Preface

This is a collection of earlier separate lecture notes in Economic Growth. The notes have been used in recent years in the course Economic Growth within the Master's Program in Economics at the Department of Economics, University of Copenhagen.

Compared with the earlier versions of the lecture notes some chapters have been extended and in some cases divided into several chapters. In addition, discovered typos and similar have been corrected. In some of the chapters a terminal list of references is at present lacking.

The lecture notes are in no way intended as a substitute for the currently applied textbook: D. Acemoglu, *Introduction to Modern Economic Growth*, Princeton University Press, 2009. The lecture notes are meant to be read along with the textbook. Some parts of the lecture notes are alternative presentations of stuff also covered in the textbook, while many other parts are complementary in the sense of presenting additional material. Sections marked by an asterisk, *, are cursory reading.

For constructive criticism I thank Niklas Brønager, class instructor since 2012, and plenty of earlier students. No doubt, obscurities remain. Hence, I very much welcome comments and suggestions of any kind relating to these lecture notes.

February 2015

Christian Groth

Chapter 1

Introduction to economic growth

This introductory lecture note is a refresher on basic concepts.

Section 1.1 defines Economic Growth as a field of economics. In Section 1.2 formulas for calculation of compound average growth rates in discrete and continuous time are presented. Section 1.3 briefly presents two sets of what is by many considered as "stylized facts" about economic growth. Finally, Section 1.4 discusses, in an informal way, the different concepts of cross-country income convergence. In his introductory Chapter 1, §1.5, Acemoglu[1] briefly touches upon these concepts.

1.1 The field

Economic growth analysis is the study of what factors and mechanisms determine the time path of *productivity* (a simple index of productivity is output per unit of labor). The focus is on

- productivity levels and
- productivity growth.

1.1.1 Economic growth theory

Economic growth theory endogenizes productivity growth via considering human capital accumulation (formal education as well as learning-by-doing)

[1] Throughout these lecture notes, "Acemoglu" refers to Daron Acemoglu, *Introduction to Modern Economic Growth*, Princeton University Press: Oxford, 2009.

and endogenous research and development. Also the conditioning role of geography and juridical, political, and cultural institutions is taken into account.

Although for practical reasons, economic growth theory is often stated in terms of national income and product account variables like per capita GDP, the term "economic growth" may be interpreted as referring to something deeper. We could think of "economic growth" as the widening of the opportunities of human beings to lead freer and more worthwhile lives.

To make our complex economic environment accessible for theoretical analysis we use economic models. What *is* an economic model? It is a way of organizing one's thoughts about the economic functioning of a society. A more specific answer is to define an economic model as a conceptual structure based on a set of mathematically formulated assumptions which have an economic interpretation and from which empirically testable predictions can be derived. In particular, an economic growth model is an economic model concerned with productivity issues. The union of connected and non-contradictory models dealing with economic growth and the theorems derived from these constitute an *economic growth theory*. Occasionally, intense controversies about the validity of different growth theories take place.

The terms "New Growth Theory" and "endogenous growth theory" refer to theory and models which attempt at explaining sustained per capita growth as an outcome of internal mechanisms in the model rather than just a reflection of exogenous technical progress as in "Old Growth Theory".

Among the themes addressed in this course are:

- How is the world income distribution evolving?
- Why do living standards differ so much across countries and regions? Why are some countries 50 times richer than others?
- Why do per capita growth rates differ over long periods?
- What are the roles of human capital and technology innovation in economic growth? Getting the questions right.
- Catching-up and increased speed of communication and technology diffusion.
- Economic growth, natural resources, and the environment (including the climate). What are the limits to growth?
- Policies to ignite and sustain productivity growth.

- The prospects of growth in the future.

The course concentrates on *mechanisms* behind the evolution of productivity in the industrialized world. We study these mechanisms as integral parts of dynamic general equilibrium models. The exam is a test of the extent to which the student has acquired understanding of these models, is able to evaluate them, from both a theoretical and empirical perspective, and is able to use them to analyze specific economic questions. The course is calculus intensive.

1.1.2 Some long-run data

Let Y denote real GDP (per year) and let N be population size. Then Y/N is GDP per capita. Further, let g_Y denote the average (compound) growth rate of Y per year since 1870 and let $g_{Y/N}$ denote the average (compound) growth rate of Y/N per year since 1870. Table 1.1 gives these growth rates for four countries.

	g_Y	$g_{Y/N}$
Denmark	2,67	1,87
UK	1,96	1,46
USA	3,40	1,89
Japan	3,54	2,54

Table 1.1: Average annual growth rate of GDP and GDP per capita in percent, 1870–2006. Discrete compounding. Source: Maddison, A: The World Economy: Historical Statistics, 2006, Table 1b, 1c and 5c.

Figure 1.1 displays the time path of annual GDP and GDP per capita in Denmark 1870-2006 along with regression lines estimated by OLS (logarithmic scale on the vertical axis). Figure 1.2 displays the time path of GDP per capita in UK, USA, and Japan 1870-2006. In both figures the average annual growth rates are reported. In spite of being based on exactly the same data as Table 1.1, the numbers are slightly different. Indeed, the numbers in the figures are slightly lower than those in the table. The reason is that discrete compounding is used in Table 1.1 while continuous compounding is used in the two figures. These two alternative methods of calculation are explained in the next section.

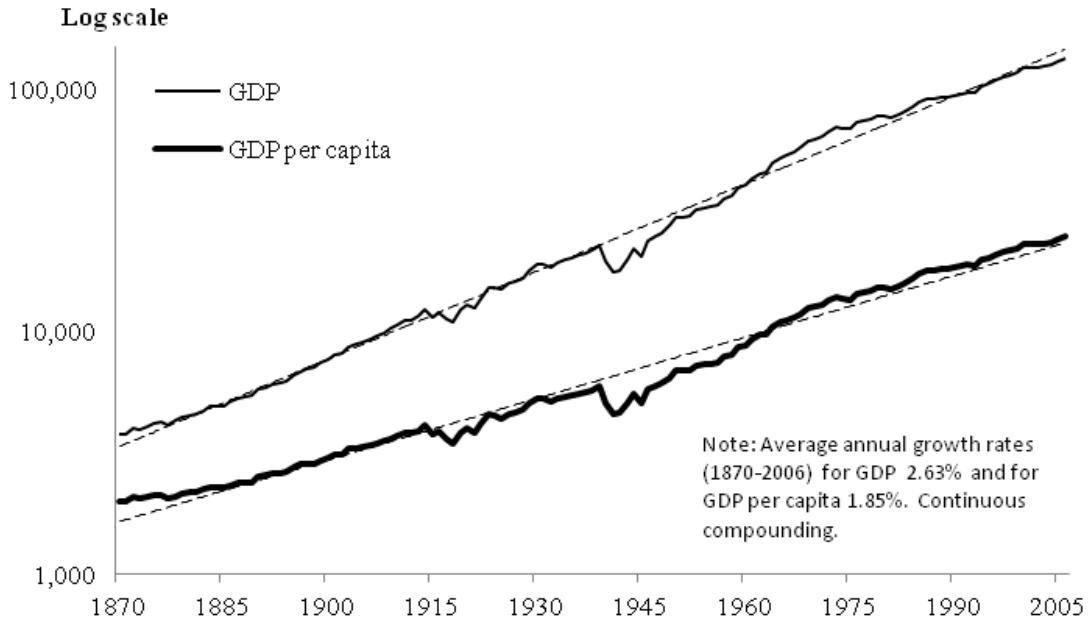


Figure 1.1: GDP and GDP per capita (1990 International Geary-Khamis dollars) in Denmark, 1870-2006. Source: Maddison, A. (2009). Statistics on World Population, GDP and Per Capita GDP, 1-2006 AD, www.ggdc.net/maddison.

1.2 Calculation of the average growth rate

1.2.1 Discrete compounding

Let y denote aggregate labor productivity, i.e., $y \equiv Y/L$, where L is employment. The average growth rate of y from period 0 to period t, with discrete compounding, is that G which satisfies

$$\begin{aligned} y_t &= y_0(1+G)^t, \qquad t = 1, 2, ... , \qquad \text{or} \qquad (1.1) \\ 1+G &= (\frac{y_t}{y_0})^{1/t}, \text{ i.e.,} \\ G &= (\frac{y_t}{y_0})^{1/t} - 1. \qquad (1.2) \end{aligned}$$

"Compounding" means adding the one-period "net return" to the "principal" before adding next period's "net return" (like with interest on interest, also called "compound interest"). The growth factor $1 + G$ will generally be less than the arithmetic average of the period-by-period growth factors. To

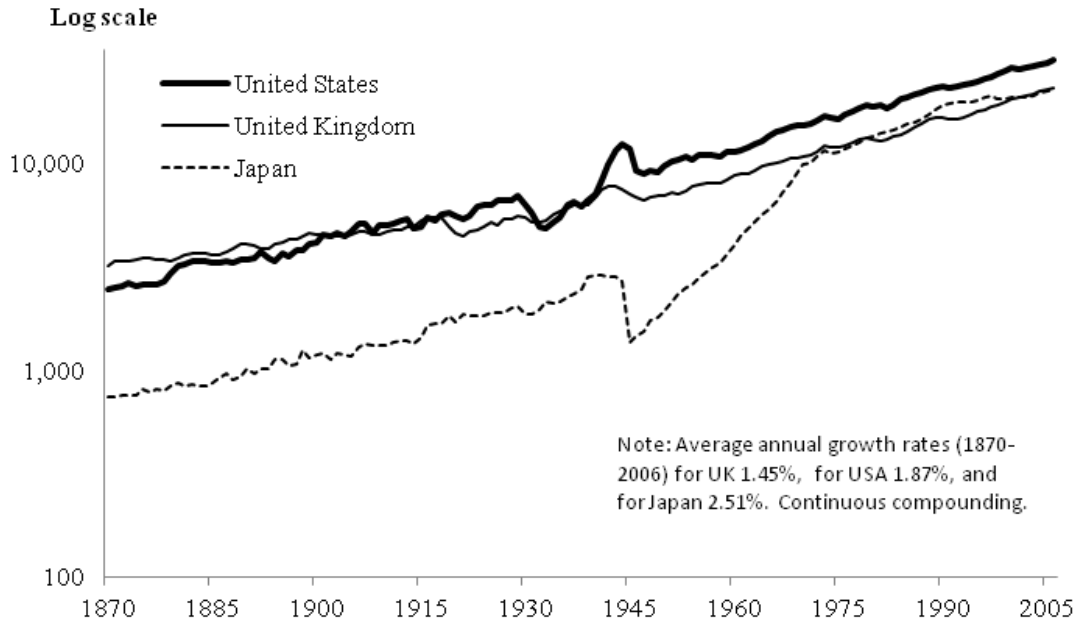


Figure 1.2: GDP per capita (1990 International Geary-Khamis dollars) in UK, USA and Japan, 1870-2006. Source: Maddison, A. (2009). Statistics on World Population, GDP and Per Capita GDP, 1-2006 AD, www.ggdc.net/maddison.

underline this difference, $1+G$ is sometimes called the "compound average growth factor" or the "geometric average growth factor" and G itself then called the "compound average growth rate" or the "geometric average growth rate"

Using a pocket calculator, the following steps in the calculation of G may be convenient. Take logs on both sides of (1.1) to get

$$\begin{aligned} \ln \frac{y_t}{y_0} &= t \ln(1+G) \Rightarrow \\ \ln(1+G) &= \frac{\ln \frac{y_t}{y_0}}{t} \Rightarrow \qquad (1.3) \\ G &= \text{antilog}(\frac{\ln \frac{y_t}{y_0}}{t}) - 1. \qquad (1.4) \end{aligned}$$

Note that t in the formulas (1.2) and (1.4) equals the number of periods *minus 1.*

1.2.2 Continuous compounding

The average growth rate of y, with continuous compounding, is that g which satisfies

$$y_t = y_0 e^{gt}, \tag{1.5}$$

where e denotes the Euler number, i.e., the base of the natural logarithm.[2] Solving for g gives

$$g = \frac{\ln \frac{y_t}{y_0}}{t} = \frac{\ln y_t - \ln y_0}{t}. \tag{1.6}$$

The first formula in (1.6) is convenient for calculation with a pocket calculator, whereas the second formula is perhaps closer to intuition. Another name for g is the "exponential average growth rate".

