\#workers}}{\text{\#people}}, \end{aligned} \tag{1.5}$$

where # stand for “number of”, and “employed workers” and “workers” stand for “full-time” people, thus weighting by the fraction of a standard man-year they actually work or at least want to work, respectively. That is, aggregate per capita income equals average labor productivity times average labor intensity times the employment rate times the workforce participation rate. An increase from one year to the next in per capita income thus 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 for instance a fall in productivity, but may reflect, say, a fall in the employment rate (a rise in unemployment) or in the participation rate due to an ageing population.

Natural resources?

A *second* conceptual issue concerning the production function in (1.1) is: what about the role of 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, to simplify, 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.

Intermediate goods?

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. It therefore seems strange to regard output as produced by only capital and labor. Again, the motivation is that putting the engineering input-output relations involving intermediate goods aside is a convenient simplification. One imagines 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

(very incomplete)

Stylized national income and product accounts

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 intermediate goods (including raw materials and 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 means of production, *labor*, and *capital* (producible durable means of production). In practice also intermediate goods are a necessary production input. As mentioned above, in simple models this input is regarded as itself produced at an early stage within the production period and then used up during the remainder of the production process. In more rigorous dynamic analyses, however, the intermediate goods are considered produced *prior* to the production process in which they are used. To see what this looks like and what it means to abstract from it in the simpler models, we here consider intermediate goods as a fourth input type produced separately in Sector 1.

....

1.4 Some terminological points

(Incomplete)

We follow the convention in macroeconomics and, unless otherwise specified, use "capital" for physical capital, that is, a (re-producible) 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 real as well as financial 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 expected stream of future labor income,[4] 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 of stock are examples. A *financial liability* of an economic agent is an obligation to transfer resources to others in the future. A mortgage loan is an example.

In spite of the described distinction between what is called physical assets and what is called financial assets, often in macroeconomics the household's "financial wealth" is used as synonymous with its non-human wealth. In this book, unless otherwise indicated, we follow this convention. Thereby, a household's *financial wealth* is the total value of its non-human assets, thus including not only its net financial assets, but also its 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) versus savings (stock).

...

1.5 Brief history of macroeconomics

Text not yet available.

—

?

?

1.6 Literature notes

....

[4] And is thus to be distinguished from *human capital*, which, as defined in Section 2.1, is a production factor.

The modern theory of economic growth ("new growth theory", "endogenous growth theory") is extensively covered in dedicated textbooks like **?**, **?**, **?**, **?**, and **?**. A good introduction to analytical development economics is Basu (1997).

?, **?**, and **?** 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 **?**, **?**, and **?**. Somewhat different perspectives, from opposite poles, are offered by **?** and **?**.

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 useful in a 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.

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:

$$\begin{aligned} F_K &> 0, \quad F_L > 0, && (2.2) \\ F_{KK} &< 0, \quad F_{LL} < 0. && (2.3) \end{aligned}$$

(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 $\left\{(K, L) \in \mathbb{R}_+ | \ F(K, L) = \bar{Y}\right\}$. 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 may 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

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

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) how much of K can be saved by applying an extra unit of labor. Hence, a cost-minimizing firm that plans to produce $\bar{Y}$ units, will choose inputs, K and L, such that the marginal rate of substitution of K for L equals the inverse factor price ratio.

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).

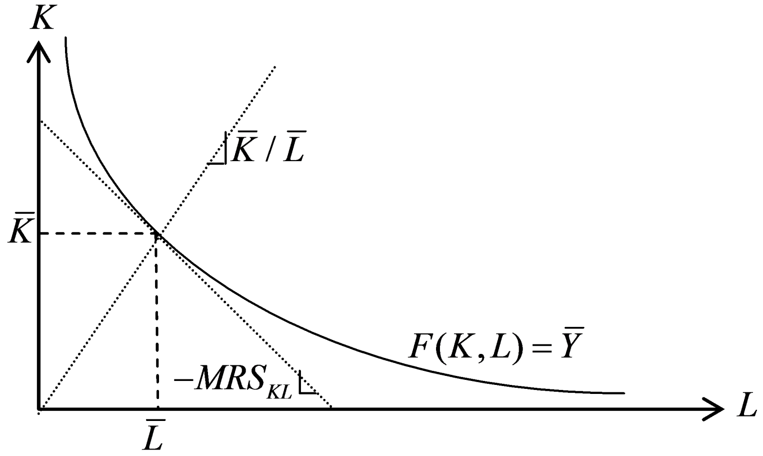

Figure 2.1: MRS_{KL} as the absolute slope of the isoquant representing $F(K, L) = \bar{Y}$.

