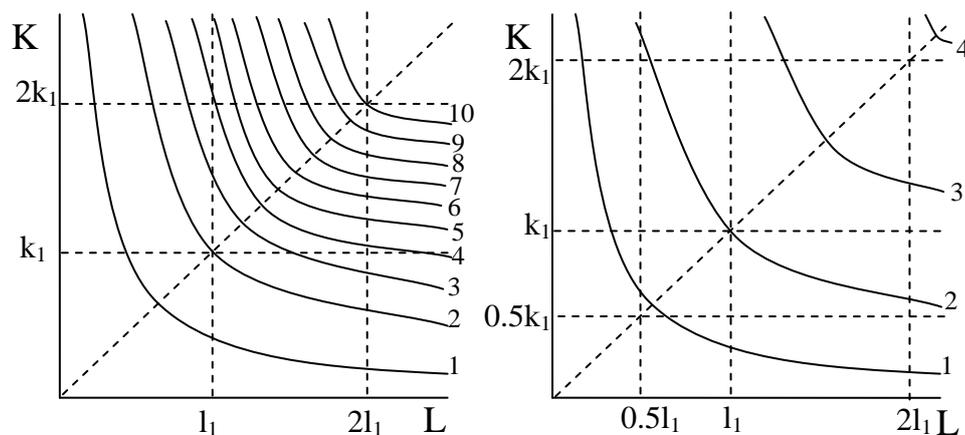


Week 5 – Production Theory

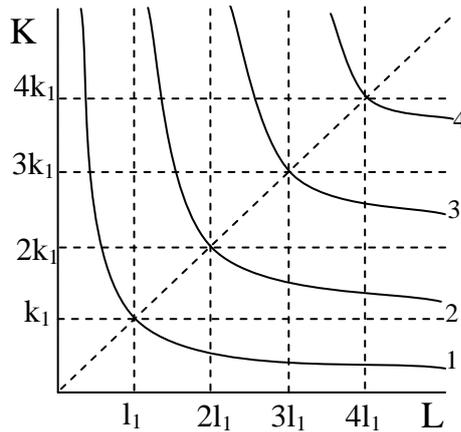
1. (a) The production function is a mathematical formula which describes the relationship between a firm's capital and labour inputs and its output. It is analogous to a utility function, except that whereas the utility values placed on indifference curves are irrelevant from the point of view of the consumer's choice of the optimal bundle (i.e. many different utility functions can represent the same preference orderings), the quantity at an isoquant is of vital importance, since it is the quantity which can be produced at a particular cost which is one of the key elements in the firm's profit maximization decision. Although there are important differences, there are also very close parallels between the problem of maximizing utility subject to remaining within a particular budget set and that of minimizing costs subject to producing a particular quantity (i.e. reaching a particular isoquant).

Just as the standard assumption in the simple two good model of consumer choice is that the two goods are imperfect substitutes, the standard assumption in production theory is that the two inputs, usually capital and labour, are imperfect substitutes. This implies that isoquants are strictly convex, so that an average between two points (i.e. a more even mixture between capital and labour) will result in a higher production level. Although it does not generally make sense to talk about the "spacing" of the indifference curves in terms of the utility difference between them, because the units in which we measure utility are arbitrary, it does make sense to talk about the spacing of isoquants because the units in which we measure quantity have a definite meaning. For example, suppose we have two firms which manufacture the same product, whose isoquant maps are shown by the diagram below. Not only can we say that the firm with more spaced out isoquants on the right is less productive at output levels higher than 1 unit than the firm on the left, we can see that the firm on the left is exhibiting increasing returns to scale (i.e. when both capital and labour inputs are doubled from l_1 and k_1 to $2l_1$ and $2k_1$, output more than doubles) whereas the firm on the right is experiencing decreasing returns to scale above an output of 2 (when capital and labour inputs are doubled from l_1 to l_2 , output less than doubles).