Again, for discrete time data the t in the formula equals the number of periods minus 1.

Comparing with (1.3) we see that $g = \ln(1+G) < G$ for $G \neq 0$. Yet, by a first-order Taylor approximation of $\ln(1+G)$ about $G = 0$ we have

$$g = \ln(1+G) \approx G \quad \text{for } G \text{ "small"}. \tag{1.7}$$

For a given data set the G calculated from (1.2) will be slightly above the g calculated from (1.6), cf. the mentioned difference between the growth rates in Table 1.1 and those in Figure 1.1 and Figure 1.2. The reason is that a given growth force is more powerful when compounding is continuous rather than discrete. Anyway, the difference between G and g is usually unimportant. If for example G refers to the annual GDP growth rate, it will be a small number, and the difference between G and g immaterial. For example, to $G = 0.040$ corresponds $g \approx 0.039$. Even if $G = 0.10$, the corresponding g is 0.0953. But if G stands for the inflation rate and there is high inflation, the difference between G and g will be substantial. During hyperinflation the monthly inflation rate may be, say, $G = 100\%$, but the corresponding g will be only 69%.

Which method, discrete or continuous compounding, is preferable? To some extent it is a matter of taste or convenience. In period analysis discrete compounding is most common and in continuous time analysis continuous compounding is most common.

For calculation with a pocket calculator the continuous compounding formula, (1.6), is slightly easier to use than the discrete compounding formulas, whether (1.2) or (1.4).

[2] Unless otherwise specified, whenever we write $\ln x$ or $\log x$, the *natural* logarithm is understood.

To avoid too much sensitiveness to the initial and terminal observations, which may involve measurement error or depend on the state of the business cycle, one can use an OLS approach to the trend coefficient, g, in the following regression:

$$\ln Y_t = \alpha + gt + \varepsilon_t.$$

This is in fact what is done in Fig. 1.1.

1.2.3 Doubling time

How long time does it take for y to double if the growth rate with discrete compounding is G? Knowing G, we rewrite the formula (1.3):

$$t = \frac{\ln \frac{y_t}{y_0}}{\ln(1+G)} = \frac{\ln 2}{\ln(1+G)} \approx \frac{0.6931}{\ln(1+G)}.$$

With $G = 0.0187$, cf. Table 1.1, we find

$$t \approx 37.4 \text{ years},$$

meaning that productivity doubles every 37.4 years.

How long time does it take for y to double if the growth rate with continuous compounding is g? The answer is based on rewriting the formula (1.6):

$$t = \frac{\ln \frac{y_t}{y_0}}{g} = \frac{\ln 2}{g} \approx \frac{0.6931}{g}.$$

Maintaining the value 0.0187 also for g, we find

$$t \approx \frac{0.6931}{0.0187} \approx 37.1 \text{ years}.$$

Again, with a pocket calculator the continuous compounding formula is slightly easier to use. With a lower g, say $g = 0.01$, we find doubling time equal to 69.1 years. With $g = 0.07$ (think of China since the early 1980's), doubling time is about 10 years! Owing to the compounding, exponential growth is extremely powerful.

1.3 Some stylized facts of economic growth

1.3.1 The Kuznets facts

A well-known characteristic of modern economic growth is structural change: unbalanced sectorial growth. There is a massive reallocation of labor from

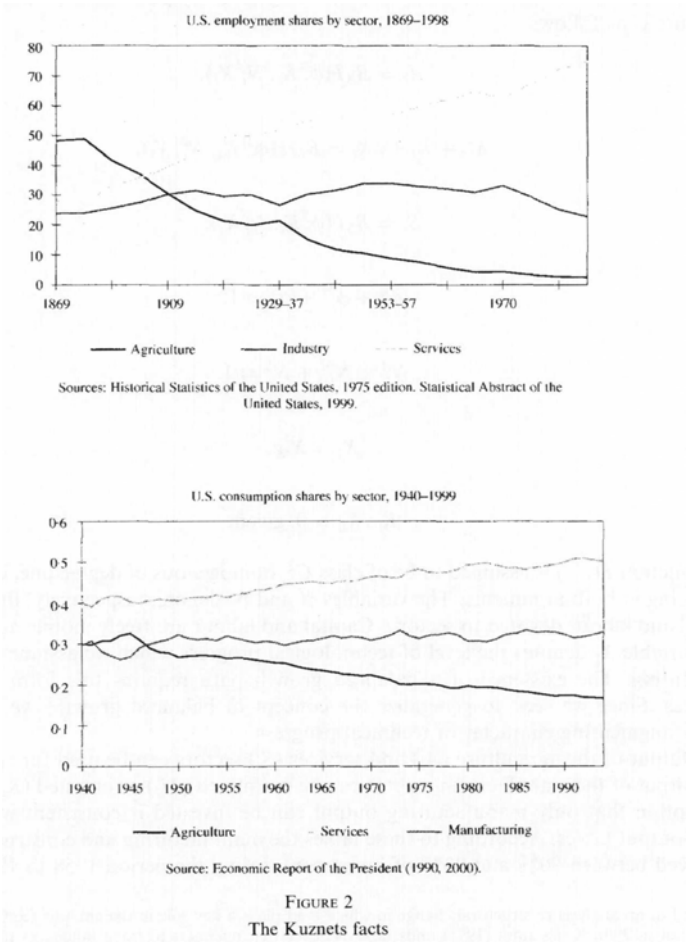


Figure 1.3: The Kuznets facts. Source: Kongsamut et al., Beyond Balanced Growth, Review of Economic Studies, vol. 68, Oct. 2001, 869-82.

agriculture into industry (manufacturing, construction, and mining) and further into services (including transport and communication). The shares of total consumption expenditure going to these three sectors have moved similarly. Differences in the demand elasticities with respect to income seem the main explanation. These observations are often referred to as the *Kuznets facts* (after Simon Kuznets, 1901-85, see, e.g., Kuznets 1957).

The two graphs in Figure 1.3 illustrate the Kuznets facts.

1.3.2 Kaldor's stylized facts

Surprisingly, in spite of the Kuznets facts, the evolution at the *aggregate* level in developed countries is by many economists seen as roughly described by what is called Kaldor's "stylized facts" (after the Hungarian-British economist Nicholas Kaldor, 1908-1986, see, e.g., Kaldor 1957, 1961)[3]:

1. Real output per man-hour grows at a more or less constant rate over fairly long periods of time. (Of course, there are short-run fluctuations superposed around this trend.)

2. The stock of physical capital per man-hour grows at a more or less constant rate over fairly long periods of time.

3. The ratio of output to capital shows no systematic trend.

4. The rate of return to capital shows no systematic trend.

5. The income shares of labor and capital (in the national accounting sense, i.e., including land and other natural resources), respectively, are nearly constant.

6. The growth rate of output per man-hour differs substantially across countries.

These claimed regularities do certainly not fit all developed countries equally well. Although Solow's growth model (Solow, 1956) can be seen as the first successful attempt at building a model consistent with Kaldor's "stylized facts", Solow once remarked about them: "There is no doubt that they are stylized, though it is possible to question whether they are facts" (Solow, 1970). Yet, for instance a relatively recent study by Attfield and Temple (2010) of US and UK data since the Second World War is not unfavorable to Kaldor's "facts". The sixth Kaldor fact is, of course, generally accepted as a well documented observation (a nice summary is contained in Pritchett, 1997).

Kaldor also proposed hypotheses about the links between growth in the different sectors (see, e.g., Kaldor 1967):

a. Productivity growth in the manufacturing and construction sectors is enhanced by output growth in these sectors (this is also known as Verdoorn's Law). Increasing returns to scale and learning by doing are the main factors behind this.

b. Productivity growth in agriculture and services is enhanced by output growth in the manufacturing and construction sectors.

[3]Kaldor presented his six regularities as "a stylised view of the facts".

1.4 Concepts of income convergence

The two most popular across-country income convergence concepts are "β convergence" and "σ convergence".

1.4.1 β convergence vs. σ convergence

Definition 1 *We say that β convergence occurs for a given selection of countries if there is a tendency for the poor (those with low income per capita or low output per worker) to subsequently grow faster than the rich.*

By "grow faster" is meant that the growth rate of per capita income (or per worker output) is systematically higher.

In many contexts, a more appropriate convergence concept is the following:

Definition 2 *We say that σ convergence, with respect to a given measure of dispersion, occurs for a given collection of countries if this measure of dispersion, applied to income per capita or output per worker across the countries, declines systematically over time. On the other hand, σ divergence occurs, if the dispersion increases systematically over time.*

The reason that σ convergence must be considered the more appropriate concept is the following. In the end, it is the question of increasing or decreasing dispersion across countries that we are interested in. From a superficial point of view one might think that β convergence implies decreasing dispersion and vice versa, so that β convergence and σ convergence are more or less equivalent concepts. But since the world is not deterministic, but stochastic, this is not true. Indeed, β convergence is only a necessary, not a sufficient condition for σ convergence. This is because over time some reshuffling among the countries is always taking place, and this implies that there will always be some extreme countries (those initially far away from the mean) that move closer to the mean, thus creating a negative correlation between initial level and subsequent growth, in spite of equally many countries moving from a middle position toward one of the extremes.[4] In this way β convergence may be observed at the same time as there is no σ

[4] As an intuitive analogy, think of the ordinal rankings of the sports teams in a league. The dispersion of rankings is constant by definition. Yet, no doubt there will allways be some tendency for weak teams to rebound toward the mean and of champions to revert to mediocrity. (This example is taken from the first edition of Barro and Sala-i-Martin, *Economic Growth*, 1995; I do not know why, but the example was deleted in the second edition from 2004.)

convergence; the mere presence of random measurement errors implies a bias in this direction because a growth rate depends negatively on the initial measurement and positively on the later measurement. In fact, β convergence may be consistent with σ *divergence* (for a formal proof of this claim, see Barro and Sala-i-Martin, 2004, pp. 50-51 and 462 ff.; see also Valdés, 1999, p. 49-50, and Romer, 2001, p. 32-34).

Hence, it is wrong to conclude from β convergence (poor countries tend to grow faster than rich ones) to σ convergence (reduced dispersion of per capita income) without any further investigation. The mistake is called "regression towards the mean" or "Galton's fallacy". Francis Galton was an anthropologist (and a cousin of Darwin), who in the late nineteenth century observed that tall fathers tended to have not as tall sons and small fathers tended to have taller sons. From this he falsely concluded that there was a tendency to averaging out of the differences in height in the population. Indeed, being a true aristocrat, Galton found this tendency pitiable. But since his conclusion was mistaken, he did not really have to worry.