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 $|Y = \bar{Y}$ 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 $F(K, L) = \bar{Y}$.

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,* 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 exhibits a problem in defending CRS of our present production function, F, by an appeal to the replication argument. Besides capital and labor, also land is a necessary input and should in principle 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 with respect to capital and labor alone. In manufacturing and services, empirically, this and other possible sources for departure from CRS with respect to capital and labor may be minor and so many macroeconomists feel comfortable enough with assuming CRS with respect to 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 a differentiable production function $F(K,L)$ 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.

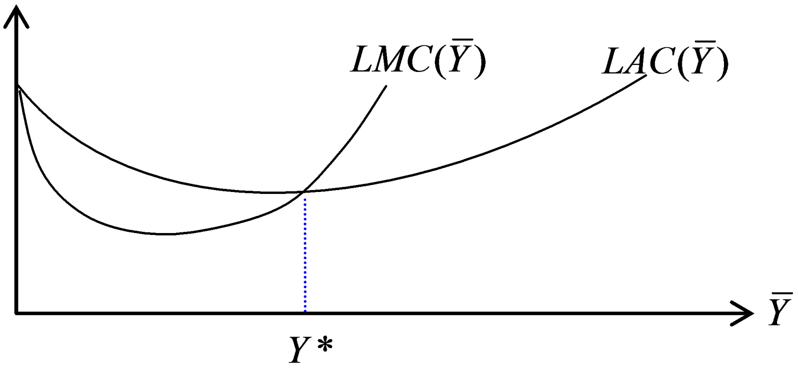

Figure 2.2: Locally CRS at optimal plant size.

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 with respect to 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, for all $(K, L) \in \mathbb{R}^2_{++}$, $F_{KL} > 0$, i.e., it exhibits *direct complementarity* between K and L.What is ruled out in the CRS case is thus that $F_{KL} < 0$ (in which case K and L are said to be *direct substitutes),* or that $F_{KL} = 0$.

(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]

[10] Proofs of claim (iii), (iv), and (v) are in Appendix B.

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

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

And the marginal productivity of labor can be written

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

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,

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

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

[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 $f'(\gamma) = (f(\beta) - f(\alpha))/(\beta - \alpha)$.

$> f(k) - kf'(k) > f(0)$ and continuity of f (so that $\lim_{k\to 0^+} f(k) = f(0)$) follows (2.18).

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 with respect to the B/A ratios, we get a labor productivity curve

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

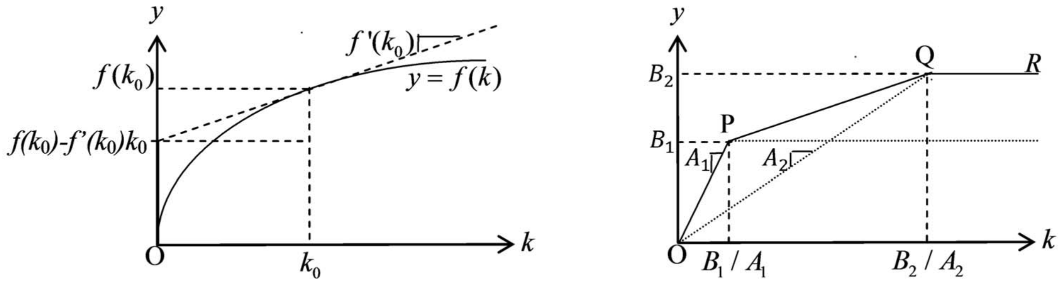

Figure 2.3: Two labor productivity curves based on CRS technologies. Left: neoclassical technology with Inada conditions for MPK satisfied; the graphical representation of MPK and MPL at $k = k_0$ as $f'(k_0)$ and $f(k_0) - f'(k_0)k_0$ are indicated. Right: the line segment PQ makes up an efficient combination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

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 $k > B_n/A_n$, $MPK = 0$,[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

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

change take a form known as *factor-augmenting* technological change:

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

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 B_t and A_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 B_t in (2.20) is a constant (which we can then normalize to one). So only A_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 with respect to 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-labor ratio or 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

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

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.

