CHAPTER 3.
CLASSICAL WELFARE ECONOMICS AND MARKET FAILURE IN HEALTH CARE, PART I

1. Introduction

The overall topic of the next two chapters, *Classical welfare economics and market failure in health care*, is dealt with in two parts. The main purpose of this chapter (Chapter 3) is to set out the classical welfare economic framework that underlies many of the public policy debates about health care provision. We will also look again at the objectives society might have with respect to health care provision, and briefly discuss the economic appraisal of health care programmes. In Chapter 4, we will examine in detail the possibility of using the market system to allocate such care: its advantages and disadvantages as a means of achieving society's objectives. In the light of market failures, we will discuss the various ways in which the state can intervene, and the associated advantages and disadvantages of intervention.

The detailed programme for Chapters 3 and 4 is as follows:

Chapter 3	Section 2. Basic consumer and producer theory.
	Section 3. Society's economic problem.
	Section 4. The three optimality conditions for efficiency.
	Section 5. The fourth optimality condition: social justice.
	Section 6. The market system.
	Section 7. Society's objectives: efficiency and social justice in health care.
	Section 8. The economic appraisal of health care programmes.
Chapter 4	Section 1. The market system and health care.
	Section 2. Sources of market failure: why do we need government intervention?
	Section 3. Government policies: regulation, taxes/subsidies, direct provision.
	Section 4. Government policies and objectives: an assessment.

2. Basic consumer and producer theory

We begin in this section by setting out the basic consumer and producer theory we will need.

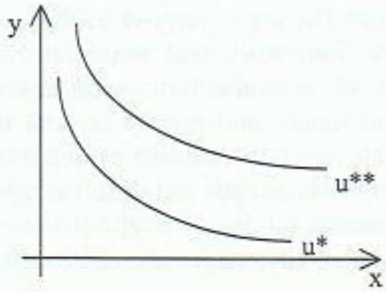
2.1. Consumer theory

At a fundamental level, economics is simply about choosing the <u>best way</u> to use <u>scarce resources</u> to produce goods and services which satisfy our material wants. Finding the 'best' way means optimising i.e. maximising or minimising. In the elementary theory of consumer choice (with which you should be familiar!), this manifests itself as the problem of maximising <u>utility</u> subject to a <u>budget constraint</u>. Consumers derive (or 'produce') <u>utility</u> (i.e. happiness, satisfaction etc.) through the consumption of various goods and services (commodities). Their consumption is limited by a <u>budget constraint</u>, which describes the different bundles the consumers can afford, given market prices and income levels. In the simple case where a consumer has to choose how much to consume of each of two commodities (call them x and y), we can use <u>indifference curve analysis</u> to describe and solve the problem. Recall that an <u>indifference curve</u> shows all the consumption bundles (i.e.

combinations of quantities of x and y) which yield the same utility to the consumer. We assume that x and y yield utility according to the function

$$u = u(x, y)$$

The indifference curve is obtained by fixing u at a particular level, say $u = u^*$, and plotting on a graph the different combinations of quantities of x and y that yield utility level u^*. The following diagram shows two such indifference curves, one corresponding to a utility level u^*, and another corresponding to a higher utility level u^{**}:



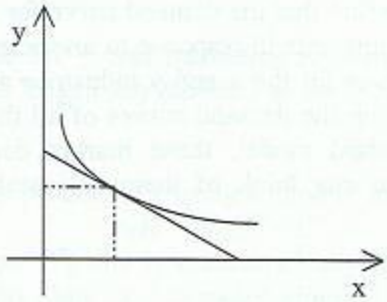
The slope of an indifference curve at any particular point is the marginal rate of substitution of y for x at that point, denoted by MRS_{yx}. This measures how much extra y the consumer needs to compensate him/her for having one less unit of x i.e. how much extra y is needed to keep utility constant at u^* if the quantity of x is reduced by one unit. Indifference curves are assumed to be strictly convex to the origin, implying that as the consumer gets more and more y (and less and less x), he/she values the next unit of y (the 'marginal' unit) less and less, and so has to be given increasingly bigger quantities of y every time x is reduced. This results in the 'convex' shape of the indifference curves above. We can express this idea in terms of diminishing marginal utility. The marginal utility of x, denoted by MU_x, is the increase in utility produced by consuming one more unit of x. (Formally, it is the *partial derivative* of u with respect to x i.e. $\partial u/\partial x$). Similarly, the marginal utility of y, MU_y, is the increase in utility produced by consuming one more unit of y (formally, the partial derivative $\partial u/\partial y$). We assume that these marginal quantities decrease as the amount consumed increases, reflecting the fact that consumers value extra units of a particular commodity less and less as the quantity consumed increases. Conversely, marginal utilities increase as the quantities consumed decrease, reflecting the fact that consumers attach greater marginal value to commodities when they have less of them. It can be shown that

$$MRS_{yx} = -\frac{MU_x}{MU_y}$$

Reducing x will increase MU_x, and increasing y will reduce MU_y, so the indifference curve will become steeper (i.e. more 'negatively sloped') as the quantity of x falls and the quantity of y rises. It is assumed that consumers prefer more of any particular commodity to less, and any point on a higher indifference curve (such as the curve for u^{**} in the diagram above) is preferred to any point on a lower indifference curve (such as the curve for u^*). The budget constraint can be drawn as a downward sloping budget line, whose equation is

$$p_x x + p_y y = m$$

where p_x is the price of x, p_y is the price of y, and m is income. Maximising utility subject to this budget constraint involves finding the highest indifference curve that is still 'touching' the budget line. This happens at the values of x and y for which the slope of the indifference curve equals the slope of the budget line i.e. at the point where the budget line is tangential to the indifference curve:



Formally, what you are doing is solving the following constrained optimisation problem:

$$\max_{x,y} \; u = u(x, y)$$

$$\text{subject to } p_x x + p_y y = m$$

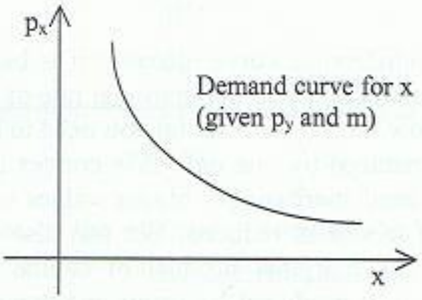
This can be done mathematically using partial differentiation and Lagrange-multiplier methods. At the optimum, the slope of the budget line equals the marginal rate of substitution of y for x. It can easily be shown that the slope of the budget line is

$$\frac{dy}{dx} = -\frac{p_x}{p_y}$$

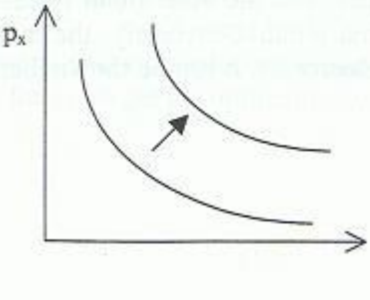
Thus, the optimality condition for utility maximisation subject to a budget constraint is

$$MRS_{yx} = -\frac{p_x}{p_y}$$

The optimal values of x and y that you get by solving this problem depend on p_x, p_y and m. These are called the parameters of the problem. We can derive the demand curve for good x by seeing what happens to the solution of the above optimisation problem as we change the value of the parameter p_x, keeping p_y and m constant. Typically, we would find that the optimal value of x is smaller, the higher is p_x. If we plot a graph of the optimal values of x for different values of p_x (assuming that p_y and m are constant) we get the familiar downward sloping demand curve:



Thus, for given p_y and m, lower quantities of x are demanded at higher values of the price p_x. If we now consider increasing the level of income m (keeping p_y constant), we will normally find that the demand curve for x shifts outwards, indicating that more x is demanded at each price p_x:



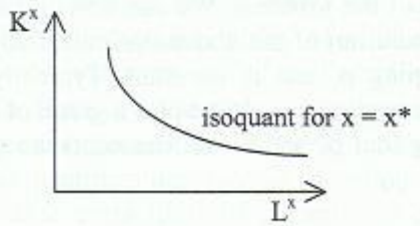
When this happens, we call x a 'normal good'. Similarly, we can derive the demand curve for good y by examining what happens to the optimal value of y as p_y varies, keeping p_x and m constant. Again, we will typically find that the demand curve for y is downward sloping, and that the demand curve will shift outwards in response to an increase in income (all else held constant). To get the demand curves for the x and y industries as a whole (i.e. the 'market' demand curves), we can just add up the demand curves of all the indvidual consumers of x and y respectively. In the standard model, these market demand curves will also be downward-sloping; basically, you can think of them as just horizontal translates of the individual demand curves above.

2.2. Producer theory

We can visualise producers as solving a very similar constrained optimisation problem. They try to combine inputs of capital, K, and labour, L, to produce as much output as they can, for any given level of money cost. For example, let K^x and L^x denote the quantity of capital and labour used to produce commodity x. Also let w_K denote the price that producers have to pay for each unit of capital they use, and let w_L denote the wage rate (the price producers have to pay for each unit of labour they use). The technology of production (i.e. how you have to combine inputs to produce output) is represented by a production function. The production function for commodity x is written as

$$x = x(K^x, L^x)$$

It tells you how much x you can get, given particular quantities of capital and labour. Note that its role in producer theory is exactly the same as that of the utility function in consumer theory. If we fix the quantity of x at a particular level, say $x = x^*$, we can plot a graph called an isoquant to describe all the combinations of K^x and L^x that produce the quantity x^*:



It looks exactly the same as an indifference curve, because it is based on the same sorts of assumptions. The slope of the isoquant is called the marginal rate of substitution of K^x for L^x, denoted by MRS^x_{KL}. It tells you how much more capital you need to keep the production level at x^*, if the amount of labour is reduced by one unit. It is convex to the origin (just like an indifference curve) because you need increasingly bigger values of K^x to keep production constant at x^*, as the amount of labour is reduced. We can also express this in terms of diminishing marginal products. The marginal product of capital in the production of x, denoted by MP^x_K, is the increase in x produced by using one more unit of capital, keeping labour constant (formally, it is the *partial derivative* of x with respect to K^x). Similarly, the marginal product of labour in the production of x, MP^x_L, is the increase in x produced by using one more unit of labour, keeping capital constant (formally, it is the partial derivative of x with respect to L^x). We assume that the marginal product of an input decreases as the quantity of the input increases, because with the other input fixed, you can produce less and less by continuing to increase the first input. Conversely, the marginal product of an input increases as the quantity of the input decreases. It can be shown that

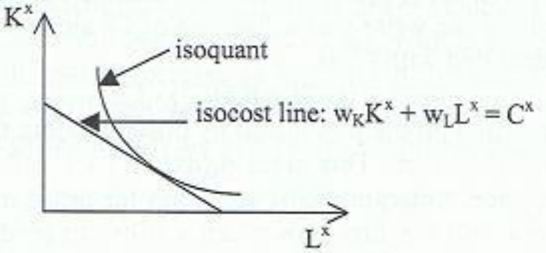
$$MRS^x_{KL} = -\frac{MP^x_L}{MP^x_K}$$

Reducing L^x will increase MP^x_L, and increasing K^x will reduce MP^x_K, so the isoquant will become steeper (i.e. more 'negatively sloped') as the quantity of L^x falls and the quantity of K^x rises.

