Economic Growth Lecture Note 2

08.02.2012. Christian Groth

Terminology concerning technology. Continuous time modeling

In the lecture yesterday I presented the terminology around technology which I find useful and will use in the lectures. Section 1 of this lecture note is a recapitulation of this and of how it relates to some of the definitions in Acemoglu's book. Section 2 provides a refresher of stuff that should be well-known from earlier courses in micro- and macroeconomics. Section 3 can be used as a formula manual for the case of CRS. Finally, in Section 4 we describe the transition from discrete time to continuous time analysis.

1 A two-factor production function

Consider a two-factor production function given by

$$Y = F(K, L), \tag{1}$$

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 (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 (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 (cf. Exercise Problem I.3).

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 continuity (and the "no input, no output" property).

1.1 A neoclassical production function

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:

$$F_K > 0, \quad F_L > 0, \tag{2}$$

$$F_{KK} < 0, \quad F_{LL} < 0. \tag{3}$$

(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 form qualitatively similar to that shown in Fig. 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 $\Delta Y \approx (\partial Y/\partial K)\Delta K = \partial Y/\partial K$.

1.2 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

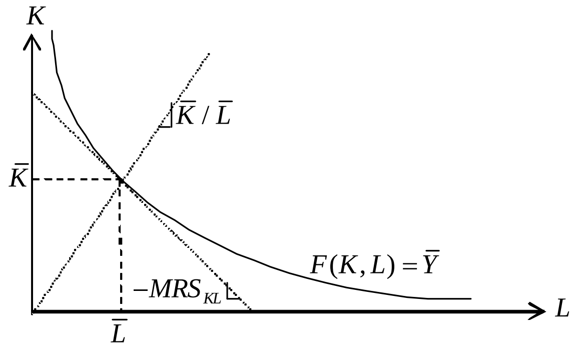


Figure 1: MRS_{KL} as the absolute slope of the isoquant.

L at the point (K, L) is defined as the absolute slope of the isoquant at that point, cf. Fig. 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

$$MRS_{KL} = -\frac{dK}{dL}_{|Y=\bar{Y}} = \frac{F_L(K, L)}{F_K(K, L)} > 0. \tag{4}$$

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.[1] 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).

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 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 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.

[1] The subscript $|Y = \bar{Y}$ in (4) indicates that we are moving along a given isoquant, $F(K, L) = \bar{Y}$.

1.3 The Inada conditions

A continuously differentiable production function is said to satisfy the *Inada conditions*[2] if

$$\lim_{K\to 0} F_K(K,L) = \infty, \lim_{K\to\infty} F_K(K,L) = 0, \tag{5}$$

$$\lim_{L\to 0} F_L(K,L) = \infty, \lim_{L\to\infty} F_L(K,L) = 0. \tag{6}$$

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 vanishing when the input of the factor increases without bound. Actually, (5) and (6) express *four* conditions, which it is preferable to consider separately and label one by one. In (5) we have two *Inada conditions for MPK* (the marginal productivity of capital), the first being a *lower*, the second an *upper* Inada condition for MPK. And in (6) 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. Some economic growth textbooks, *including Acemoglu's*, define a neoclassical production function more narrowly by including the Inada conditions as a requirement for calling the production function neoclassical. In contrast, when in a given context we need one or another Inada condition, we state it explicitly as an additional assumption.

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 $\lambda > 0$,

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

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

As all inputs are scaled up or down by some factor, output is scaled up or down by the same factor. In their definition of a neoclassical production function some textbooks, *including Acemoglu's*, add constant returns to scale as a requirement. In contrast, when in a given context we need an assumption of constant returns to scale, this is stated as an additional assumption.

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, 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.

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

$$Y = AK^{\alpha}L^{\beta}, \qquad A > 0, 0 < \alpha < 1, 0 < \beta < 1, \tag{7}$$

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 the "total factor productivity". As an exercise, the reader should verify that (7) satisfies (a) and (b) above and is therefore a neoclassical production function. The function in (7) 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. $\square$

EXAMPLE 2 The production function

$$Y = \min(AK, BL), \qquad A > 0, B > 0, \tag{8}$$

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*. 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. As (8) 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. $\square$

2.1 The replication argument

The assumption of CRS is widely used in macroeconomics. The model builder may appeal to the *replication argument* saying that by, conceptually, doubling all the inputs, we should always be able to double the output, since we just "replicate" what we are already doing. One should be aware that in principle the CRS assumption is about *technology* − limits to the availability of resources is another question. The CRS assumption and the replication argument presuppose that *all* the relevant inputs are explicit as arguments in the production function and that these are changed equiproportionately. Concerning our present production function $F(\cdot)$, one could easily argue that besides capital and labor, also land is a necessary input and should appear as a separate argument.[3] Then,

[3] We think of "capital" as producible means of production, whereas "land" refers to non-producible natural resources, including for example building sites. If an industrial firm decides to duplicate what it

on the basis of the replication argument we should in fact expect DRS wrt. capital and labor alone. In manufacturing and services, empirically, this and other possible sources for departure from CRS may be minor and so many macroeconomists feel comfortable enough with assuming CRS wrt. 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 and an important part of national wealth.