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 $B_t = A_t$ 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 obscure reason this form became known as *Solow-neutral* technological progress.[16] This association of (2.23) to Solow's name may easily confuse people, however. In his famous growth model,[17] well-known from introductory macroeconomics, Solow assumed Harrod-neutral technological progress. And in another famous contribution, Solow generalized the concept of Harrod-neutrality to the case of *embodied* technological change and capital of *different vintages,* see below.

It is easily shown (Exercise 2.5) that if F in (2.20) is a Cobb-Douglas production function, then F satisfies all three neutrality criteria at the same time, if it satisfies one of them (which requires that 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 way in which it materializes. For instance, the occurrence of Harrod-neutrality should not be interpreted as if something miraculous has happened to the labor input. It only means that technological innovations predominantly are such that not only do labor and

[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).

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).)

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 pharmaceutical formula are examples. (Note that the distinction rival versus nonrival is different from the distinction excludable versus 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 has to do with the legal status of a piece of knowledge and does not interfere with its technical character as a nonrival input. Finally, a good that is both non-rival and non-excludable is called a *pure public good.*)

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 with respect to 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 with respect to the rival inputs and T taken together.

From the perspective of the theory of economic growth, the important distinction between a rival and a non-rival input can be exemplified this way. Adding a new tractor to the economy benefits one farmer. But adding a new idea − a new piece of technical knowledge − benefits everyone that wants to use it. In brief: the economic value of an idea is proportional to the number of users.

The perpetual inventory method

Before proceeding, a brief remark about how the capital stock K_t can be in principle 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 versus 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 of high importance. Hence, unless explicitly specified otherwise, technological change is understood to be disembodied.

2.3 The concepts of representative firm and aggregate production function

Many macroeconomic models make use of the simplifying, and not unproblematic, notion of a *representative firm.* By this is meant a fictional firm whose production "represents" the 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, $i = 1, 2, \dots, n$. Define $Y \equiv \Sigma_{i=1}^n Y_i$, $K \equiv \Sigma_{i=1}^n K_i$, and $L \equiv \Sigma_{i=1}^n L_i$. Let the firms maximize profits, taking input and output prices as given. Suppose the aggregate variables are then related through some function, F^*, such that we can write

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

and such that the input choices of a single fictional 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 economy. If this is possible, we call $F^*(K, L)$ 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 this fictional single firm.

A simple example where an aggregate production function is well-defined is the following. Suppose all the firms have the *same* production function so that $Y_i = F(K_i, L_i)$, $i = 1, 2, \dots, n$. 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 the factor prices, profit-maximizing firms will choose the same capital intensity $k_i = k$ for all i (but not necessarily the same level of production since under CRS, this is indeterminate). 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 case an aggregate production function immediately appears and turns out to be exactly the same as the identical CRS production functions of the individual firms. Moreover, given F is neoclassical, the common capital-labor ratio $k_i = k$, for all i, implies that $\partial Y_i/\partial K_i = f'(k_i) = f'(k) = F_K(K, L) = \partial Y/\partial K$ for all i. So each firm's marginal productivity of capital is the same as the marginal productivity of capital calculated on the basis of the aggregate production function.

A less trivial case is the following. Let the firms have *different* concave neoclassical production functions at firm level. Define the function F by

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

Then $F(K,L)$ is a "well-behaved" aggregate production function. Indeed, the n individual firms will choose inputs such that both $\partial Y_i/\partial K_i$ $(= F_K^i(K_i,L_i)$ and $\partial Y_i/\partial L_i$ $(= F_L^i(K_i,L_i)$ are the same across firms, namely equal to the cost per unit of capital and the cost per unit of labor, respectively. By the envelope theorem (see Math Tools) it can then be shown that F will be such that $\partial Y/\partial K$ $= F_K(K,L)$ and $\partial Y/\partial L$ will equal $\partial Y_i/\partial K_i$ and $\partial Y_i/\partial L_i$, respectively.

A next step is to allow also for the existence of different output goods (either within or across the single firms), different capital goods, and different types of labor. This makes the issue much 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 the given prices, the aggregate profit is maximized on the basis of the firms' combined production technology.[18]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 at least simple qualitative properties from its constituent parts) are very 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 lower rate of interest always induces a higher aggregate capital-labor ratio. 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

[18] See Mas-Colell (1995).

devastating problems only under exceptional circumstances. In the end this is a matter of empirical assessment.[19]

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, tolerable in a first rough approximation. Of course there are cases where this "as if" approach is clearly inappropriate and 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. A brief example is given in 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 notion of a representative household may be a useful simplifying concept. On the other hand, if we aim at understanding, say, the *interaction* between lending and borrowing households, perhaps via financial intermediaries, the existence of *different* categories of households should be taken into account. Similarly, if the theme is conflicts of interests between firm owners and employees. And if we want to assess the welfare costs of business cycle fluctuations, we should take into account that exposure to unemployment risk tends to be very unevenly distributed in the population.