The total money cost C^x of producing x using K^x and L^x at prices w_K and w_L is given by

$$w_K K^x + w_L L^x = C^x$$

For any given money cost C^x, this is mathematically identical to the budget constraint in consumer theory, and also plots as downward sloping line. A firm having the production function $x(K^x, L^x)$ will always want to produce as much x as it can, for any given money cost C^x. This involves finding the highest isoquant that is still 'touching' the cost line corresponding to C^x. This happens at the values of K^x and L^x for which the slope of the isoquant equals the slope of the cost line i.e. at the point where the cost line is tangential to the isoquant:



Formally, what you are doing is solving the following <u>constrained optimisation problem</u>:

$$\max_{K,L}\ x = x(K^x, L^x)$$

$$\text{subject to } w_K K^x + w_L L^x = C^x$$

As in consumer theory, this can be done mathematically using partial differentiation and Lagrange-multiplier methods. At the optimum, the <u>slope</u> of the cost line equals the marginal rate of substitution of capital for labour. It can easily be shown that the slope of the cost line is

$$\frac{dK^x}{dL^x} = -\frac{w_L}{w_K}$$

Thus, the <u>optimality condition</u> for maximisation of output subject to a cost constraint is

$$MRS^x_{KL} = -\frac{w_L}{w_K}$$

This looks very similar to the optimality condition in consumer theory because it is the solution to a mathematically identical problem.

Precisely the same arguments apply to the production of commodity y. We can assume that the production function is

$$y = y(K^y, L^y)$$

where K^y is the amount of capital used to produce y, and L^y is the amount of labour used to produce y. The cost constraint in this case becomes

$$w_K K^y + w_L L^y = C^y$$

where C^y is the total money cost of producing y using K^y and L^y. Solving the same sort of optimisation problem as for x, we get the optimality condition

$$MRS^y_{KL} = -\frac{w_L}{w_K}$$

For the consumer, this was the end of the problem. However, in producer theory we have to go one step further: we must consider profits (i.e. revenues minus costs). The total revenue a firm gets from selling an amount x is simply $p_x x$. Its profit is the difference between this and total cost:

$$\text{profit} = p_x x - C^x$$

Thus, the solution to the above constrained optimisation problem for given values of w_K, w_L and C^x not only determines the optimal values of K^x and L^x and the associated output level $x(K^x, L^x)$; for any given p_x, it also determines the profit level $p_x x - C^x$ at that level of output.

Both the total revenue $p_x x$ and the total cost C^x rise and fall together (the more you produce, the higher are your total costs), but in general the distance between them (i.e. the profit) varies as C^x varies. For given p_x, w_K and w_L, we imagine that the firm solves its constrained optimisation problem for all possible values of C^x and then simply 'chooses' the C^x which has the highest associated profit level $p_x x - C^x$. Clearly, each C^x implies a value of x, so with everything else constant we can write x as a function of C^x alone i.e. $x = x(C^x)$. We can then write the profit level associated with C^x as

$$\text{profit} = p_x x(C^x) - C^x$$

The firm's pofit maximisation problem is solved by differentiating this with respect to C^x and setting the derivative equal to zero. This gives $p_x(dx/dC^x) - 1 = 0$ which rearranges to $p_x = dC^x/dx$ i.e. the familiar price = marginal cost condition for profit maximisation. Solving this first-order condition for x tells the firm how much x it has to produce in order to maximise profits. Similarly, a firm in the y industry has produce the level of y which satisfies $p_y = dC^y/dy$ to maximise profits.

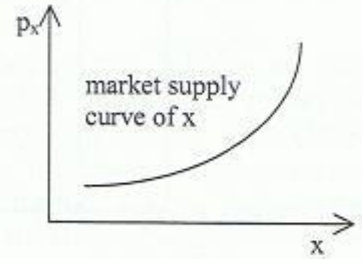
However, finding the x which satisfies the first-order condition $p_x = dC^x/dx$ is not enough. Clearly, it will not be worthwhile for a firm to produce any level of x at which the total revenue $p_x x$ is not high enough to cover the total cost C^x. This would involve making a loss instead of a profit ! Thus, a firm will only be willing to produce x if $p_x x$ is at least as big as C^x i.e. if

$$p_x x - C^x \geq 0$$

If $p_x x$ is always less than C^x, no matter what the value of x is, then the firm is better off not producing anything. Clearly, the higher is p_x (everything else held constant), the more likely it is that $p_x x$ is going to be at least as big as C^x for at least some values of x. Similarly, a firm in the y industry will only produce the profit-maximising level of y if $p_y y - C^y \geq 0$ at that level.