Since σ convergence comes closer to what we are ultimately looking for, from now, when we speak of just "income convergence", σ convergence is understood.

In the above definitions of σ convergence and β convergence, respectively, we were vague as to what kind of selection of countries is considered. In principle we would like it to be a representative sample of the "population" of countries that we are interested in. The population could be all countries in the world. Or it could be the countries that a century ago had obtained a certain level of development.

One should be aware that historical GDP data are constructed retrospectively. Long time series data have only been constructed for those countries that became relatively rich during the after-WWII period. Thus, if we as our sample select the countries for which long data series exist, what is known as *selection bias* is involved which generates a spurious convergence. A country which was poor a century ago will only appear in the sample if it grew rapidly over the next 100 years. A country which was relatively rich a century ago will appear in the sample unconditionally. This selection bias problem was pointed out by DeLong (1988) in a criticism of widespread false interpretations of Maddison's long data series (Maddison 1982).

1.4.2 Measures of dispersion

Our next problem is: *what* measure of dispersion is to be used as a useful descriptive statistics for σ convergence? Here there are different possibilities.

To be precise about this we need some notation. Let

$$\begin{aligned} y &\equiv \frac{Y}{L}, \qquad \text{and} \\ q &\equiv \frac{Y}{N}, \end{aligned}$$

where Y = real GDP, L = employment, and N = population. If the focus is on living standards, Y/N, is the relevant variable.[5] But if the focus is on (labor) productivity, it is Y/L, that is relevant. Since most growth models focus on Y/L rather than Y/N, let os take y as our example.

One might think that the standard deviation of y could be a relevant measure of dispersion when discussing whether σ convergence is present or not. The *standard deviation* of y across n countries in a given year is

$$\sigma_y \equiv \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{1.8}$$

where

$$\bar{y} \equiv \frac{\sum_i y_i}{n}, \tag{1.9}$$

i.e., $\bar{y}$ is the average output per worker. However, if this measure were used, it would be hard to find *any* group of countries for which there is income convergence. This is because y tends to grow over time for most countries, and then there is an inherent tendency for the variance also to grow; hence also the square root of the variance, σ_y, tends to grow. Indeed, suppose that for all countries, y is doubled from time t_1 to time t_2. Then, automatically, σ_y is also doubled. But hardly anyone would interpret this as an increase in the income inequality across the countries.

Hence, it is more adequate to look at the standard deviation of *relative* income levels:

$$\sigma_{y/\bar{y}} \equiv \sqrt{\frac{1}{n}\sum_{i}(\frac{y_i}{\bar{y}} - 1)^2}. \tag{1.10}$$

This measure is the same as what is called the *coefficient of variation*, CV_y, usually defined as

$$CV_y \equiv \frac{\sigma_y}{\bar{y}}, \tag{1.11}$$

[5] Or perhaps better, Q/N, where $Q \equiv GNP \equiv GDP - rD - wF$. Here, rD, denotes net interest payments on foreign debt and wF denotes net labor income of foreign workers in the country.

that is, the standard deviation of y standardized by the mean. That the two measures are identical can be seen in this way:

$$\frac{\sigma_y}{\bar{y}} \equiv \frac{\sqrt{\frac{1}{n}\sum_i (y_i - \bar{y})^2}}{\bar{y}} = \sqrt{\frac{1}{n}\sum_i (\frac{y_i - \bar{y}}{\bar{y}})^2} = \sqrt{\frac{1}{n}\sum_i (\frac{y_i}{\bar{y}} - 1)^2} \equiv \sigma_{y/\bar{y}}.$$

The point is that the coefficient of variation is "scale free", which the standard deviation itself is not.

Instead of the coefficient of variation, another scale free measure is often used, namely the standard deviation of $\ln y$, i.e.,

$$\sigma_{\ln y} \equiv \sqrt{\frac{1}{n}\sum_i (\ln y_i - \ln y^*)^2}, \tag{1.12}$$

where

$$\ln y^* \equiv \frac{\sum_i \ln y_i}{n}. \tag{1.13}$$

Note that y^* is the geometric average, i.e., $y^* \equiv \sqrt[n]{y_1 y_2 \cdots y_n}$. Now, by a first-order Taylor approximation of $\ln y$ around $y = \bar{y}$, we have

$$\ln y \approx \ln \bar{y} + \frac{1}{\bar{y}}(y - \bar{y})$$

Hence, as a very rough approximation we have $\sigma_{\ln y} \approx \sigma_{y/\bar{y}} = CV_y$, though this approximation can be quite poor (cf. Dalgaard and Vastrup, 2001). It may be possible, however, to defend the use of $\sigma_{\ln y}$ in its own right to the extent that y tends to be approximately lognormally distributed across countries.

Yet another possible measure of income dispersion across countries is the *Gini index* (see for example Cowell, 1995).

1.4.3 Weighting by size of population

Another important issue is whether the applied dispersion measure is based on a *weighting of the countries by size of population*. For the world as a whole, when no weighting by size of population is used, then there is a slight tendency to income divergence according to the $\sigma_{\ln q}$ criterion (Acemoglu, 2009, p. 4), where q is per capita income ($\equiv Y/N$). As seen by Fig. 4 below, this tendency is not so clear according to the CV_q criterion. Anyway, when there *is* weighting by size of population, then in the last twenty years there has been a tendency to income convergence at the global level (Sala-i-Martin

2006; Acemoglu, 2009, p. 6). With weighting by size of population (1.12) is modified to

$$\sigma^w_{\ln q} \equiv \sqrt{\sum_i w_i (\ln q_i - \ln q^*)^2},$$

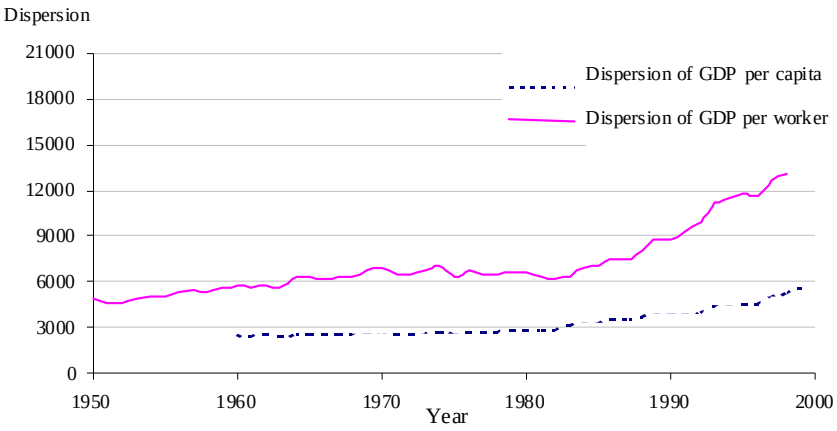
where

$$w_i = \frac{N_i}{N} \quad \text{and} \quad \ln q^* \equiv \sum_i w_i \ln q_i.$$

1.4.4 Unconditional vs. conditional convergence

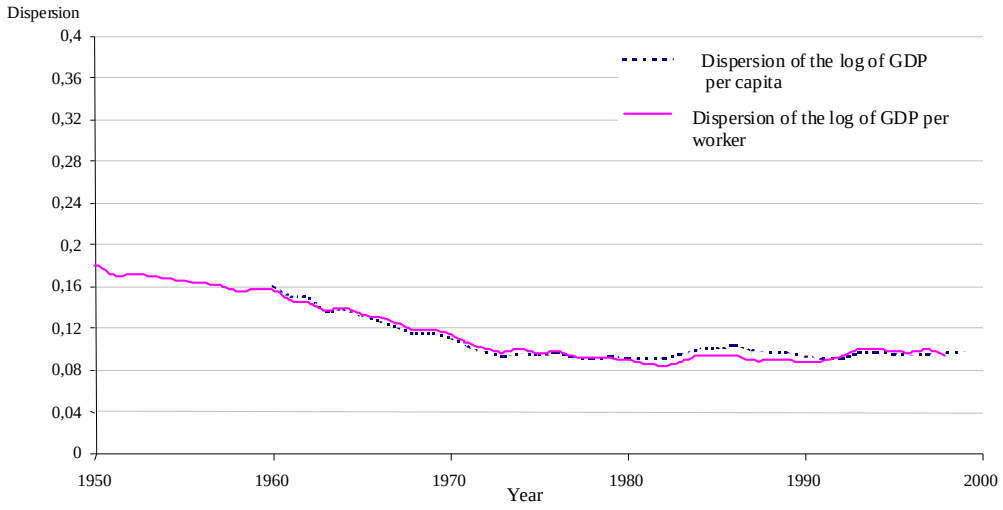
Yet another distinction in the study of income convergence is that between unconditional (or absolute) and conditional convergence. We say that a large heterogeneous group of countries (say the countries in the world) show *unconditional* income convergence if income convergence occurs for the whole group without conditioning on specific characteristics of the countries. If income convergence occurs only for a subgroup of the countries, namely those countries that in advance share the same "structural characteristics", then we say there is *conditional* income convergence. As noted earlier, when we speak of just income "convergence", income "σ convergence" is understood. If in a given context there might be doubt, one should of course be explicit and speak of unconditional or conditional σ convergence. Similarly, if the focus for some reason is on β convergence, we should distinguish between unconditional and conditional β convergence.

What the precise meaning of "structural characteristics" is, will depend on what model of the countries the researcher has in mind. According to the Solow model, a set of relevant "structural characteristics" are: the aggregate production function, the initial level of technology, the rate of technical progress, the capital depreciation rate, the saving rate, and the population growth rate. But the Solow model, as well as its extension with human capital (Mankiw et al., 1992), is a model of a closed economy with exogenous technical progress. The model deals with "within-country" convergence in the sense that the model predicts that a closed economy being initially below or above its steady state path, will over time converge towards its steady state path. It is far from obvious that this kind of model is a good model of cross-country convergence in a globalized world where capital mobility and to some extent also labor mobility are important and some countries are pushing the technological frontier further out, while others try to imitate and catch up.



Remarks: Germany is not included in GDP per worker. GDP per worker is missing for Sweden and Greece in 1950, and for Portugal in 1998. The EU comprises Belgium, Denmark, Finland, France, Greece, Holland, Ireland, Italy, Luxembourg, Portugal, Spain, Sweden, Germany, the UK and Austria.
Source: Pwt6, OECD Economic Outlook No. 65 1999 via Eco Win and World Bank Global Development Network Growth Database.

Figure 1.4: Standard deviation of GDP per capita and per worker across 12 EU countries, 1950-1998.



Remarks: Germany is not included in GDP per worker. GDP per worker is missing for Sweden and Greece in 1950, and for Portugal in 1998. The EU comprises Belgium, Denmark, Finland, France, Greece, Holland, Ireland, Italy, Luxembourg, Portugal, Spain, Sweden, Germany, the UK and Austria.
Source: Pwt6, OECD Economic Outlook No. 65 1999 via Eco Win and World Bank Global Development Network Growth Database.

Figure 1.5: Standard deviation of the log of GDP per capita and per worker across 12 EU countries, 1950-1998.