Another problem with the replication argument is the following. 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 precise form. This gives occasion for introducing a useful local measure of returns to scale.

2.2 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

$$\eta(K, L) = \frac{\theta}{F(K, L)} \frac{dF(\theta K, \theta L)}{d\theta} \approx \frac{\Delta F(\theta K, \theta L)/F(K, L)}{\Delta\theta/\theta}, \text{ evaluated at } \theta = 1. \tag{9}$$

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

$$\text{if } \eta(K, L) \begin{cases} > 1, \text{ then there are locally } \textit{IRS,} \\ = 1, \text{ then there are locally } \textit{CRS,} \\ < 1, \text{ then there are locally } \textit{DRS.} \end{cases} \tag{10}$$

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

has been doing, it needs a piece of land to build another plant like the first.

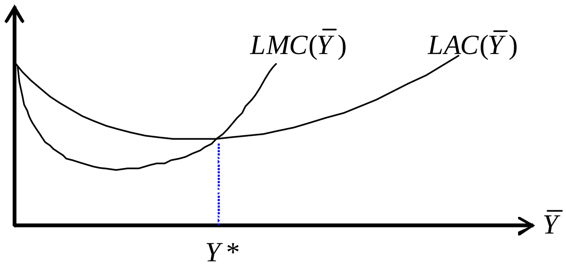


Figure 2: Locally CRS at optimal plant size.

partial output elasticities at that point:

$$\eta(K,L) = \frac{F_K(K,L)K}{F(K,L)} + \frac{F_L(K,L)L}{F(K,L)}. \tag{11}$$

This follows from the definition in (9) by taking into account that

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

Fig. 2 illustrates a popular case from microeconomics, a U-shaped average cost curve from the perspective of the individual firm (or plant): at low levels of output there are falling average costs (thus IRS), at higher levels rising average costs (thus DRS). 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 $F_L(\bar{K}, \bar{L})/F_K(\bar{K}, \bar{L}) = w_L/w_K$. From microeconomics we know that the elasticity of scale at $(\bar{K}, \bar{L})$ will satisfy:

$$\eta(\bar{K}, \bar{L}) = \frac{LAC(\bar{Y})}{LMC(\bar{Y})}, \tag{12}$$

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.

This provides a more subtle replication argument for CRS at the aggregate level. Even though technologies may differ across firms, the surviving firms in a competitive market will have the same average costs at the optimal plant size. In the medium and long run, changes in aggregate output will take place primarily by entry and exit of optimal-size 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.

The empirical evidence concerning returns to scale is mixed (see the literature notes at the end of the chapter). Notwithstanding the theoretical and empirical ambiguities, the assumption of CRS wrt. 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.

3 Properties of the production function under CRS

Expedient inferences of the CRS assumption include:

(i) marginal costs are constant and equal to average costs (so the right-hand side of (12) 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 (11) 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 "gross-complementarity" between K and L;

(v) a two-factor production function known to have CRS and 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.[4]

[4] Proofs of these claims are 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 the *capital intensity* and $f(k)$ is the *production function in intensive form* (sometimes named the per capita production function). Thus output per unit of labor depends only on the capital intensity:

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

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

$$F_K(K, L) = \frac{\partial Y}{\partial K} = \frac{\partial [Lf(k)]}{\partial K} = Lf'(k)\frac{\partial k}{\partial K} = f'(k). \tag{13}$$

And the marginal productivity of labor can be written

$$\begin{aligned} F_L(K, L) &= \frac{\partial Y}{\partial L} = \frac{\partial [Lf(k)]}{\partial L} = f(k) + Lf'(k)\frac{\partial k}{\partial L} \\ &= f(k) + Lf'(k)K(-L^{-2}) = f(k) - f'(k)k. \end{aligned} \tag{14}$$

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 (13) and (2). And the property $f'' < 0$ follows from (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

$$f(k) - f'(k)k > 0 \text{ for all } k > 0, \tag{15}$$

as well as

$$\lim_{k \to 0} [f(k) - f'(k)k] = f(0). \tag{16}$$

Indeed, from the mean value theorem[5] 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 $f'' < 0$. In view of $f(0) \geq 0$, this establishes (15). And from $f(k) > f(k) - f'(k)k > f(0)$ and continuity of f follows (16).