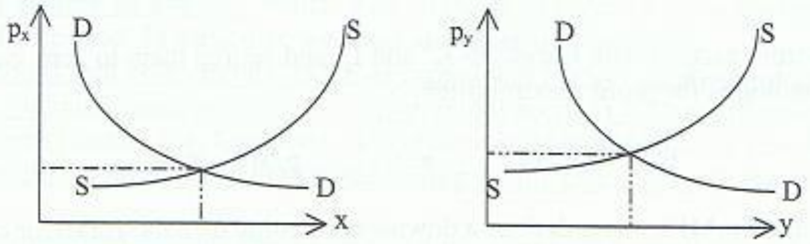
Note that the condition $p_x x - C^x \geq 0$ can be rearranged to read $C^x/x \leq p_x$ which says that the firm will only produce the profit-maximising level of output if it is such that the average cost is less than or equal to the price. But we know from the first-order condition that $p_x = dC^x/dx$, so the firm will only produce the profit-maximising level of output if average cost is less than or equal to marginal cost i.e. if $C^x/x \leq dC^x/dx$. It is easy to show that if the average cost is less than or equal to the marginal cost then the average cost must be non-decreasing (just differentiate C^x/x with respect to x to get $[dC^x/dx - C^x/x]/x$, and note that this derivative will be non-negative i.e. average cost will be non-decreasing, if and only if $C^x/x \leq dC^x/dx$). This is a very important result, because it tells us that our market model requires 'non-increasing returns to scale' in order to work (recall that returns to scale are increasing when the average cost curve is falling, constant when the average cost curve is horizontal, and decreasing when the average cost curve is rising). As we will see later, increasing returns to scale are a major source of 'market failure' in practice.

Now, returning to the condition $p_x x - C^x \geq 0$, note that different firms differ in their skill at combining inputs to produce outputs. Firms which are more 'technically efficient' will need to buy lower quantities of capital and labour to produce any given output level than firms which are less 'technically efficient', so they will be able to produce a bigger x for any given total money cost C^x. In other words, at the levels of K^x and L^x which are associated with the total money cost C^x (given input prices w_K and w_L), more efficient firms will be able to produce higher levels of x than less efficient firms. At a given price level p_x, this means that more efficient firms will make bigger profits for any given total money cost C^x than less efficient firms. At very low values of the price p_x, only the most technically efficient (i.e. low cost) producers will be able to find any values of x at which profits are non-negative. Thus, the market supply of x will be relatively low. As the price p_x rises, two things will happen. First, existing suppliers of commodity x will find it worthwhile to try to increase their technical efficiency (i.e. the quantity of x they can produce at any given money cost C^x) by investing in new technology which enables them to use less inputs. Second, producers who were previously too inefficient to make a profit at lower values of p_x will now find that it is profitable to supply x, and will begin to do so. Thus, with everything else held constant, the market supply of x will tend to increase as the price p_x increases. If we plot a graph of the quantity of x supplied at different values of p_x, we get the familiar upward sloping market supply curve, indicating that more is supplied at higher prices than at lower prices:



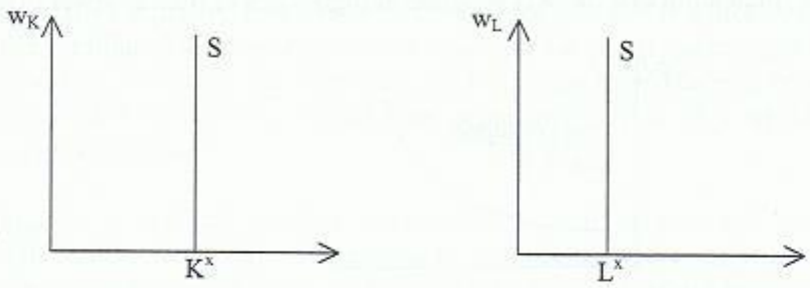
Exactly the same applies to commodity y. In general, we would expect to find that higher levels of y are supplied at higher values of p_y, and less y is supplied when p_y is relatively low.

We have now derived the market demand and supply curves for commodities x and y from first principles. All that remains is to put them together to produce the familiar demand and supply diagrams from elementary partial equilibrium analysis:



Equilibrium in the market for x occurs at the values of p_x and x for which demand = supply. Similarly, the equilibrium price and quantity of y are the ones at which the demand and supply curves intersect. An equilibrium is a 'state of rest' in which there is no incentive for any of the participants to change their behaviour (note: 'state of rest' does not mean that nothing is happening!). Since you should be familiar with the basics of demand and supply by now, we will leave it at that.

So far we have only considered demand and supply for the two output goods x and y. However, there is also a demand and supply story for the two input goods K and L. In the input (or 'factor') markets, our actors switch roles: the firms are the ones who do the 'demanding' and the consumers in our model above are the ones who now 'supply' capital and labour. The story is as follows. Each consumer i in the economy is assumed to own a quantity K^i of capital, and a quantity L^i of labour. This is his/her 'factor endowment'. Adding up everyone's quantity of capital and everyone's quantity of labour gives us the total endowments of these factors in the economy as a whole, denoted by $\overline{K}$ and $\overline{L}$. The consumer's money income on the right-hand side of the budget constraint is then computed as $m^i = w_K K^i + w_L L^i$, so that consumers are viewed as deriving their income from selling their factor endowments to producers. Each consumer i is maximising utility $u^i(x^i, y^i)$ subject to a budget constraint $p_x x^i + p_y y^i = m^i$, where we now interpret m^i as the money income derived from individual i's factor endowment. In our simple model, each consumer ends up selling the whole of his/her endowment at the prevailing prices w_K and w_L (which each consumer regards as being outside his/her control). Adding up the total amounts of capital and labour supplied to the x industry gives us two vertical supply curves:



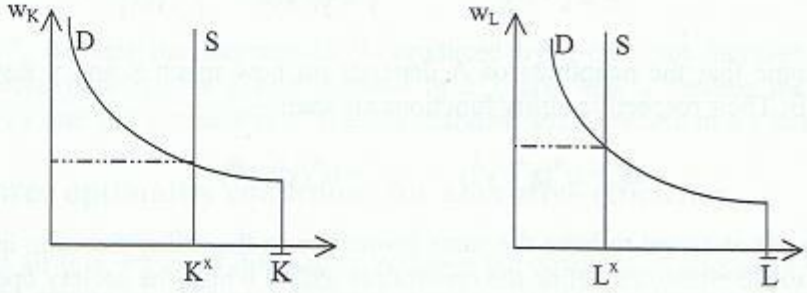
The demand curves in the factor markets are derived directly from firms' profit-maximising behaviour. We modelled this behaviour as consisting of two stages. First, the firm chooses K^x and L^x to maximise x given w_K, w_L and C^x. For given p_x, this implies a profit level $p_x x - C^x$. Then, the firm chooses the pair x and C^x at which $p_x x - C^x$ is maximised, and produces this x if $p_x x - C^x \geq 0$. However, solving these steps separately is equivalent to solving the following single problem:

$$\text{choose } K^x \text{ and } L^x \text{ to max. profit} = p_x x(K^x, L^x) - w_K K^x - w_L L^x$$

Taking first-order partials with respect to K^x and L^x and setting them to zero gives us the two first-order conditions for profit maximisation:

$$p_x MP^x_K = w_K \qquad \text{and} \qquad p_x MP^x_L = w_L$$

The first equation $p_x MP^x_K = w_K$ defines a downward sloping demand curve for capital in the x industry, and the second equation $p_x MP^x_L = w_L$ defines a downward sloping demand curve for labour in the x industry. To see this, suppose we increase w_K on the right-hand side of the first equation. To restore the equality, the left-hand side (which is called the 'marginal revenue product' of capital) must rise, but since p_x is constant, it can only rise if MP^x_K rises. Because of the diminishing marginal product property of production functions in our model, the only way to raise MP^x_K is to *reduce* K^x. Thus, a rise in w_K on the right-hand side of the equation $p_x MP^x_K = w_K$ above must be matched by a fall in K^x on the left-hand side. Conversely, a fall in w_K must lead to rise in K^x, so we must have a downward sloping demand curve for capital. Exactly the same applies to labour in the equation $p_x MP^x_L = w_L$. If w_L rises, the marginal revenue product of labour on the left-hand side must rise to restore the equality, but this can only happen if L^x falls. Conversely, if w_L falls, L^x must rise to restore the profit-maximising equality. Plotting the two demand curves together with the corresponding supply curves gives the partial equilibrium diagrams for the factor markets in the x industry:



Note that the demand curves are truncated at the total endowments $\overline{K}$ and $\overline{L}$. Following exactly the same lines of argument, we could obtain similar demand and supply diagrams for K^y and L^y.

Finally, note that the equations $p_x MP^x_K = w_K$ and $p_x MP^x_L = w_L$ imply $p_x = w_K/MP^x_K = w_L/MP^x_L$. But we know that the profit maximising condition is that $p_x = dC^x/dx$, so it must be the case that dC^x/dx = marginal cost of $x = w_K/MP^x_K = w_L/MP^x_L$. Similarly, marginal cost of y $= w_K/MP^y_K = w_L/MP^y_L$. The intuition for this is straightforward. Suppose, for example, that a firm with a production function $x(K^x, L^x)$ employs one more unit of capital. Its costs rise by w_K, and its output of x by MP^x_K. Therefore the increase in the cost per unit of extra output at the margin (i.e. the marginal cost) is just w_K/MP^x_K. So each individual firm will produce x until $p_x = w_K/MP^x_K$.

3. Society's economic problem

In the previous section, we saw that individual consumers and producers have to solve constrained optmisation problems in order to decide how much to consume or produce, given the resource constraints they face. In this section, we will set up a constrained optimisation problem <u>for society as a whole</u>, which will allow us to consider how a society's resources should ideally be used. In particular, we want to answer two questions:
(1). <u>How should factors or 'inputs' (such as capital and labour) be allocated among different products or 'outputs' (such as x and y in the previous section) ?</u> This will determine both the quantity of each product (i.e. *how much ?*), and the mix of inputs required (i.e. *how ?*).
(2). <u>How should the products be distributed among the different citizens?</u> (i.e. *for whom ?*)