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. Aggregate physical capital is just the accumulated amount of what is left of the produced good after aggregate

[19] 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."

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) All firms are alike and maximize profit subject to the same 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. By definition this means that 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 this setup the *neoclassical competitive one-sector setup*. It is certainly an abstraction from the diversity and multitude of frictions in the real world. Nevertheless, the outcome under the described 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 firm in a single period. The firm has 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).[20]

Since F is neoclassical, we have $F_{KK} < 0$ and $F_{LL} < 0$ everywhere. To obtain 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).

Let us 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.

The DRS case

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

$$D > 0.$$

[20] 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 $F(K,L) \geq \lambda F(K_1,L_1) + (1-\lambda)F(K_2,L_2)$.

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[21] 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

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

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

in view of $F_{LK} = F_{KL}$.[22]

In contrast to the cases of CRS and IRS (for a two-factor production function), here we cannot be sure that direct complementarity between K and L (i.e., $F_{KL} >$

[21] 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.

[22] 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.

0) holds everywhere; this explains the "if" in (2.33) and (2.34). In any event, the rule 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 with respect to k and the numerator is the elasticity of f' with respect to 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.

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 does not take us far. 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

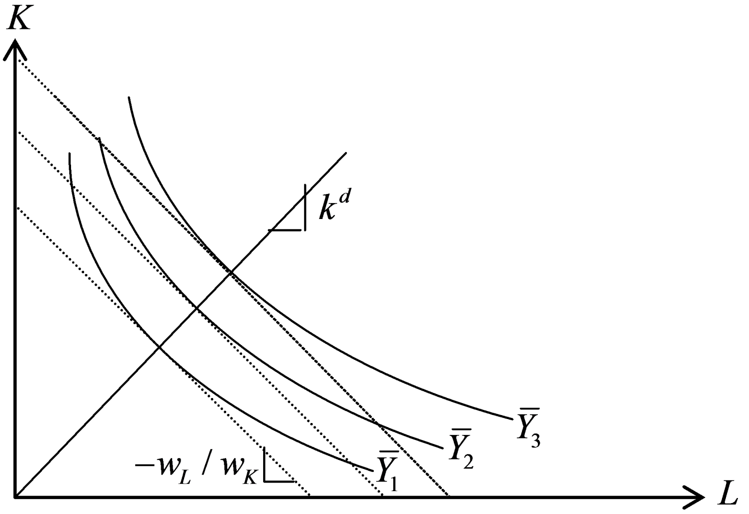

Figure 2.4: Constancy of MRS along rays when the production function is homogeneous of degree h (the cost-minimizing capital intensity is the same at all output levels).

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 with respect to 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. With respect to capital this is a "natural" assumption since in a closed economy in the short run the available amount of capital will be *predetermined,* that is, historically determined by the accumulated previous investment in the economy. With respect to 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. Since we now consider the aggregate level, we interpret K^d and L^d as factor demands by a representative firm.

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 factor 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.[23] 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 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* adjustment process.[24] We imagine that our period is sub-divided into many short time intervals $(t, t+\Delta t)$. We still interpret K^d and L^d as factor demands per time unit by a representative firm. In the initial short time interval the factor markets

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

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

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 $\Delta t \rightarrow 0$. A dot over a variable denotes the time derivative, i.e., $\dot{x}(t) = dx(t)/dt$. The adjustment process 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 $t \rightarrow \infty$?

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,

$$\begin{aligned}
\dot{x}(t) &= 0.5 - w(t), && (2.40) \\
\dot{w}(t) &= x(t) - 1. && (2.41)
\end{aligned}$$

Using the theory of coupled linear differential equations, the solution is[25]

$$\begin{aligned}
x(t) &= 1 + (x(0) - 1)\cos t - (w(0) - 0.5)\sin t, && (2.42) \\
w(t) &= 0.5 + (w(0) - 0.5)\cos t + (x(0) - 1)\sin t. && (2.43)
\end{aligned}$$

[25] 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 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 depreciation or the *rate of geometric decay.*[26] When changes in relative prices can occur, the rate of decay 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. Note that if the capital good itself is the numeraire, so that $p_{t-1} = p_t = 1$, then the economic depreciation coincides with the rate of geometric decay, δ.