1.4.5 A bird's-eye view of the data

In the following no serious econometrics is attempted. We use the term "trend" in an admittedly loose sense.

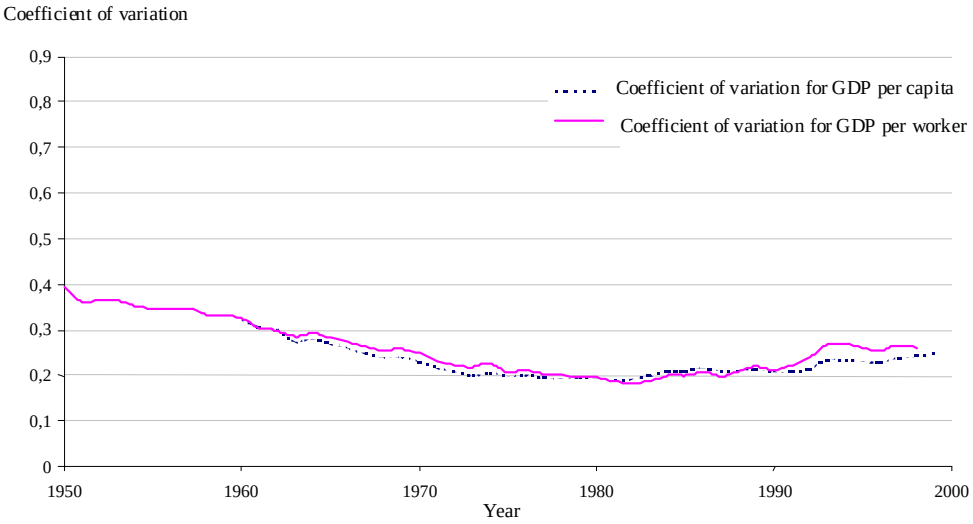
Figure 1.4 shows the time profile for the standard deviation of y itself for 12 EU countries, whereas Figure 1.5 and Figure 1.6 show the time profile of the standard deviation of $\log y$ and the time profile of the coefficient of variation, respectively. Comparing the upward trend in Figure 1.4 with the downward trend in the two other figures, we have an illustration of the fact that the movement of the standard deviation of y itself does not capture income convergence. To put it another way: although there seems to be conditional income convergence with respect to the two scale-free measures, Figure 1.4 shows that this tendency to convergence is *not* so strong as to produce a narrowing of the absolute distance between the EU countries.[6]

Figure 1.7 shows the time path of the coefficient of variation across 121 countries in the world, 22 OECD countries and 12 EU countries, respectively. We see the lack of unconditional income convergence, but the presence of conditional income convergence. One should not over-interpret the observation of convergence for the 22 OECD countries over the period 1950-1990. It is likely that this observation suffer from the selection bias problem mentioned in Section 1.4.1. A country that was poor in 1950 will typically have become a member of OECD only if it grew relatively fast afterwards.

1.4.6 Other convergence concepts

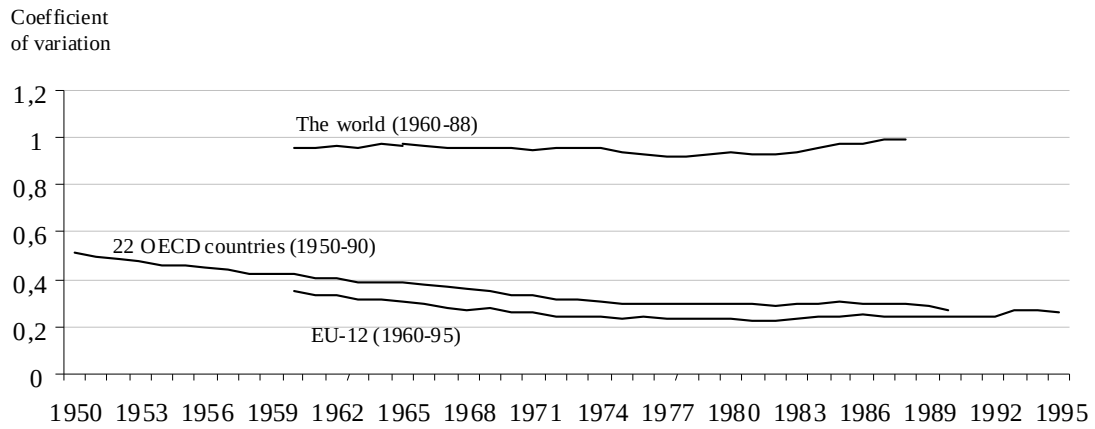
Of course, just considering the time profile of the first and second moments of a distribution may sometimes be a poor characterization of the evolution of the distribution. For example, there are signs that the distribution has polarized into *twin peaks* of rich and poor countries (Quah, 1996a; Jones, 1997). Related to this observation is the notion of club convergence. If income convergence occurs *only* among a subgroup of the countries that to some extent share the same initial conditions, then we say there is *club-convergence.* This concept is relevant in a setting where there are *multiple* steady states toward which countries can converge. At least at the theoretical level multiple steady states can easily arise in overlapping generations models. Then the initial condition for a given country matters for which of these steady states this country is heading to. Similarly, we may say that *conditional club-convergence* is present, if income convergence occurs *only*

[6] Unfortunately, sometimes misleading graphs or texts to graphs about across-country income convergence are published. In the collection of exercises, Chapter 1, you are asked to discuss some examples of this.



Remarks: Germany is not included in GDP per worker. GDP per worker is missing for Sweden and Greece in 1950, and for Portugal in 1998. The EU comprises Belgium, Denmark, Finland, France, Greece, Holland, Ireland, Italy, Luxembourg, Portugal, Spain, Sweden, Germany, the UK and Austria.
Source: Pwt6, OECD Economic Outlook No. 65 1999 via Eco Win and World Bank Global Development Network Growth Database.

Figure 1.6: Coefficient of variation of GDP per capita and GDP per worker across 12 EU countries, 1950-1998.



Remarks: The world' comprises 121 countries (not weighed by size) where complete time series for GDP per capita exist.
The OECD countries exclude South Korea, Hungary, Poland, Iceland, Czech Rep., Luxembourg and Mexico.
EU-12 comprises: Benelux, Germany, France, Italy, Denmark, Ireland, UK, Spain, Portugal og Greece.
Source: Penn World Table 5.6 and OECD Economic Outlook, Statistics on Microcomputer Disc, December 1998.

Figure 1.7: Coefficient of variation of income per capita across different sets of countries.

for a subgroup of the countries, namely countries sharing similar structural characteristics (this may to some extent be true for the OECD countries) *and*, within an interval, similar initial conditions.

Instead of focusing on income convergence, one could study *TFP convergence* at aggregate or industry level.[7] Sometimes the less demanding concept of *growth rate convergence* is the focus.

The above considerations are only of a very elementary nature and are only about descriptive statistics. The reader is referred to the large existing literature on concepts and econometric methods of relevance for characterizing the evolution of world income distribution (see Quah, 1996b, 1996c, 1997, and for a survey, see Islam 2003).

1.5 Literature

Acemoglu, D., 2009, *Introduction to Modern Economic Growth*, Princeton University Press: Oxford.

Acemoglu, D., and V. Guerrieri, 2008, Capital deepening and nonbalanced

[7] See, for instance, Bernard and Jones 1996a and 1996b.

economic growth, *J. Political Economy,* vol. 116 (3), 467- .

Attfield, C., and J.R.W. Temple, 2010, Balanced growth and the great ratios: New evidence for the US and UK, *Journal of Macroeconomics*, vol. 32, 937-956.

Barro, R. J., and X. Sala-i-Martin, 1995, *Economic Growth,* MIT Press, New York. Second edition, 2004.

Bernard, A.B., and C.I. Jones, 1996a, ..., *Economic Journal.*

- , 1996b, Comparing Apples to Oranges: Productivity Convergence and Measurement Across Industries and Countries, *American Economic Review*, vol. 86 (5), 1216-1238.

Cowell, Frank A., 1995, *Measuring Inequality. 2. ed.,* London.

Dalgaard, C.-J., and J. Vastrup, 2001, On the measurement of σ-convergence, *Economics letters,* vol. 70, 283-87.

Dansk økonomi. Efterår 2001, (Det økonomiske Råds formandskab) Kbh. 2001.

Deininger, K., and L. Squire, 1996, A new data set measuring income inequality, *The World Bank Economic Review*, 10, 3.

Delong, B., 1988, ... *American Economic Review.*

Handbook of Economic Growth, vol. 1A and 1B, ed. by S. N. Durlauf and P. Aghion, Amsterdam 2005.

Handbook of Income Distribution, vol. 1, ed. by A.B. Atkinson and F. Bourguignon, Amsterdam 2000.

Islam, N., 2003, What have we learnt from the convergence debate? *Journal of Economic Surveys 17*, 3, 309-62.

Kaldor, N., 1957, A model of economic growth, *The Economic Journal,* vol. 67, pp. 591-624.

- , 1961, "Capital Accumulation and Economic Growth". In: F. Lutz, ed., *Theory of Capital,* London: MacMillan.

- , 1967, *Strategic Factors in Economic Development*, New York State School of Industrial and Labor Relations, Cornell University.

Kongsamut et al., 2001, Beyond Balanced Growth, *Review of Economic Studies*, vol. 68, 869-882.

Kuznets, S., 1957, Quantitative aspects of economic growth of nations: II, *Economic Development and Cultural Change*, Supplement to vol. 5, 3-111.

Maddison, A., 1982,

Mankiw, N.G., D. Romer, and D.N. Weil, 1992,

Pritchett, L., 1997, Divergence – big time, *Journal of Economic Perspectives*, vol. 11, no. 3.

Quah, D., 1996a, Twin peaks ..., *Economic Journal,* vol. 106*,* 1045-1055.

-, 1996b, Empirics for growth and convergence, *European Economic Review*, vol. 40 (6).

-, 1996c, Convergence empirics ..., *J. of Ec. Growth,* vol. 1 (1).

-, 1997, Searching for prosperity: A comment, *Carnegie-Rochester Conferende Series on Public Policy,* vol. 55, 305-319.

Romer, D., 2012, *Advanced Macroeconomics*, 4th ed., McGraw-Hill: New York.

Sala-i-Martin, X., 2006, The World Distribution of Income, *Quarterly Journal of Economics 121*, No. 2.

Solow, R.M., 1970, *Growth theory. An exposition,* Clarendon Press: Oxford. Second enlarged edition, 2000.

Valdés, B., 1999, *Economic Growth. Theory, Empirics, and Policy*, Edward Elgar.

On measurement problems, see: http://www.worldbank.org/poverty/inequal/methods/index.htm

Chapter 2

Review of technology

The aim of this chapter is, first, to introduce the terminology concerning firms' technology and technological change used in the lectures and exercises of this course. At a few points I deviate somewhat from definitions in Acemoglu's book. Section 1.3 can be used as a formula manual for the case of CRS.

Second, the chapter contains a brief discussion of the somewhat controversial notions of a representative firm and an aggregate production function.