Most of the important insights can be obtained by setting up a simple model involving only two people, called A and B. The economy is endowed with a certain (fixed) quantity of capital, denoted by $\overline{K}$, and a certain (fixed) quantity of labour time, denoted by $\overline{L}$. We will assume that these can be used to produce two outputs, x and y. Using the same notation as in the previous section, the allocation of factors between x and y is constrained by

$$\overline{K} = K^x + K^y \qquad \overline{L} = L^x + L^y$$

The production functions are the same as in the previous section:

$$x = x(K^x, L^x) \qquad y = y(K^y, L^y)$$

The total outputs of x and y must be allocated between A and B. If x^A is the amount of x allocated to A and so on, then

$$x = x^A + x^B \qquad y = y^A + y^B$$

We shall assume that the happiness of A depends on how much x and y he/she gets, and similarly for B. Their respective utility functions are then

$$u^A = u^A(x^A, y^A) \qquad u^B = u^B(x^B, y^B)$$

Both of these are assumed to have the same properties as the utility function in the previous section. The above equations define the constraints within which the society operates. Capital and labour are limited in supply, technology (represented by the production functions) limit the quantity of goods that can be produced by given factors, and tastes determine the happiness that can be obtained from consuming goods. We now introduce a <u>social welfare function</u>, which is a kind of utility function for 'society as a whole'. It is assumed that the only variables which affect social welfare are the utilities of individuals A and B, so the social welfare function is of the form

$$W = W(u^A, u^B)$$

This has exactly the same (mathematical) properties as the ordinary utility functions above. The only difference is that W is a utility function which measures society's happiness with respect to levels of u^A and u^B, rather than levels of goods or services. Society's constrained optimisation problem can now be written as follows:

Find the values of $K^x, K^y, L^x, L^y, x^A, x^B, y^A, y^B$ *which maximise* $W = W(u^A, u^B)$ *subject to*

$$\overline{K} = K^x + K^y \qquad \overline{L} = L^x + L^y \qquad \textit{(factor endowments)}$$
$$x^A + x^B = x(K^x, L^x) \qquad y^A + y^B = y(K^y, L^y) \qquad \textit{(technology)}$$
$$u^A = u^A(x^A, y^A) \qquad u^B = u^B(x^B, y^B) \qquad \textit{(tastes)}$$

Just as in the consumer and producer theories discussed in the previous section, society's constrained optimisation problem can be solved mathematically using partial differentiation and Lagrange-multiplier techniques. We get four optimality conditions which are fundamental to welfare economics, and tell us everything we need to know in order to answer the two questions posed at the beginning of this section. We will list the four conditions here, and discuss them in the next two sections.

(1). <u>Efficient consumption:</u> $\left(\dfrac{MU_x}{MU_y}\right)^A = \left(\dfrac{MU_x}{MU_y}\right)^B$

(2). Efficient production: $\left(\frac{MP_L^x}{MP_K^x}\right) = \left(\frac{MP_L^y}{MP_K^y}\right)$

(3). Efficient product-mix: $\left(\frac{MU_x}{MU_y}\right)^A = \left(\frac{MP_K^y}{MP_K^x}\right)$

(4). Social justice: $\frac{(MU_x)^B}{(MU_x)^A} = \frac{MW_u^A}{MW_u^B}$

where $MW^A{}_u$ denotes the increase in W produced by a one unit increase in u^A (i.e. the marginal social benefit of a 1 unit increase in u^A), and $MW^B{}_u$ denotes the increase in W produced by a one unit increase in u^B (i.e. the marginal social benefit of a 1 unit rise in u^B).

4. The three optimality conditions for allocative efficiency

To interpret all this, we must distinguish between the concepts of allocative efficiency (also called Pareto optimality or Pareto efficiency) and equity (also called social justice). A social state is said to be allocatively efficient if it is impossible to make one individual better off, without making another individual worse off. To be in such a position is simply to avoid wasting an opportunity of getting something for nothing. Any other position we call wasteful or inefficient. An allocation is said to be equitable if it also maximises the social welfare function (subject to the constraints discussed above). The three conditions necessary for allocative efficiency are those set out in the first three optimality conditions in the previous section. We consider them one by one.

4.1. Efficient consumption

Recall that the marginal rate of substitution of y for x is written in terms of marginal utilities as

$$MRS_{yx} = -\frac{MU_x}{MU_y}$$

What this tells you is how much y the consumer has to be given in order to compensate him/her for a one unit reduction in x. Thus, for example, if $MRS_{yx} = -1/2$, the consumer must be given 1/2 units of y in order to keep his/her utility level constant after a one unit reduction in x. In other words, the value of x to the consumer is half that of y. So MRS_{yx} measures the relative value attached to x and y by the consumer. The optimality condition for efficient consumption, i.e.

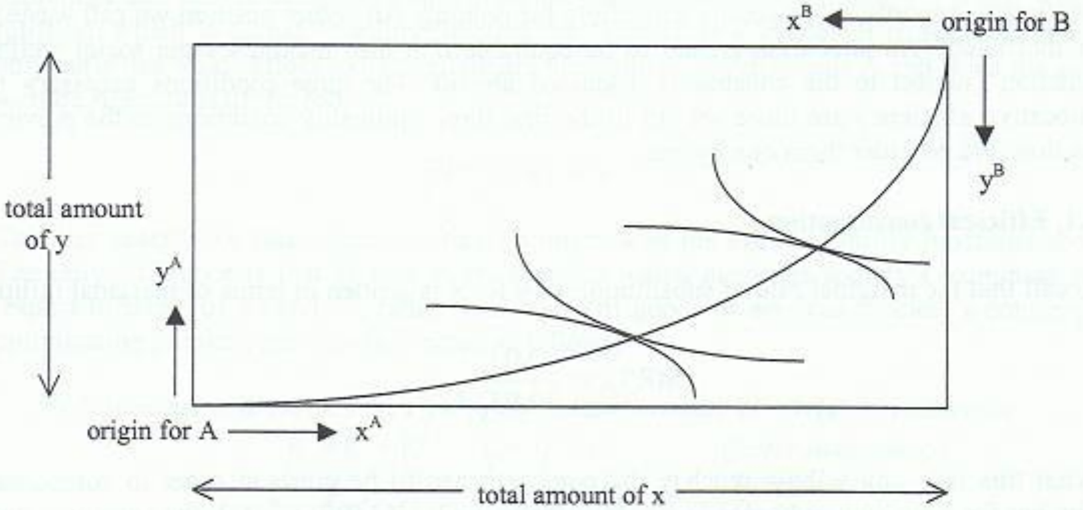
$$\left(\frac{MU_x}{MU_y}\right)^A = \left(\frac{MU_x}{MU_y}\right)^B$$