Under CRS the Inada conditions for MPK can be written

$$\lim_{k \to 0} f'(k) = \infty, \qquad \lim_{k \to \infty} f'(k) = 0. \tag{17}$$

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

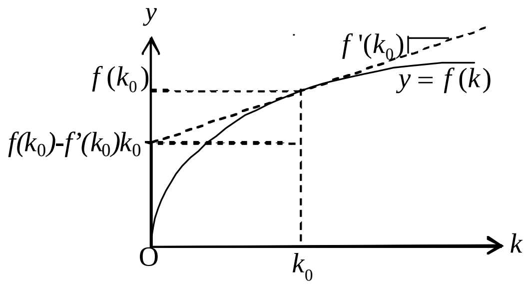
An input which must be positive for positive output to arise is called an *essential input.* The second part of (17), representing the upper Inada condition for MPK under CRS, has the implication that *labor* is an essential input; but capital need not be, as the production function $f(k) = a + bk/(1 + k)$, $a > 0, b > 0$, illustrates. Similarly, under CRS the upper Inada condition for MPL implies that *capital* is an essential input.[6] Combining these results, when *both* the upper Inada conditions hold and CRS obtains, then both capital and labor are essential inputs.[7]

Fig. 3 is drawn to provide an intuitive understanding of a neoclassical CRS production function and at the same time illustrate that the lower Inada conditions are more questionable than the upper Inada conditions. The left panel of Fig. 3 shows output per unit of labor for a CRS neoclassical production function satisfying the Inada conditions for MPK. The $f(k)$ in the diagram could for instance represent the Cobb-Douglas function in Example 1 with $\beta = 1 - \alpha$, i.e., $f(k) = Ak^{\alpha}$. The right panel of Fig. 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 Fig. 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 wrt. the B/A ratios, we get a labor productivity curve looking more like that of a neoclassical CRS production function. On the one hand, this gives some intuition of what lies behind the assumption of a neoclassical CRS production function. On the other hand, it remains true that for all $k > B_n/A_n$, $MPK = 0$,[8] whereas for $k \to 0$, MPK stays equal to $A_1 < \infty$, thus questioning the lower Inada condition.

[6] Proofs of these claims are in the Appendix to Chapter 2 of my Lecture Notes in Macroeconomics.

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

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



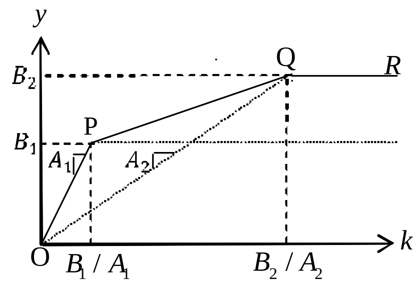


Figure 3: Two labor productivity curves based on CRS technologies. Left: neoclassical technology with Inada conditions for MPK satisfied. Right: a combination of two efficient Leontief techniques.

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* can be shown to imply, under CRS, that output has no upper bound when either input goes to infinity for fixed amount of the other input.

4 On continuous time analysis

Because dynamic analysis is often easier in continuous time, most growth models are stated in continuous time. In this section we describe some of the conceptual aspects of continuous time analysis.

Let us start from a discrete time framework: the run of time is divided into successive periods of constant length, taken as the time-unit. Let financial wealth at the beginning of period i be denoted a_i, $i = 0, 1, 2, ...$ Then wealth accumulation in discrete time can be written

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

where s_i is (net) saving in period i.

4.1 Transition to continuous time

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 be equal to one time unit. Then $a(i\Delta t) = a_i$. 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 $\Delta t \to 0$. Assuming $a(\cdot)$ is a differentiable function, then $\lim_{\Delta t \to 0} \Delta a(t)/\Delta t$ exists and is denoted the *derivative of* $a(\cdot)$ at t, usually written $da(t)/dt$ or just $\dot{a}(t)$. That is,

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

By implication, wealth accumulation in continuous time is written

$$\dot{a}(t) = s(t), \qquad a(0) = a_0 \text{ given,} \tag{18}$$

where $s(t)$ is the saving 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 let us choose one year. Going back to discrete time, if wealth grows at the constant rate $g > 0$ per year, then after i periods of length one year (with annual compounding)

$$a_i = a_0(1 + g)^i, \qquad i = 0, 1, 2, \ldots \,. \tag{19}$$

When compounding is n times a year, corresponding to a period length of $1/n$ year, then after i *such* periods:

$$a_i = a_0(1 + \frac{g}{n})^i. \tag{20}$$

With t still denoting time (measured in years) that has passed since the initial date (here date 0), we have $i = nt$ periods. Substituting into (20) 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 (and so m) $\to \infty$. Then, in the limit there is continuous compounding and

$$a(t) = a_0 e^{gt}, \tag{21}$$

where e is the "exponential" defined as $e \equiv \lim_{m \to \infty}(1 + 1/m)^m \simeq 2.718281828....$

The formula (21) is the analogue in continuous time (with continuous compounding) to the discrete time formula (19) with annual compounding. Thus, a geometric growth factor is replaced by an exponential growth factor.