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-factor production function given by

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

where Y is output (value added) per time unit, K is capital input per time unit, and L is labor input per time unit ($K \geq 0$, $L \geq 0$). We may think of (2.1) as describing the output of a firm, a sector, or the economy as a whole. It is in any case a very simplified description, ignoring the heterogeneity of output, capital, and labor. Yet, for many macroeconomic questions it may be a useful first approach. Note that in (2.1) not only Y but also K and L represent *flows,* that is, quantities per unit of time. If the time unit is one year, we think of K as measured in machine hours per year. Similarly, we think of L as measured in labor hours per year. Unless otherwise specified, it is understood that the rate of utilization of the production factors is constant over time and normalized to one for each production factor. As explained in Chapter 1, we can then use the same symbol, K, for the *flow* of capital services as for the *stock* of capital. Similarly with L.

2.1.1 A neoclassical production function

By definition, 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 specific functional forms are used to represent a production function, that function may not be defined at points where $K = 0$ or $L = 0$ or both. In such a case we adopt the convention that the domain of the function is understood extended to include such boundary points whenever it is possible to assign function values to them such that continuity is maintained. For instance the function $F(K, L) = \alpha L + \beta KL/(K + L)$, where $\alpha > 0$ and $\beta > 0$, is not defined at $(K, L) = (0, 0)$. But by assigning the function value 0 to the point $(0, 0)$, we maintain both continuity and the "no input, no output" property, cf. Exercise 2.4.

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

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

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

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

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

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

[1] For any fixed $\bar{Y} \geq 0$, the associated *isoquant* is the level set $\{(K, L) \in \mathbb{R}_+ |\ F(K, L) = \bar{Y}\}$.

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, and F is strictly quasiconcave (i.e., the upper contour sets are strictly convex, cf. Appendix A).

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

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

where the superscript on F indicates that the production function may shift over time, due to changes in technology. We then say that $F^t(\cdot)$ is a neoclassical production function if it satisfies the conditions (a) and (b) for all pairs (K_t, L_t). *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 at that point, cf. Figure 2.1. The equation $F(K, L) = \bar{Y}$ defines K as an implicit function of L. By implicit differentiation we find $F_K(K, L)dK/dL + F_L(K, L) = 0$, from which follows

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

That is, MRS_{KL} measures the amount of K that can be saved (approximately) by applying an extra unit of labor. In turn, this equals the ratio of the marginal productivities of labor and capital, respectively.[2] 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).

[2]The subscript $|Y = \bar{Y}$ in (2.5) indicates that we are moving along a given isoquant, $F(K, L) = \bar{Y}$. Expressions like, e.g., $F_L(K, L)$ or $F_2(K, L)$ mean the partial derivative of F w.r.t. the second argument, evaluated at the point (K, L).

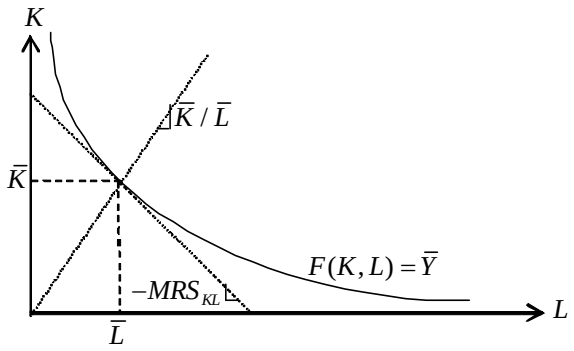


Figure 2.1: MRS_{KL} as the absolute slope of the isoquant.

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 $K = (\bar{K}/\bar{L})L$. That is, in this case the expansion path is a straight line.

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

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

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 gave above is quite common in macroeconomic journal articles and convenient because of its flexibility. 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 course, 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) and all $\lambda > 0$,

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

As all inputs are scaled up or down by some factor > 1, output is scaled up or down by the same factor.[4] The assumption of CRS is often defended by the *replication argument.* 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) and all $\lambda > 1$,

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

That is, IRS is present if, when all inputs are scaled up by some factor > 1, output is scaled up by *more* than this factor. The existence of gains by specialization and division of labor, synergy effects, etc. sometimes speak in support of this assumption, at least up to a certain level of production. The assumption is also called the *economies of scale* assumption.

[4] In their definition of a neoclassical production function some textbooks add constant returns to scale as a requirement besides (a) and (b). This course follows the alternative terminology where, if in a given context an assumption of constant returns to scale is needed, this is stated as an additional assumption.

Another possibility is *decreasing returns to scale* (DRS). This is said to occur when for all (K, L) 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 $\alpha + \beta > 1$, 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

$$Y = \min(AK, BL), \qquad A > 0, B > 0, \tag{2.9}$$

where A and B are given parameters, is called a *Leontief production function* 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 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 are limited.[5] As (2.9) 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.* To explain the content of this argument we have to first clarify the distinction between rival and nonrival inputs or more generally the distinction between rival and nonrival goods. 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. In contrast, however, technical knowledge like a farmaceutical formula or an engineering principle is *nonrival.* An unbounded number of factories can simultaneously use the same farmaceutical formula.

The replication argument now says 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. One should be aware that the CRS assumption is about *technology* in the sense of functions linking inputs to outputs − limits to the *availability* of input resources is an entirely different matter. The fact that for example managerial talent may be in limited supply does not preclude the thought experiment that *if* a firm could double all its inputs, including the number of talented managers, then the output level could also be doubled.

The replication argument presupposes, first, that *all* the relevant inputs are explicit as arguments in the production function; second, that these are changed equiproportionately. This, however, exhibits the weakness of the replication argument as a defence for assuming CRS of our present production function, $F(\cdot)$. One could easily make the case that besides capital and labor, also land is a necessary input and should appear as a separate argument.[6] If an industrial firm decides to duplicate what it has been doing, it needs a piece of land to build another plant like the first. Then, on the basis of the replication argument we should in fact expect DRS w.r.t. capital and labor alone. In manufacturing and services, empirically, this and other possible

[5] Cf. Section 2.4.

[6] We think of "capital" as producible means of production, whereas "land" refers to non-producible natural resources, including for example building sites.

sources for departure from CRS may be minor and so many macroeconomists feel comfortable enough with assuming CRS w.r.t. K and L alone, at least as a first approximation. This approximation is, however, less applicable to poor countries, where natural resources may be a quantitatively important production factor.

There is a further problem with the replication argument. Strictly speaking, the CRS claim is that by changing all the inputs equiproportionately by *any* positive factor, λ, which does not have to be an integer, the firm should be able to get output changed 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 size 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* To allow for indivisibilities and mixed cases (for example IRS at low levels of production and CRS or DRS at higher levels), we need a local measure of returns to scale. One defines the *elasticity of scale*, $\eta(K,L)$, of F at the point (K,L), where $F(K,L)>0$, as

$$\eta(K,L)=\frac{\lambda}{F(K,L)}\frac{dF(\lambda K,\lambda L)}{d\lambda}\approx\frac{\Delta F(\lambda K,\lambda L)/F(K,L)}{\Delta\lambda/\lambda},\text{ evaluated at }\lambda=1. \tag{2.10}$$

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

$$\text{if }\eta(K,L)\begin{cases}>1,\text{ then there are locally } \textit{IRS,}\\ =1,\text{ then there are locally } \textit{CRS,}\\ <1,\text{ then there are locally } \textit{DRS.}\end{cases} \tag{2.11}$$

The production function *may* have the same elasticity of scale everywhere. This is the case if and only if the production function is homogeneous. If F is homogeneous of degree h, then $\eta(K,L)=h$ and h is called the *elasticity of scale parameter*.

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

$$\eta(K,L)=\frac{F_K(K,L)K}{F(K,L)}+\frac{F_L(K,L)L}{F(K,L)}. \tag{2.12}$$

This follows from the definition in (2.10) by taking into account that

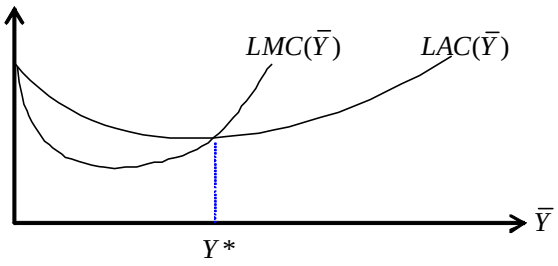


Figure 2.2: Locally CRS at optimal plant size.

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

Figure 2.2 illustrates a popular case from introductory economics, an average cost curve which from the perspective of the individual firm (or plant) is U-shaped: at low levels of output there are falling average costs (thus IRS), at higher levels rising average costs (thus DRS).[7] Given the input prices, w_K and w_L, and a specified output level, $\bar{Y}$, 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:

$$\eta(\bar{K},\bar{L}) = \frac{LAC(\bar{Y})}{LMC(\bar{Y})}, \tag{2.13}$$

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.

Anyway, we have here a more subtle replication argument for CRS w.r.t. K and L 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

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

aggregate output will take place primarily by entry and exit of optimal-size 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 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. Notwithstanding the theoretical and empirical ambiguities, the assumption of CRS w.r.t. capital and labor has a prominent role in macroeconomics. In many contexts it is regarded as an acceptable approximation and a convenient simple background for studying the question at hand.

Expedient inferences of the CRS assumption include:

(i) marginal costs are constant and equal to average costs (so the right-hand side of (2.13) 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 (2.12) equals unity);

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

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

(v) a two-factor production function known to have CRS and to be 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.[8]

[8] Proofs of these claims can be found in intermediate microeconomics textbooks and in the Appendix to Chapter 2 of my Lecture Notes in Macroeconomics.

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:

$$F_K(K, L) = \frac{\partial Y}{\partial K} = \frac{\partial \left[Lf(k)\right]}{\partial K} = Lf'(k)\frac{\partial k}{\partial K} = f'(k). \tag{2.14}$$

And the marginal productivity of labor can be written

$$\begin{aligned} F_L(K, L) &= \frac{\partial Y}{\partial L} = \frac{\partial \left[Lf(k)\right]}{\partial L} = f(k) + Lf'(k)\frac{\partial k}{\partial L} \\ &= f(k) + Lf'(k)K(-L^{-2}) = f(k) - f'(k)k. \end{aligned} \tag{2.15}$$

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 (2.14) 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

$$f(k) - f'(k)k > 0 \text{ for all } k > 0, \tag{2.16}$$

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

$$\lim_{k \to 0} \left[f(k) - f'(k)k\right] = f(0). \tag{2.17}$$

Indeed, from the mean value theorem[9] we know there exists a number $a \in (0, 1)$ such that for any given $k > 0$ we have $f(k) - f(0) = f'(ak)k$. From this follows $f(k) - f'(ak)k = f(0) < f(k) - f'(k)k$, since $f'(ak) > f'(k)$ by $f'' < 0$.

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

In view of $f(0) \geq 0$, this establishes (2.16). And from $f(k) > f(k) - f'(k)k > f(0)$ and continuity of f follows (2.17).