says that both A and B must place the same relative values on x and y. If this condition is not satisfied, then it must be possible to make one consumer better of without making the other worse off, so the social state cannot be 'allocatively efficient'.

To illustrate this, suppose that A and B place different relative values on x and y. Suppose that A must be given 2 units of y to keep his/her utility constant after a one unit reduction in x, whereas B needs only 1 unit of y to compensate him/her for a one unit reduction in x. Thus, B values x and y equally. If B gives A 1 unit of x in exchange for 1 unit of y, then B's utility is unaffected. But A is made better off, because A gets a unit of x which is worth twice as much to A as the 1 unit of y given to B. Thus, we have made A better off without making B worse off, so the social state could not have been allocatively efficient. This sort of thing cannot happen when A and B place the same relative values on x and y. Therefore:

Efficient consumption requires that all individuals place the same relative values on all products (values being assessed at the margin).

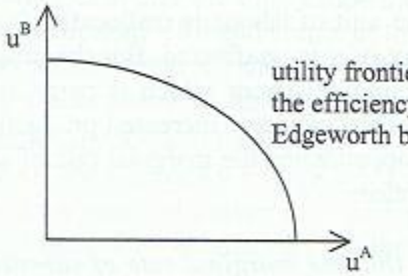
It is important to note that this is a necessary but not a sufficient condition for a social optimum. The reason is that, for any given quantities of the goods x and y, we can always find an infinite number of ways of allocating the goods between A and B such that the efficient consumption condition is satisfied ! Furthermore, the efficient consumption condition alone cannot tell us how much of each of the goods x and y we should produce. All we can say is that if the efficient consumption condition is not satisfied, then we can make one individual better off without making the other worse off. We can illustrate the fact that an infinite number of allocations will satisfy the efficient consumption condition using the economic tool known as Edgeworth's box.



The dimensions of the box are the total quantity of x in the economy, measured along the horizontal axis, and the total quantity of y, measured along the vertical axis. We make the bottom left-hand corner of the box the 'origin' of measurement for A's allocations of x and y, and the top right-hand corner the 'origin' for B's allocations. Any point in the box then represents a division of the total x into x^A and x^B, and likewise for y. With each point, there is therefore an associated utility level for A given by $u^A = u^A(x^A, y^A)$ and a utility level for B given by $u^B = u^B(x^B, y^B)$. We can draw indifference curves for A and B, A's emanating from the bottom left-hand corner, and B's emanating from the top right.

Now, the efficient consumption condition is satisfied by all the points in the box which are at a point of tangency between A's and B's indifference curves (because at all such points, the two individuals' marginal rates of substitution of y for x are equal). Two such points of tangency are shown in the above diagram. If we trace out all of them we get a continuous line from the bottom left-hand corner to the top right-hand corner of Edgeworth's box known as the efficiency locus. All the allocations represented by the infinite number of points on the efficiency locus satisfy the efficient consumption condition ! Furthermore, there are an infinite number of possible Edgeworth boxes like the one above, each corresponding to different amounts of x and y in the economy. Not only can the efficient consumption condition alone not tell us which point on an efficiency locus is best; it cannot even tell us which Edgeworth box we should be looking at !

The concept of an 'efficiency locus' in Edgeworth box diagrams makes it clear that there are many allocatively efficient combinations of u^A and u^B which are not equitable or socially just. For example, the point in the bottom left-hand corner of the Edgeworth box represents an allocatively efficient allocation in which we give everything to B and nothing to A. By any reasonable standard this is an unjust allocation, but it is nevertheless an allocatively efficient one because taking something away from B and giving it to A will inevitably make B worse off. We can make this trade-off between A's and B's utilities along the efficiency locus clearer by plotting the combinations of utilities along the efficiency locus in 'utility space' (in the above Edgeworth box, they are plotted in 'product space'). The result is called a utility frontier (for given x and y):



The utility frontier shows the maximum utility that B can attain along the efficiency locus for any given utility level for A (and for given amounts of x and y in the economy). The vertical intercept of the utility frontier represents the situation in which we give everything to B and nothing to A. It is impossible to make A better off (i.e. increase u^A) without making B worse off (i.e. decreasing u^B), so this patently unjust allocation is nevertheless an allocatively efficient one.

4.2. Efficient production

This condition is mathematically identical to the condition for efficient consumption. Essentially, it says that it must not be possible to increase the production of one commodity, without reducing the production of the other commodity. Recall that the marginal rate of substitution of capital for labour in the production of x can be written in terms of marginal products as

$$MRS^x{}_{KL} = -\frac{MP_L^x}{MP_K^x}$$

A similar expression can be written for y. What this tells you is how many units of capital are required to keep the output of x constant after a one unit reduction in labour. For example, if $MRS^x_{KL} = -1/2$, then an extra unit of capital increases the output of x by twice as much as an extra unit of labour, so that 1/2 units of capital are required to keep output constant after a one unit reduction in labour. In other words, labour is only half as productive as capital in the production of x. The optimality condition for efficient production, i.e.