We can also view the formulas (19) and (21) as the solutions to a difference equation and a differential equation, respectively. Thus, (19) is the solution to the simple linear difference equation $a_{i+1} = (1+g)a_i$, given the initial value a_0. And (21) is the solution to the simple linear differential equation $\dot{a}(t) = ga(t)$, given the initial condition $a(0) = a_0$. With a time-dependent growth rate, $g(t)$, the corresponding differential equation is $\dot{a}(t) = g(t)a(t)$ with solution

$$a(t) = a_0 e^{\int_0^t g(\tau)d\tau}, \tag{22}$$

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 (22) is called the *basic growth formula* in continuous time and the factor $e^{\int_0^t g(\tau)d\tau}$ is called the *growth factor* or the *accumulation factor*.

Notice that 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

$$\dot{K}(t) = I(t) - \delta K(t), \qquad K(0) = K_0 \text{ given}, \tag{23}$$

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, in (23) $\delta > 1$ is conceptually allowed. Indeed, suppose for simplicity that $I(t) = 0$ for all $t \geq 0$; then (23) gives $K(t) = K_0 e^{-\delta t}$ (exponential decay). 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, 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 overlapping generations 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. Then with 30 years as our time unit, inserting into the formula $1/2 = e^{-\delta/3}$ gives $\delta = (\ln 2) \cdot 3 \simeq 2$.

4.2 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 and saving). A *stock* variable is a variable measured as just 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 $I(t)$ above are flow variables.

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

Thus, in continuous time analysis expressions like $K(t) + I(t)$ or $K(t) + \dot{K}(t)$ are 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 contains 50 liters of water and the tap pours $\frac{1}{2}$ liter per second into the tub. 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 $50\ \ell + \frac{1}{2} \cdot 60\ ((\ell/\text{sec.})\times\text{sec.}) = 90\ \ell$. 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 $[0, \infty)$. 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, $\int_{t_0}^{t_0+\Delta t} I(t)dt \approx I(t)\Delta t$. Similarly, $s(t)$, that is, the flow of individual saving, 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 Fig. 3.

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. But, it is really 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.

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

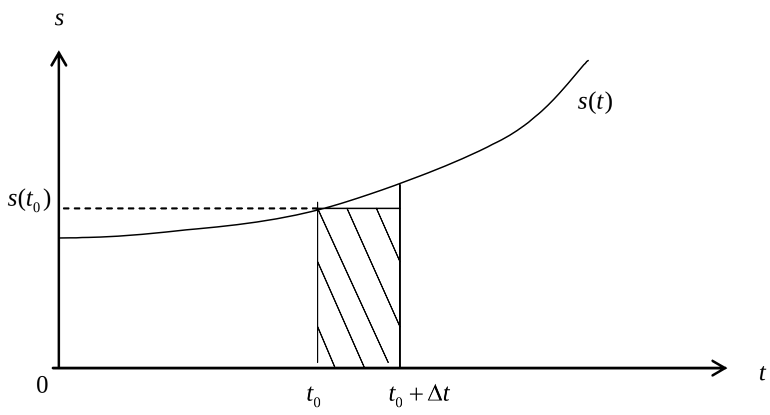


Figure 4: With Δt "small" the integral of $s(t)$ from t_0 to $t_0 + \Delta t$ is $\approx$ the hatched area.

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 be a good reason for putting at least microeconomic analysis in period terms. Yet, 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 and then a formulation in continuous time may be preferable. This also allows variation *within* the usually artificial periods in which the data are chopped up.[9] For example, stock markets (markets for bonds and shares) are more naturally modelled in continuous time because such markets equilibrate almost instantaneously; they respond immediately to new information.

In his discussion of this modelling issue, Allen (1967) concluded that from the point of view of the economic contents, the choice between discrete time or continuous time analysis may be a matter of taste. But from the point of view of mathematical convenience, the continuous time formulation, which has worked so well in the natural sciences, is preferable.[10]

5 References

Allen, R.G.D., 1967, *Macro-economic Theory. A mathematical Treatment*, Macmillan, London.

[9] Allowing for such variations may be necessary to avoid the *artificial* oscillations which sometimes arise in a discrete time model due to a too large period length.

[10] At least this is so in the absence of uncertainty. For problems where uncertainty is important, discrete time formulations are easier if one is not familiar with stochastic calculus.