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).[27]

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

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

[26] The latter name comes from the fact that if no investment occurs, then $K_t = K_{t-1} - \delta K_{t-1}$ and thus $K_t = (1-\delta)^t K_0$.

[27] 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 $p(t)$. Here δ appears as the rate of exponential decay, since, in case of no investment, $\dot{K}(t) = \delta K(t)$, hence $K(t) = K(0)e^{-\delta t}$.

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

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

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 versus 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 $(A_i, B_i) = (f(k_i)/k_i, f(k_i))$, $i = 1, 2, \ldots$.[28] We say that ex ante, depending on the relative factor prices as they are "now" and are expected to evolve in the future, 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

[28] 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.

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, $i = 1, 2, \dots, n$, 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.[29] 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).

[29] We ignore production for inventory holding.

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

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

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.[30] 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.

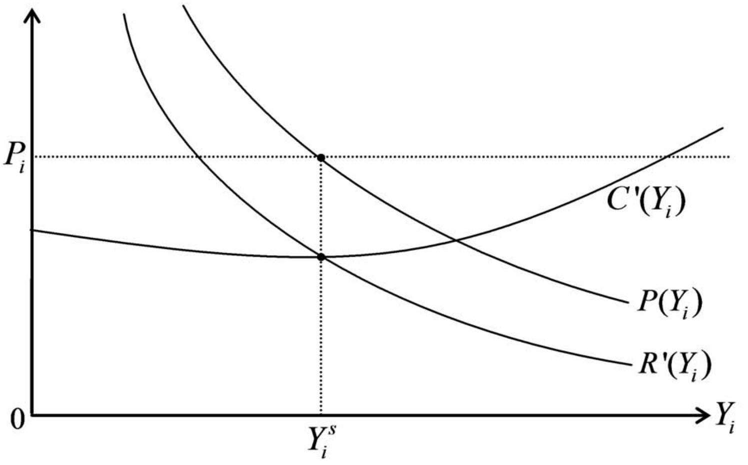

Figure 2.5: Determination of the monopolist price and output.

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

[30] 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.

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

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

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 that the markup is thus a *constant.*

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 $P_i = P$, $i = 1, 2, \dots, n$. 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.[31] 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, **?** offers an account of the econometric difficulties associated with estimating production functions. Studies by **?** and **?** suggest returns to scale are about constant or decreasing. Studies by **?**, **?**, **?**, **?**, and **?** 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 received more attention in contemporary 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 with respect to inputs are time-independent and thereby not affected by technological change. For the case where the inputs in question are renewable and nonrenewable natural resources,

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

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.

Embodied technological progress, sometimes called investment-specific technological progress, is explored in, for instance, Solow (1960), Greenwood et al. (1997), and Groth and Wendner (2015).

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 **?** and **?**. Reviews of the "Cambridge Controversy" are contained in Mas-Colell (1989) and **?**. 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), **?**, Jorgenson et al. (2005), and **?**. For a stochastic approach to aggregation, see e.g. Gallegati et al., 2006.

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 **?**, where further perspectives are discussed. It may be argued questions about stability 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. This is a complicated field, and the theory seems not settled. Yet it may be 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 Porter (1975), **?**, Osborne and Rubinstein (1990), **?**, and Foley (2010) 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., **?**.

2.7 Appendix

A. Strict quasiconcavity

Consider a function $f : \mathcal{A} \to \mathbb{R}$, where $\mathcal{A}$ is a convex set, $\mathcal{A} \subseteq \mathbb{R}^n$.[32] 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, $i = 1, \ldots, n$.[33] 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[34] 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

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

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

[33] 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).

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

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

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

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 with respect to 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

Claim (iv) in Section 2.1.3 is 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 with respect to K and L, respectively, gives, after ordering,

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

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

$$\begin{aligned} F_{KK} &< 0 \text{ implies } F_{KL} > 0, & (2.57)\\ F_{LL} &< 0 \text{ implies } F_{KL} > 0. & (2.58)\end{aligned}$$

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 microeconomic 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.

Claim (v) in Section 2.1.3 is 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,

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

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

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

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