- (b) A firm which experiences constant returns to scale has a production function such that the output is multiplied by t whenever all inputs are multiplied by t . A commonly used example is a Cobb-Douglas production

function, which, for the simple two-input model, takes the form $y = Ak^\alpha l^{1-\alpha}$ (where A is a constant) Suppose starting from output k_1 and l_1 , so that $y_1 = Ak_1^\alpha l_1^{1-\alpha}$, both inputs are multiplied by t. This means $y_2 = A(tk_1)^\alpha (tl_1)^{1-\alpha} = At(k_1^\alpha l_1^{1-\alpha}) = ty_1$. Diminishing returns to each factor apply because if either capital or labour input is multiplied by t, with the other input left fixed, the overall amount produced is not multiplied by t. For example, if $y = k^{0.5} l^{0.5}$ then multiplying k or l by t only multiplies the output by the square root of t. The diagram below illustrates a strictly convex production technology with constant returns to scale.



To understand better the meaning of diminishing returns to a factor, we need to look at the derivatives of the production function with respect to capital and labour. If we take the general Cobb-Douglas production function:-

$$y = Ak^\alpha l^{1-\alpha}$$

$$\frac{\partial y}{\partial k} = \alpha Ak^{\alpha-1} l^{1-\alpha} = A\alpha (l/k)^{1-\alpha}$$

$$\frac{\partial y}{\partial l} = (1-\alpha) Ak^\alpha l^{-\alpha} = A(1-\alpha) (k/l)^\alpha$$

As l increases (with k held constant), the derivative with respect to l decreases, and as k increases (with l held constant), the derivative with respect to k decreases. The economic intuition behind the mathematics is that if you hold the amount of capital fixed and give the firm extra units of labour, each additional unit of labour is less and less useful to the firm as you give it more and more because the firm has a convex production technology, meaning that capital and labour are imperfect substitutes. i.e. more productive when used in a balanced proportion than when there are extreme amount of one or the other. A similar argument would apply if you hold the amount of labour fixed and give it more capital; each additional unit of capital becomes less and less useful because the firm does not have the labour to use in complement with it. Note that although there are diminishing returns to labour and capital, with a Cobb-Douglas production function it is still the case that an extra unit of labour and capital is always useful to some degree (i.e. Cobb-Douglas production functions do not allow for the possibility that sooner or later extra workers with no machines to use or machines with no workers to operate them would actually get in the way and *reduce* productivity).

(c) It has been argued that decreasing returns to scale cannot occur in practice because it should always be possible to replicate the firm's activities at the optimal level. For example, suppose that the isoquant map on the right of the previous page represents the technology of a shoe factory. Suppose that the isoquants are numbered in thousands of pair of shoes per week. Suppose

the firm starts off with 1 factory producing 2000 pairs of shoes each week using inputs (l_1, k_1) . We can see that by halving both inputs the output is more than halved, and that by doubling both inputs, output is less than doubled. However, if a firm wanted to produce, say, 8000 pairs of shoes, arguably all it would have to do would be to set up 4 separate factories, so that the isoquant map represents the technology for each individual factory, and then set each factory's inputs at (l_1, k_1) , so that each factory produces 2000 pairs of shoes. This exercise could be repeated for any number of pairs of shoes. So, to produce 100000 pairs of shoes the firm would open 50 factories, each with inputs (l_1, k_1) and each producing 2000 pairs of shoes.

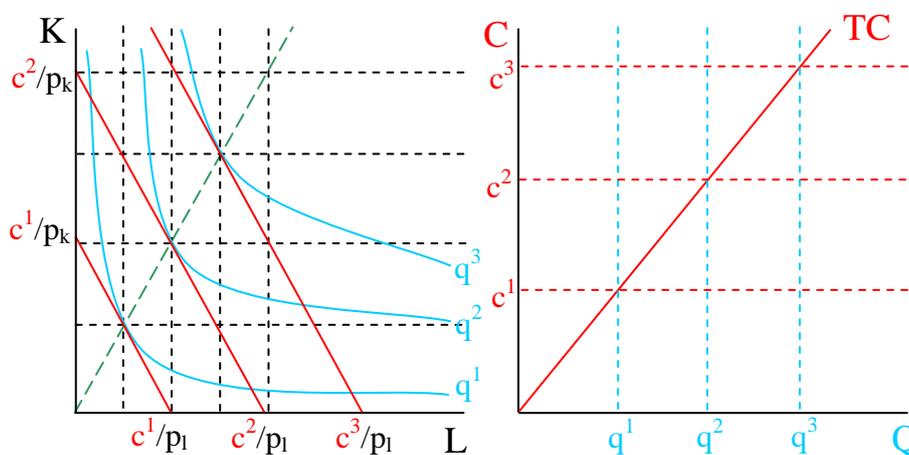
The above argument suggests that once a firm has reached the minimum efficient scale (i.e. the point where economies of scale have been fully exploited, as with 2000 pairs of shoes in the example above, which is the same as the point of minimum average cost, as we shall see later) then it should exhibit constant returns to scale from then on. However, if this were the case, it would be difficult to explain why one single firm does not control the entire economy, since if production can always be carried out at the optimal scale as described here, the surely the cost savings from not having to duplicate management structures would mean that there would be strong economic incentives for firms to amalgamate into one big firm.