Under CRS the Inada conditions for MPK can be written

$$\lim_{k\to 0} f'(k) = \infty, \qquad \lim_{k\to\infty} f'(k) = 0. \tag{2.18}$$

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 (2.18), 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.[10]

Figure 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 Figure 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 Figure 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 $k < B_1/A_1$ 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 Figure 2.3. Note that for $k \to 0$, MPK stays equal to $A_1 < \infty$, whereas for all $k > B_2/A_2$, $MPK = 0$. A similar feature remains true, when we consider *many,* say n, alternative efficient Leontief techniques available. Assuming these techniques cover a considerable range w.r.t. the B/A ratios,

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

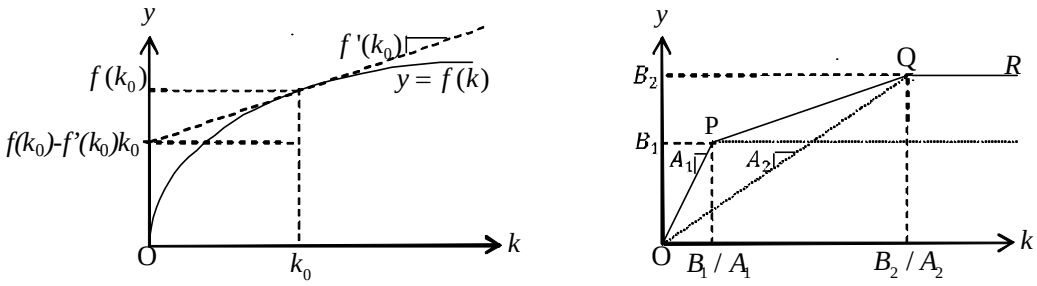


Figure 2.3: Two labor productivity curves based on CRS technologies. Left: neoclassical technology with Inada conditions for MPK satisfied; the graphical representation of MPK and MPL at $k = k_0$.as $f'(k_0)$ and $f(k_0) - f'(k_0)k_0$ are indicated. Right: a combination of two efficient Leontief techniques.

we get a labor productivity curve looking more like that of a neoclassical CRS production function. On the one hand, this gives some intuition of what lies behind the assumption of a neoclassical CRS production function. On the other hand, it remains true that for all $k > B_n/A_n$, $MPK = 0$,[11] 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 to 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*.

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

Concepts of neutral technological change

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

$$Y_t = F(a_t K_t, b_t L_t), \tag{2.19}$$

where F is a (time-independent) neoclassical production function, Y_t, K_t, and L_t are output, capital, and labor input, respectively, at time t, while a_t and b_t are time-dependent efficiencies of capital and labor, respectively, reflecting technological change. In macroeconomics an even more specific form is often assumed, namely the form of *Harrod-neutral technological change.*[12] This amounts to assuming that a_t in (2.19) is a constant (which we can then normalize to one). So only b_t, which we will then denote T_t, is changing over time, and we have

$$Y_t = F(K_t, T_t L_t). \tag{2.20}$$

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 (2.20) represents *Harrod-neutral technological progress.* An alternative name for this is *labor-augmenting* technological progress (technological change acts *as if* the labor input were augmented).

If the function F in (2.20) is homogeneous of degree one (so that the technology exhibits CRS w.r.t. capital and labor), we may write

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

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

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

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

[12] The name refers to the English economist Roy F. Harrod, 1900−1978.

where, again, F is a (time-independent) neoclassical production function, while T_t is the growing technology level.[13] 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 (2.21) as $Y_t = F(T_tK_t, T_tL_t)$. Comparing with (2.19), we see that in this case Hicks-neutrality is equivalent with $a_t = b_t$ in (2.19), whereby technological change is said to be *equally factor-augmenting.*

Finally, in a kind of symmetric analogy with (2.20), *Solow-neutral* technological progress[14] is often in textbooks presented by a formula like:

$$Y_t = F(T_tK_t, L_t). \tag{2.22}$$

Another name for the same is *capital-augmenting* technological progress (because here technological change acts as if the capital input were augmented). Solow's original concept[15] of neutral technological change is not well portrayed this way, however, since it is related to the notion of *embodied* technological change and capital of *different vintages,* see below.

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

Note that the neutrality concepts do not say anything about the *source* of technological progress, only about the quantitative form in which it materializes. For instance, the occurrence of Harrod-neutrality should not be interpreted as indicating that the technological change emanates specifically from the labor input in some sense. Harrod-neutrality only means that technological innovations predominantly are such that not only do labor and capital in combination become more productive, but this happens to *manifest itself* in the form (2.20). Similarly, if indeed an improvement in the quality of the labor input occurs, this "labor-specific" improvement may be manifested in a higher a_t, b_t, or both.

Before proceeding, we briefly comment on how the capital stock, K_t, is typically measured. While data on gross investment, I_t, is available in national income and product accounts, data on K_t usually is not. One ap-

[13] The name refers to the English economist and Nobel Prize laureate John R. Hicks, 1904−1989.

[14] The name refers to the American economist and Nobel Prize laureate Robert Solow (1924−).

[15] Solow (1960).

proach to the measurement of K_t is the *perpetual inventory method* which builds upon the accounting relationship

$$K_t = I_{t-1} + (1-\delta)K_{t-1}. \tag{2.23}$$

Assuming a constant capital depreciation rate δ, backward substitution gives

$$K_t = I_{t-1}+(1-\delta)\left[I_{t-2} + (1-\delta)K_{t-2}\right] = \ldots = \sum_{i=1}^{N}(1-\delta)^{i-1}I_{t-i}+(1-\delta)^N K_{t-N}. \tag{2.24}$$

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 (2.24) becomes very small.

Embodied vs. disembodied technological progress

There exists an additional taxonomy of technological change. 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 way formalize embodied technological progress by writing capital accumulation in the following way:

$$K_{t+1} - K_t = q_t I_t - \delta K_t, \tag{2.25}$$

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 (2.25) 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 (2.20) 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 (2.20), (2.21), and (2.22) above refers to the total, historically accumulated capital stock as calculated by (2.24), 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 (2.24) to better take estimated quality improvements into account. Otherwise the resulting K will not indicate the capital stock measured in efficiency units.

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

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

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

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

and such that the choices of a single firm facing this production function coincide with the aggregate outcomes, $\Sigma_{i=1}^n Y_i$, $\Sigma_{i=1}^n K_i$, and $\Sigma_{i=1}^n L_i$, in the original economy. Then $F^*(K, L)$ is called the *aggregate production function*

or the production function of the *representative* firm. It is *as if* aggregate production is the result of the behavior of such a single firm.

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

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

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

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

In this (trivial) case the aggregate production function is well-defined and turns out to be exactly the same as the identical CRS production functions of the individual firms.

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

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

Allowing also the existence of different output goods, different capital goods, and different types of labor makes the issue more intricate, of course. Yet, if firms are price taking profit maximizers and there are nonincreasing returns to scale, we at least know that the aggregate outcome is *as if,* for given prices, the firms jointly maximize aggregate profit on the basis of their combined production technology (Mas-Colell et al., 1955). The problem is, however, that the conditions needed for this to imply existence of an aggregate production function which is *well-behaved* (in the sense of inheriting simple qualitative properties from its constituent parts) are restrictive.

Nevertheless macroeconomics often treats aggregate output as a single homogeneous good and capital and labor as being two single and homogeneous inputs. There was in the 1960s a heated debate about the problems involved in this, with particular emphasis on the aggregation of different kinds of equipment into one variable, the capital stock "K". The debate is known as the "Cambridge controversy" because the dispute was between a group of

economists from Cambridge University, UK, and a group from Massachusetts Institute of Technology (MIT), which is located in Cambridge, USA. The former group questioned the theoretical robustness of several of the neoclassical tenets, including the proposition that rising aggregate capital intensity tends to be associated with a falling rate of interest. Starting at the disaggregate level, an association of this sort is not a logical necessity because, with different production functions across the industries, the relative prices of produced inputs tend to change, when the interest rate changes. While acknowledging the possibility of "paradoxical" relationships, the latter group maintained that in a macroeconomic context they are likely to cause devastating problems only under exceptional circumstances. In the end this is a matter of empirical assessment.[16]

To avoid complexity and because, for many important issues in growth theory, there is today no well-tried alternative, we shall in this course most of the time use aggregate constructs like "Y", "K", and "L" as simplifying devices, hopefully acceptable in a first approximation. There are cases, however, where some disaggregation is pertinent. When for example the role of imperfect competition is in focus, we shall be ready to disaggregate the production side of the economy into several product lines, each producing its own differentiated product. We shall also touch upon a type of growth models where a key ingredient is the phenomenon of "creative destruction" meaning that an incumbent technological leader is competed out by an entrant with a qualitatively new technology.

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 importance 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* marginal propensity to consume out of income, aggregation is straightforward and the representative household is a meaningful concept. On the other hand, if we aim at understanding, say, the *interaction* between lending and borrowing households, perhaps via financial intermediaries, the representative household is not a useful starting point. Similarly, if the theme is conflicts of interests between firm owners and employees, the existence of *different* types of households should be taken into account.

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

2.4 Long-run vs. short-run production functions*

Is the substitutability between capital and labor the same "ex ante" and "ex post"? 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, of for instance the Solow or the Ramsey model, 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,

$$Y = \min(Au\bar{K}, BL), \qquad A > 0, B > 0, \tag{2.26}$$

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.

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 $u = Y^d/(A\bar{K}) < 1$ and $L = Y^d/B$. 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 (2.26). These pairs satisfy

$$f(k) = Ak = B. \tag{2.27}$$

So, for an increasing sequence of k's, $k_1, k_2, \ldots, k_i, \ldots$, the corresponding

pairs are $(A_i, B_i) = (f(k_i)/k_i, f(k_i))$, $i = 1, 2, \ldots$.[17] 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 (2.26). 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.

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 2.1 gives an overview of the alternative cases.

Table 2.1. 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

The putty-clay case is generally considered the realistic case. As time proceeds, technological progress occurs. To take this into account, we may replace (2.27) and (2.26) 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

[17] 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.

least one strict equality), it may pay the firm to invest in the new technology at the same time as some old machinery is scrapped. Real wages tend to rise along with technological progress and the scrapping occurs because the revenue from using the old machinery in production no longer covers the associated labor costs.

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

2.5 Literature notes

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

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

Informative history of technology is contained in Ruttan (2001) and Smil (2003). For more general economic history, see, e.g., Clark (2008) and Persson (2010). Forecasts of technological development in the next decades are contained in, for instance, Brynjolfsson and McAfee (2014).

Embodied technological progress, sometimes called investment-specific technological progress, is explored in, for instance, Solow (1960), Greenwood

et al. (1997), and Groth and Wendner (2014). Hulten (2001) surveys the literature and issues related to measurement of the direct contribution of capital accumulation and technological change, respectively, to productivity growth.

Conditions ensuring that a representative household is admitted and the concept of Gorman preferences are discussed in Acemoglu (2009). Another useful source, also concerning the conditions for the representative firm to be a meaningful notion, is Mas-Colell et al. (1995). For general discussions of the limitations of representative agent approaches, see Kirman (1992) and Gallegati and Kirman (1999). Reviews of the "Cambridge Controversy" are contained in Mas-Colell (1989) and Felipe and Fisher (2003). The last-mentioned authors find the conditions required for the well-behavedness of these constructs so stringent that it is difficult to believe that actual economies are in any sense close to satisfy them. For a less distrustful view, see for instance Ferguson (1969), Johansen (1972), Malinvaud (1998), Jorgenson et al. (2005), and Jones (2005).

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

2.6 References

(incomplete)

Brynjolfsson, E., and A. McAfee, 2014, *The Second Machine Age*, Norton.

Clark, G., 2008, *A Farewell to Alms: A Brief Economic History of the World,* Princeton University Press.

Persson, K. G., 2010, *An economic history of Europe. Knowledge, institutions and growth, 600 to the present*, Cambridge University Press: Cambridge.

Ruttan, V. W. , 2001, *Technology, Growth, and Development: An Induced Innovation Perspective*, Oxford University Press: Oxford.