Clearly, something is missing from the simple model of production technology laid out above. One of the most important is that there is much more to the firm than just putting labour and capital together in the correct quantities. Firms are bureaucracies consisting of a number of individuals. Unlike the model of consumer behaviour, where there are good reasons to think that individuals will, generally at least, behave rationally according to their preferences as represented by their indifference curve map, there is much less a priori reason to think that firms will behave like a rational individual. Rather, we would expect that the individuals within the firm will act to maximize their own utilities. The classic argument why firms will eventually experience decreasing returns to scale is the same as one of the common arguments against the planned economy; in a large bureaucracy where employees are paid according to hours worked rather than the results produced, it becomes more and more difficult to provide the incentives to work hard as the bureaucracy becomes larger and larger. So, a firm which owns thousands of factories is more likely to suffer from bureaucratic "flab" than one which owns a single factory. The other factors which produce economies of scale may still outweigh this effect, but sooner or later we would generally expect decreasing returns to scale to set in.

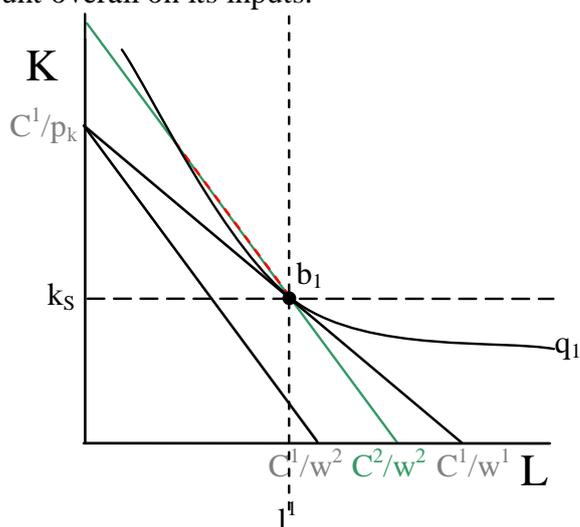
(d) The concepts of increasing returns to scale and economies of scale are related, but not the same. Increasing returns to scale refers to the production function; a technology exhibits increasing returns to scale if the output increases more than proportionately with all of the inputs. Economies of scale refer to the cost function; a firm is experiencing economies of scale if its long run average cost is decreasing in its production quantity. Increasing returns to scale provide one possible cause of economies of scale (see question 3 for an example). A firm which is experiencing increasing returns to scale must also be taking advantage of economies of scale as it approaches its quantity where average cost is minimized.

However, it is possible for economies of scale to exist even though there are constant or decreasing returns to scale in the firm's production technology. This can occur if the firm has fixed costs which it must pay independently of the amount it produces. For example, suppose a firm uses a constant returns to scale technology so that the marginal cost is always C . Suppose it must pay a fixed cost F in order to produce at all. The total cost will be given by $TC=CQ+F$. The average cost will therefore be $AC=TC/Q=C+F/Q$. Here the firm experiences economies of scale even though its technology exhibits constant returns to scale, because AC is always decreasing in Q . Intuitively, if the firm must pay a fixed cost F , then by producing more units it spreads this cost out more and so must pay a lower cost per unit.

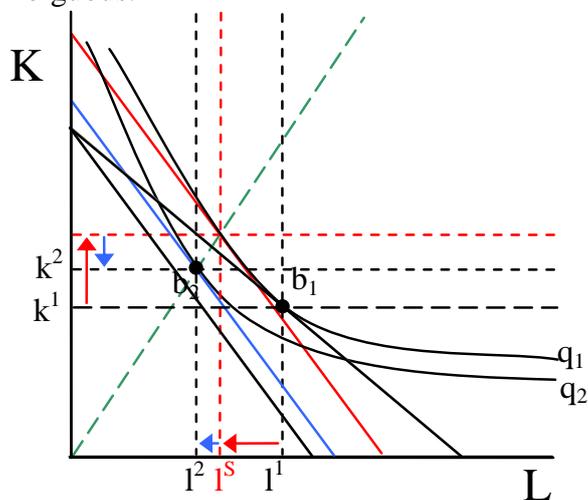
2. (a) The firm's cost minimization problem is very similar to a consumer's utility maximization problem. The budget constraint is replaced by an *isocost line*, whose slope is determined by the relative prices of capital and labour, just as the slope of the budget constraint is determined by the relative prices of the two goods. The main difference between the two situations is that whereas the consumer is told their budget set and must choose a bundle which gets them to the highest indifference curve, the firm, when trying to work out its long run total cost curve, is told the quantity it must produce (and therefore the isoquant upon which the optimal bundle must lie) and must choose the bundle which lies on the lowest isocost line. As with the consumer's problem, the optimal choice must satisfy a tangency condition between the isoquant and the isocost line. The diagram below illustrates the process of mapping out the firm's long run total cost curve, where the price of capital is p_k and the price of labour is p_l . Here we are assuming that the firm experiences constant returns to scale at all levels of output and price ratios of capital and labour. A necessary and sufficient condition for this is that the production function be homogeneous of degree 1 (like the Cobb-Douglas function discussed above). The definition in terms of the production function is $f(tl_1, tk_1) = tf(l_1, k_1)$ for all k_1, l_1 and t . This implies that all of the optimal production "bundles" generated by holding the price ratio of labour and capital (i.e. the slope of the isocost lines) fixed and altering the desired production quantity lie on a straight line from the origin (the green dashed line below). Changing the price ratio changes the slope of this line. The total cost (TC) curve will be a straight line (assuming no fixed cost), and so the marginal cost will be constant.



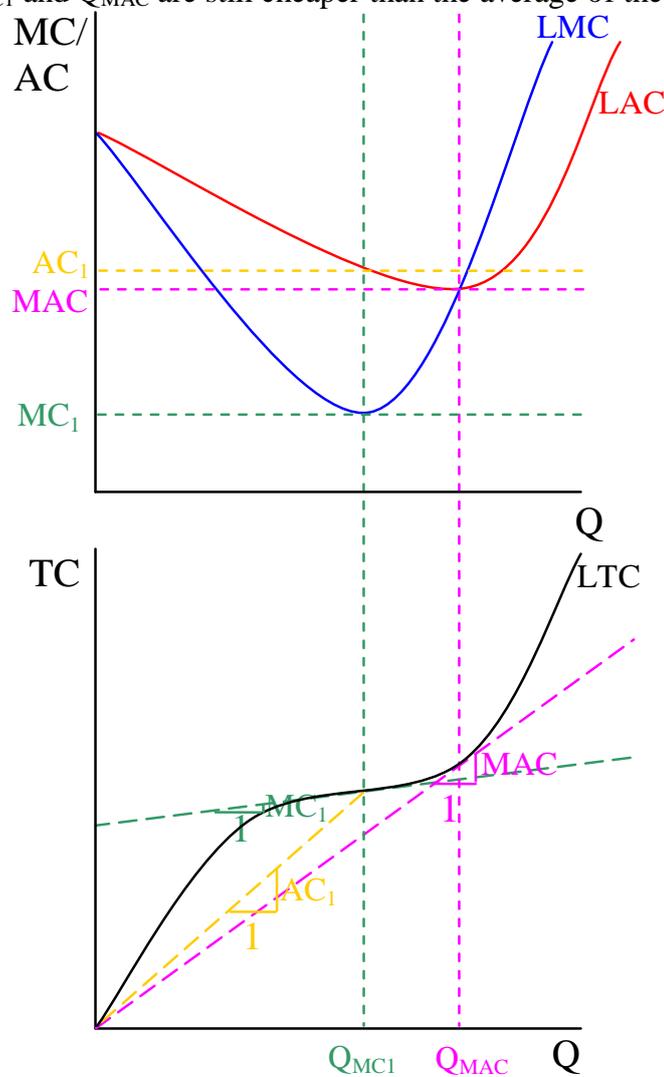
(b) If the wage increases from w^1 to w^2 whilst the capital stock is fixed at k_s in the short run, then in order to continue producing the same quantity q_1 , the firm will have to continue employing l^1 units of labour. This means that the total cost of production will unambiguously increase from C^1 to C^2 . Note that the firm is no longer at its long run output maximizing mixture of capital and labour given its current production costs. If the firm were able to freely exchange its capital and labour inputs at the current prices, then it could produce a higher output somewhere along the dotted line whilst spending the same amount overall on its inputs.