Smil, V., 2003, *Energy at the crossroads. Global perspectives and uncertainties*, MIT Press: Cambridge Mass.

Chapter 3

Continuous time analysis

Because dynamic analysis is generally easier in continuous time, growth models are often stated in continuous time. This chapter gives an account of the conceptual aspects of continuous time analysis. Appendix A considers simple growth arithmetic in continuous time. And Appendix B provides solution formulas for linear first-order differential equations.

3.1 The transition from discrete time to continuous time

We start from a discrete time framework. The run of time is divided into successive periods of equal length, taken as the time-unit. Let us here index the periods by $i = 0, 1, 2, ...$ Thus financial wealth accumulates according to

$$a_{i+1} - a_i = s_i, \qquad a_0 \text{ given},$$

where s_i is (net) saving in period i.

3.1.1 Multiple compounding per year

With time flowing continuously, we let $a(t)$ refer to financial wealth at time t. Similarly, $a(t+\Delta t)$ refers to financial wealth at time $t+\Delta t$. To begin with, let Δt equal one time unit. Then $a(i\Delta t)$ equals $a(i)$ and is of the same value as a_i. Consider the *forward* first difference in a, $\Delta a(t) \equiv a(t+\Delta t) - a(t)$. It makes sense to consider this change in a in relation to the length of the time interval involved, that is, to consider the *ratio* $\Delta a(t)/\Delta t$. As long as $\Delta t = 1$, with $t = i\Delta t$ we have $\Delta a(t)/\Delta t = (a_{i+1} - a_i)/1 = a_{i+1} - a_i$. Now, keep the time unit unchanged, but let the length of the time interval $[t, t+\Delta t)$

approach zero, i.e., let $\Delta t \to 0$. When a is a differentiable function of t, we have

$$\lim_{\Delta t \to 0} \frac{\Delta a(t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{a(t+\Delta t) - a(t)}{\Delta t} = \frac{da(t)}{dt},$$

where $da(t)/dt$, often written $\dot{a}(t)$, is known as the *derivative of* $a(\cdot)$ at the point t. Wealth accumulation in continuous time can then be written

$$\dot{a}(t) = s(t), \qquad a(0) = a_0 \text{ given}, \tag{3.1}$$

where $s(t)$ is the saving flow at time t. For Δt "small" we have the approximation $\Delta a(t) \approx \dot{a}(t)\Delta t = s(t)\Delta t$. In particular, for $\Delta t = 1$ we have $\Delta a(t) = a(t+1) - a(t) \approx s(t)$.

As time unit choose one year. Going back to discrete time, if wealth grows at a constant rate $g > 0$ per year, then after i periods of length one year, with annual compounding, we have

$$a_i = a_0(1+g)^i, \qquad i = 0, 1, 2, \ldots . \tag{3.2}$$

If instead compounding (adding saving to the principal) occurs n times a year, then after i periods of length $1/n$ year and a growth rate of g/n per such period,

$$a_i = a_0(1 + \frac{g}{n})^i. \tag{3.3}$$

With t still denoting time measured in years passed since date 0, we have $i = nt$ periods. Substituting into (3.3) gives

$$a(t) = a_{nt} = a_0(1 + \frac{g}{n})^{nt} = a_0 \left[(1 + \frac{1}{m})^m\right]^{gt}, \qquad \text{where } m \equiv \frac{n}{g}.$$

We keep g and t fixed, but let $n \to \infty$. Thus $m \to \infty$. Then, in the limit there is continuous compounding and it can be shown that

$$a(t) = a_0 e^{gt}, \tag{3.4}$$

where e is a mathematical constant called the base of the natural logarithm and defined as $e \equiv \lim_{m\to\infty}(1 + 1/m)^m \simeq 2.7182818285....$

The formula (3.4) is the continuous-time analogue to the discrete time formula (3.2) with annual compounding. A geometric growth factor is replaced by an exponential growth factor, e^{gt}, and this growth factor is valid for any t in the time interval $(-\tau_1, \tau_2)$ for which the growth rate of a equals the constant g (τ_1 and τ_2 being some positive real numbers).

We can also view the formulas (3.2) and (3.4) as the solutions to a difference equation and a differential equation, respectively. Thus, (3.2) is the solution to the linear difference equation $a_{i+1} = (1+g)a_i$, given the initial value

a_0. And (3.4) is the solution to the linear differential equation $\dot{a}(t) = ga(t)$, given the initial condition $a(0) = a_0$. Now consider a time-dependent growth rate, $g(t)$. The corresponding differential equation is $\dot{a}(t) = g(t)a(t)$ and it has the solution

$$a(t) = a(0)e^{\int_0^t g(\tau)d\tau}, \tag{3.5}$$

where the exponent, $\int_0^t g(\tau)d\tau$, is the definite integral of the function $g(\tau)$ from 0 to t. The result (3.5) is called the *basic accumulation formula* in continuous time and the factor $e^{\int_0^t g(\tau)d\tau}$ is called the *growth factor* or the *accumulation factor*.

3.1.2 Compound interest and discounting

Let $r(t)$ denote the *short-term real interest rate in continuous time* at time t. To clarify what is meant by this, consider a deposit of $V(t)$ euro on a drawing account in a bank at time t. If the general price level in the economy at time t is $P(t)$ euro, the *real* value of the deposit is $a(t) = V(t)/P(t)$ at time t. By definition the *real rate of return* on the deposit in continuous time (with continuous compounding) at time t is the (proportionate) instantaneous rate at which the real value of the deposit expands per time unit when there is no withdrawal from the account. Thus, if the instantaneous nominal interest rate is $i(t)$, we have $\dot{V}(t)/V(t) = i(t)$ and so, by the fraction rule in continuous time (cf. Appendix A),

$$r(t) = \frac{\dot{a}(t)}{a(t)} = \frac{\dot{V}(t)}{V(t)} - \frac{\dot{P}(t)}{P(t)} = i(t) - \pi(t), \tag{3.6}$$

where $\pi(t) \equiv \dot{P}(t)/P(t)$ is the instantaneous inflation rate. In contrast to the corresponding formula in discrete time, this formula is exact. Sometimes $i(t)$ and $r(t)$ are referred to as the nominal and real *interest intensity*, respectively, or the nominal and real *force of interest*.

Calculating the terminal value of the deposit at time $t_1 > t_0$, given its value at time t_0 and assuming no withdrawal in the time interval $[t_0, t_1]$, the accumulation formula (3.5) immediately yields

$$a(t_1) = a(t_0)e^{\int_{t_0}^{t_1} r(t)dt}.$$

When calculating *present values* in continuous time analysis, we use compound discounting. We reverse the accumulation formula and go from the compounded or terminal value to the present value $a(t_0)$. Similarly, given a consumption plan, $(c(t))_{t=t_0}^{t_1}$, the present value of this plan as seen from time

t_0 is

$$PV = \int_{t_0}^{t_1} c(t)\ e^{-rt} dt, \tag{3.7}$$

presupposing a constant interest rate. Instead of the geometric discount factor, $1/(1+r)^t$, from discrete time analysis, we have here an exponential discount factor, $1/(e^{rt}) = e^{-rt}$, and instead of a sum, an integral. When the interest rate varies over time, (3.7) is replaced by

$$PV = \int_{t_0}^{t_1} c(t)\ e^{-\int_{t_0}^{t} r(\tau)d\tau} dt.$$

In (3.7) $c(t)$ is discounted by $e^{-rt} \approx (1+r)^{-t}$ for r "small". This might not seem analogue to the discrete-time discounting in (**??**) where it is c_{t-1} that is discounted by $(1+r)^{-t}$, assuming a constant interest rate. When taking into account the timing convention that payment for c_{t-1} in period $t-1$ occurs at the end of the period (= time t), there is no discrepancy, however, since the continuous-time analogue to this payment is $c(t)$.

3.2 The allowed range for parameter values

The allowed range for parameters may change when we go from discrete time to continuous time with continuous compounding. For example, the usual equation for aggregate capital accumulation in continuous time is

$$\dot{K}(t) = I(t) - \delta K(t), \qquad K(0) = K_0 \text{ given}, \tag{3.8}$$

where $K(t)$ is the capital stock, $I(t)$ is the gross investment at time t and $\delta \geq 0$ is the (physical) capital depreciation rate. Unlike in discrete time, here $\delta > 1$ is conceptually allowed. Indeed, suppose for simplicity that $I(t) = 0$ for all $t \geq 0$; then (3.8) gives $K(t) = K_0 e^{-\delta t}$. This formula is meaningful for any $\delta \geq 0$. Usually, the time unit used in continuous time macro models is one year (or, in business cycle theory, rather a quarter of a year) and then a realistic value of δ is of course < 1 (say, between 0.05 and 0.10). However, if the time unit applied to the model is large (think of a Diamond-style OLG model), say 30 years, then $\delta > 1$ may fit better, empirically, if the model is converted into continuous time with the same time unit. Suppose, for example, that physical capital has a half-life of 10 years. With 30 years as our time unit, inserting into the formula $1/2 = e^{-\delta/3}$ gives $\delta = (\ln 2) \cdot 3 \simeq 2$.

In many simple macromodels, where the level of aggregation is high, the relative price of a unit of physical capital in terms of the consumption good is 1 and thus constant. More generally, if we let the relative price of the

capital good in terms of the consumption good at time t be $p(t)$ and allow $\dot{p}(t) \neq 0$, then we have to distinguish between the physical depreciation of capital, δ, and the *economic depreciation*, that is, the loss in economic value of a machine per time unit. The economic depreciation will be $d(t) = p(t)\delta - \dot{p}(t)$, namely the economic value of the physical wear and tear (and technological obsolescence, say) minus the capital gain (positive or negative) on the machine.

Other variables and parameters that by definition are bounded from below in discrete time analysis, but not so in continuous time analysis, include rates of return and discount rates in general.

3.3 Stocks and flows

An advantage of continuous time analysis is that it forces the analyst to make a clear distinction between *stocks* (say wealth) and *flows* (say consumption or saving). Recall, a *stock* variable is a variable measured as a quantity at a given point in time. The variables $a(t)$ and $K(t)$ considered above are stock variables. A *flow* variable is a variable measured as quantity *per time unit* at a given point in time. The variables $s(t)$, $\dot{K}(t)$ and $I(t)$ are flow variables.

One can not add a stock and a flow, because they have *different denominations*. What is meant by this? The elementary measurement units in economics are *quantity units* (so many machines of a certain kind or so many liters of oil or so many units of payment, for instance) and *time units* (months, quarters, years). On the basis of these elementary units we can form *composite measurement units*. Thus, the capital stock, K, has the denomination "quantity of machines", whereas investment, I, has the denomination "quantity of machines per time unit" or, shorter, "quantity/time". A growth rate or interest rate has the denomination "(quantity/time)/quantity" = "time^{-1}". If we change our time unit, say from quarters to years, the value of a flow variable as well as a growth rate is changed, in this case quadrupled (presupposing annual compounding).