(c) If the profit maximization output level falls as a result of the wage rise, then provided the production function is homogeneous, we can say that the amount of labour demanded will unambiguously fall. The argument goes as follows: The increase in the price of labour (i.e. wage), whilst holding the output fixed at q_1 , leads to a substitution towards capital and away from labour (this is a bit like a Hicksian substitution effect). The decrease in the optimal production quantity from q_1 to q_2 will then cause a shift down the dotted diagonal ray towards the origin (due to the assumption that the production function is homogeneous). This is a bit like an income effect, and is shown in blue. Both effects work in the same direction for labour, so labour demand is unambiguously reduced, whilst the two effects work in opposite directions on the level of capital, so the overall effect of the wage increase on capital demand is ambiguous.



3. (a) The long run total cost curve exhibits economies of scale up to the point of minimum average cost Q_{MAC} , and then diseconomies of scale beyond this point. The average cost at each point on the total cost curve is the slope of the line connecting each point to the origin (because $LAC=LTC/Q$). The marginal cost is the slope of the tangent to the total cost curve. At the point of minimum average cost (MAC), the line connecting the point on the total cost curve to the origin has the same slope as the tangent to the total cost curve. This corresponds to the intersection between the LMC and LAC curves at the point of MAC. Assuming no fixed costs, the LAC and LMC curves also meet at the point of 0 output, because in the limit the slope of the line connecting the origin to a point on the LTC curve comes closer and closer to the slope of the tangent to the LTC curve as that point gets closer and closer to 0. (If there is a fixed cost, then the AC curve must go to infinity as the quantity goes to 0). The quantity at which marginal cost is minimized Q_{MC1} corresponds to the point where the LTC curve has the shallowest gradient. Note therefore that increasing returns to scale are exhausted before economies of scale are exhausted, because additional units between Q_{MC1} and Q_{MAC} are still cheaper than the average of the previous units.



- (b) The short run total cost (STC) curve must always lie above the long run total cost (LTC) curve except at the quantity Q_{MAC} because it is only at this quantity that the fixed level of the capital stock is the same as the optimal quantity that would be

chosen by the firm in the long run. At all other output levels, the STC must be higher because the restriction on altering the capital stock prevents the firm from fully minimizing its costs in the short run (and we know from the assumption of homogeneous production function that the firm will want to increase the level of capital above Q_{MAC} and decrease it to produce below Q_{MAC}), so costs are higher than at the long run cost minimizing mixture of capital and labour. The STC curve cuts the y axis at a positive amount because even if 0 units are produced, the opportunity cost of the fixed capital stock still has to be counted as a cost. This means that the SAC must go to infinity as Q goes to 0. The short run marginal cost (SMC) curve lies above the LMC curve to the right of Q_{MAC} but below the LMC curve to the left of LMC because when output is below Q_{MAC} in the short run it is less costly to expand production because the capital supply is already fixed above the level the firm would optimally choose, and so less has to be paid on extra capital per marginal unit, whilst when short run output is above Q_{MAC} , it is more costly at the margin to expand output because capital cannot be freely adjusted to get the best mix of inputs for each extra unit. The fact that the SMC curve lies below the LMC curve to the left of Q_{MAC} and above LAC to the right of it can also be seen from the fact that the STC curve has a shallower slope to the left of Q_{MAC} and a steeper slope to the right of it. The SAC curve always lies above the LAC curve, except at Q_{MAC} , which is where the optimal short run capital stock is the same as the fixed level.