In continuous time analysis expressions like $K(t)+I(t)$ or $K(t)+\dot{K}(t)$ are thus illegitimate. But one can write $K(t+\Delta t) \approx K(t)+(I(t)-\delta K(t))\Delta t$, or $\dot{K}(t)\Delta t \approx (I(t) - \delta K(t))\Delta t$. In the same way, suppose a bath tub at time t contains 50 liters of water and that the tap pours $\frac{1}{2}$ liter per second into the tub for some time. Then a sum like $50\ \ell + \frac{1}{2}\ (\ell/\text{sec})$ does not make sense. But the *amount* of water in the tub after one minute is meaningful. This amount would be $50\ \ell + \frac{1}{2} \cdot 60\ ((\ell/\text{sec})\times\text{sec}) = 80\ \ell$. In analogy, economic flow variables in continuous time should be seen as *intensities* defined for every t in the time interval considered, say the time interval $[0,\ T)$ or perhaps

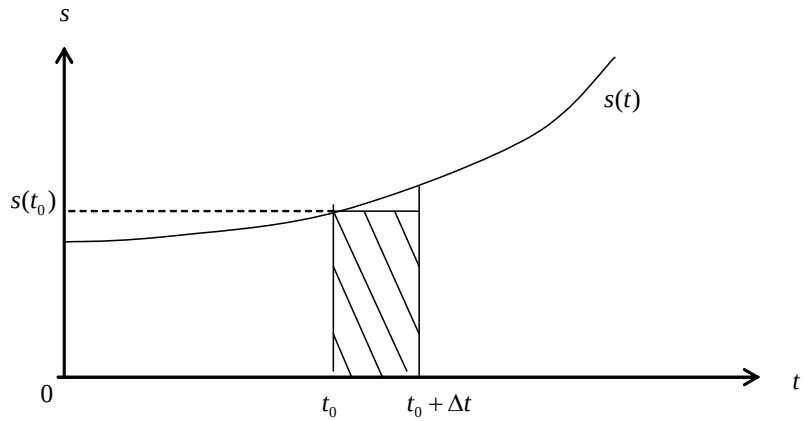


Figure 3.1: With Δt "small" the integral of $s(t)$ from t_0 to $t_0+\Delta t$ is $\approx$ the hatched area.

$[0,\ \infty)$. For example, when we say that $I(t)$ is "investment" at time t, this is really a short-hand for "investment intensity" at time t. The actual investment in a time interval $[t_0, t_0 + \Delta t)$, i.e., the invested amount *during* this time interval, is the integral, $\int_{t_0}^{t_0+\Delta t} I(t)dt \approx I(t_0)\Delta t$. Similarly, the flow of individual saving, $s(t)$, should be interpreted as the saving *intensity* at time t. The actual saving in a time interval $[t_0, t_0 + \Delta t)$, i.e., the saved (or accumulated) amount during this time interval, is the integral, $\int_{t_0}^{t_0+\Delta t} s(t)dt$. If Δt is "small", this integral is approximately equal to the product $s(t_0)\cdot\Delta t$, cf. the hatched area in Figure 3.1.

The notation commonly used in discrete time analysis blurs the distinction between stocks and flows. Expressions like $a_{i+1} = a_i + s_i$, without further comment, are usual. Seemingly, here a stock, wealth, and a flow, saving, are added. In fact, however, it is wealth at the beginning of period i and the saved *amount during* period i that are added: $a_{i+1} = a_i + s_i \cdot \Delta t$. The tacit condition is that the period length, Δt, is the time unit, so that $\Delta t = 1$. But suppose that, for example in a business cycle model, the period length is one quarter, but the time unit is one year. Then saving in quarter i is $s_i = (a_{i+1} - a_i) \cdot 4$ per year.

3.4 The choice between discrete and continuous time formulation

In empirical economics, data typically come in discrete time form and data for flow variables typically refer to periods of constant length. One could argue that this discrete form of the data speaks for discrete time rather than continuous time modelling. And the fact that economic actors often think and plan in period terms, may seem a good reason for putting at least microeconomic analysis in period terms. Nonetheless real time is continuous. And, as for instance Allen (1967) argued, it can hardly be said that the *mass* of economic actors think and plan with one and the same period. In macroeconomics we consider the *sum* of the actions. In this perspective the continuous time approach has the advantage of allowing variation *within* the usually artificial periods in which the data are chopped up. And centralized asset markets equilibrate very fast and respond immediately to new information. For such markets a formulation in continuous time seems a better approximation.

There is also a risk that a discrete time model may generate *artificial* oscillations over time. Suppose the "true" model of some mechanism is given by the differential equation

$$\dot{x} = \alpha x, \qquad \alpha < -1. \tag{3.9}$$

The solution is $x(t) = x(0)e^{\alpha t}$ which converges in a monotonic way toward 0 for $t \to \infty$. However, the analyst takes a discrete time approach and sets up the seemingly "corresponding" discrete time model

$$x_{t+1} - x_t = \alpha x_t.$$

This yields the difference equation $x_{t+1} = (1+\alpha)x_t$, where $1+\alpha < 0$. The solution is $x_t = (1+\alpha)^t x_0$, $t = 0, 1, 2, \ldots$. As $(1+\alpha)^t$ is positive when t is even and negative when t is odd, oscillations arise (together with divergence if $\alpha < -2$) in spite of the "true" model generating monotonous convergence towards the steady state $x^* = 0$.

It should be added, however, that this potential problem *can* always be avoided within discrete time models by choosing a sufficiently *short* period length. Indeed, the solution to a differential equation can always be obtained as the limit of the solution to a corresponding difference equation for the period length approaching zero. In the case of (3.9) the approximating difference equation is $x_{i+1} = (1+\alpha\Delta t)x_i$, where Δt is the period length, $i = t/\Delta t$, and $x_i = x(i\Delta t)$. By choosing Δt small enough, the solution comes

arbitrarily close to the solution of (3.9). It is generally more difficult to go in the opposite direction and find a differential equation that approximates a given difference equation. But the problem is solved as soon as a differential equation has been found that has the initial difference equation as an approximating difference equation.

From the point of view of the economic contents, the choice between discrete time and continuous time may be a matter of taste. Yet, everything else equal, the clearer distinction between stocks and flows in continuous time than in discrete time speaks for the former. From the point of view of mathematical convenience, the continuous time formulation, which has worked so well in the natural sciences, is preferable. At least this is so in the absence of uncertainty. For problems where uncertainty is important, discrete time formulations are easier to work with unless one is familiar with stochastic calculus.

3.5 Appendix A: Growth arithmetic in continuous time

Let the variables z, x, and y be differentiable functions of time t. Suppose $z(t)$, $x(t)$, and $y(t)$ are positive for all t. Then:

PRODUCT RULE $z(t) = x(t)y(t) \Rightarrow \frac{\dot{z}(t)}{z(t)} = \frac{\dot{x}(t)}{x(t)} + \frac{\dot{y}(t)}{y(t)}$.

Proof. Taking logs on both sides of the equation $z(t) = x(t)y(t)$ gives $\ln z(t) = \ln x(t) + \ln y(t)$. Differentiation w.r.t. t, using the chain rule, gives the conclusion. □

The procedure applied in this proof is called *logarithmic differentiation* w.r.t. t.

FRACTION RULE $z(t) = \frac{x(t)}{y(t)} \Rightarrow \frac{\dot{z}(t)}{z(t)} = \frac{\dot{x}(t)}{x(t)} - \frac{\dot{y}(t)}{y(t)}$.

The proof is similar.

POWER FUNCTION RULE $z(t) = x(t)^\alpha \Rightarrow \frac{\dot{z}(t)}{z(t)} = \alpha \frac{\dot{x}(t)}{x(t)}$.

The proof is similar.

In continuous time these simple formulas are exactly true. In discrete time the analogue formulas are only approximately true and the approximation can be quite bad unless the growth rates of x and y are small.

3.6 Appendix B: Solution formulas for linear differential equations of first order

For a general differential equation of first order, $\dot{x}(t) = \varphi(x(t), t)$, with $x(t_0) = x_{t_0}$ and where φ is a continuous function, we have, at least for t in an interval $(-\varepsilon, +\varepsilon)$ for some $\varepsilon > 0$,

$$x(t) = x_{t_0} + \int_{t_0}^{t} \varphi(x(\tau), \tau) d\tau. \qquad (*)$$

To get a confirmation, calculate $\dot{x}(t)$ from (*).

For the special case of a linear differential equation of first order, $\dot{x}(t) + a(t)x(t) = b(t)$, we can specify the solution. Three sub-cases of rising complexity are:

1. $\dot{x}(t) + ax(t) = b$, with $a \neq 0$ and initial condition $x(t_0) = x_{t_0}$. Solution:

$$x(t) = (x_{t_0} - x^*)e^{-a(t-t_0)} + x^*, \text{ where } x^* = \frac{b}{a}.$$

If $a = 0$, we get, directly from (*), the solution $x(t) = x_{t_0} + bt$.[1]

2. $\dot{x}(t) + ax(t) = b(t)$, with initial condition $x(t_0) = x_{t_0}$. Solution:

$$x(t) = x_{t_0}e^{-a(t-t_0)} + e^{-a(t-t_0)} \int_{t_0}^{t} b(s)e^{a(s-t_0)} ds.$$

Special case: $b(t) = ce^{ht}$, with $h \neq -a$ and initial condition $x(t_0) = x_{t_0}$. Solution:

$$x(t) = x_{t_0}e^{-a(t-t_0)} + e^{-a(t-t_0)}c \int_{t_0}^{t} e^{(a+h)(s-t_0)} ds = (x_{t_0} - \frac{c}{a+h})e^{-a(t-t_0)} + \frac{c}{a+h}e^{h(t-t_0)}.$$

3. $\dot{x}(t) + a(t)x(t) = b(t)$, with initial condition $x(t_0) = x_{t_0}$. Solution:

$$x(t) = x_{t_0}e^{-\int_{t_0}^{t} a(\tau)d\tau} + e^{-\int_{t_0}^{t} a(\tau)d\tau} \int_{t_0}^{t} b(s)e^{\int_{t_0}^{s} a(\tau)d\tau} ds.$$

[1] Some non-linear differential equations can be transformed into this simple case. For simplicity let $t_0 = 0$. Consider the equation $\dot{y}(t) = \alpha y(t)^\beta$, $y_0 > 0$ given, $\alpha \neq 0, \beta \neq 1$ (a *Bernoulli equation*). To find the solution for $y(t)$, let $x(t) \equiv y(t)^{1-\beta}$. Then, $\dot{x}(t) = (1-\beta)y(t)^{-\beta}\dot{y}(t) = (1-\beta)y(t)^{-\beta}\alpha y(t)^\beta = (1-\beta)\alpha$. The solution for this is $x(t) = x_0 + (1-\beta)\alpha t$, where $x_0 = y_0^{1-\beta}$. Thereby the solution for $y(t)$ is $y(t) = x(t)^{1/(1-\beta)} = \left(y_0^{1-\beta} + (1-\beta)\alpha t\right)^{1/(1-\beta)}$, which is defined for $t > -y_0^{1-\beta}/((1-\beta)\alpha$.

Special case: $b(t) = 0$. Solution:

$$x(t) = x_{t_0} e^{-\int_{t_0}^{t} a(\tau) d\tau}.$$

Even more special case: $b(t) = 0$ and $a(t) = a$, a constant. Solution:

$$x(t) = x_{t_0} e^{-a(t-t_0)}.$$

Remark 1 For $t_0 = 0$, most of the formulas will look simpler.

Remark 2 To check whether a suggested solution *is* a solution, calculate the time derivative of the suggested solution and add an arbitrary constant. By appropriate adjustment of the constant, the final result should be a replication of the original differential equation together with its initial condition.