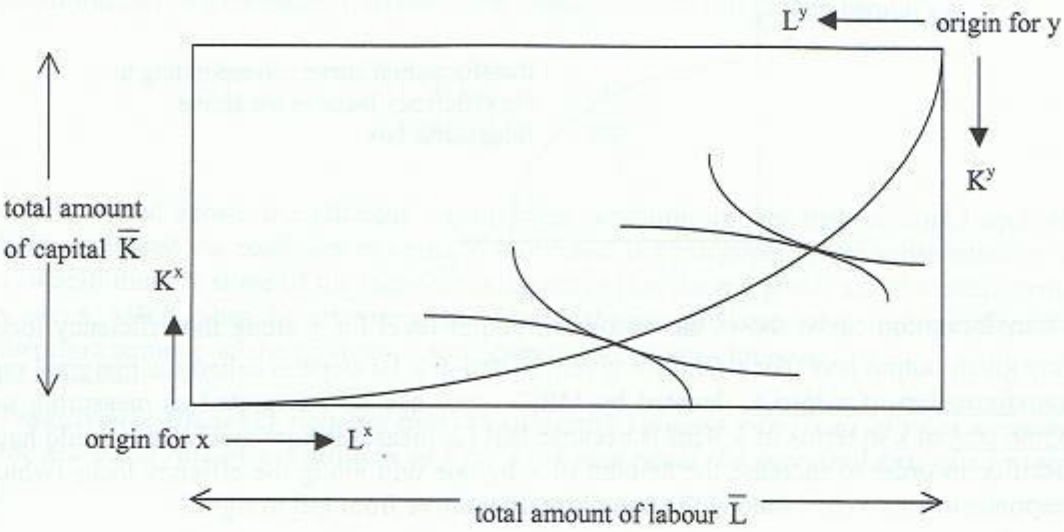
$$\left(\frac{MP_L^x}{MP_K^x}\right) = \left(\frac{MP_L^y}{MP_K^y}\right)$$

requires that the marginal rate of substitution between factors be the same in all industries. If this condition is not satisfied, then it must be possible to reallocate the factors of production in such a way that more of one commodity is produced, without reducing the production of the other commodity.

To illustrate this, suppose that labour and capital have different relative productivities in the different industries. Suppose that 2 units of capital are required to keep the output of x constant after a one unit reduction in labour, whereas only 1 unit of capital is required to keep the output of y constant after a one unit reduction in labour. Thus, labour and capital are equally productive in the y industry. If 1 unit of capital is reallocated from the production of x to the production of y, and one unit of labour is reallocated from the production of y to the production of x, then the output of y is unaffected. But the output of x must have increased, because the x industry gets a unit of labour which is twice as productive as the 1 unit of capital given to the y industry. Thus, we have increased production of x, without reducing the production of y. This cannot happen when the marginal rate of substitution between factors is the same in all industries. Therefore:

Efficient production requires that the marginal rate of substitution between factors be the same in all industries.

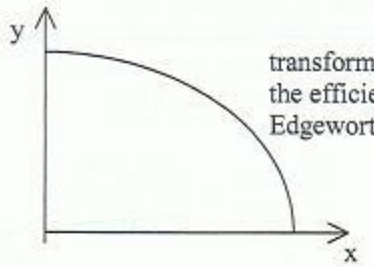
Again, the efficient production condition is necessary but not sufficient for a social optimum. For any given endowments of capital $\overline{K}$ and labour $\overline{L}$ in the economy, we can always find an infinite number of ways of allocating the factors between the x and y industries such that the efficient production condition is satisfied. On its own, it cannot tell us which is the best allocation of capital and labour between the x and y industries (so it cannot tell us how much of each of the goods x and y we should produce). All we can say is that if the efficient production condition is not satisfied, then we can make more of one good without sacrificing any of the other good. Again, we can use Edgewoth's box to illustrate the fact that an infinite number of allocations of capital and labour between the x and y industries will satisfy the efficient production condition.



The dimensions of the box are now the total endowment of labour $\overline{L}$ in the economy, measured along the horizontal axis, and the total endowment of capital $\overline{K}$, measured along the vertical axis. We make the bottom left-hand corner of the box the 'origin' of measurement for the x industry's allocations of captial K^x and labour L^x, and the top right-hand corner the 'origin' for the y industry's allocations K^y and L^y. Any point in the box then represents a division of the total $\overline{K}$ into K^x and K^y, and likewise for $\overline{L}$. With each point, there is therefore an associated production level for x given by $x = x(K^x, L^x)$, and a production level for y given by $y = y(K^y, L^y)$. We can draw isoquant curves for x and y, those for x emanating from the bottom left-hand corner, and those for y emanating from the top right.

As was the case with the efficient consumption condition, the efficient production condition is satisfied by all the points in the box which are at a point of tangency between the isoquants for x and the isoquants for y (because at all such points, the marginal rates of substitution of capital for labour in the two industries are equal). Two such points of tangency are shown in the above diagram. If we trace out all of them we get an efficiency locus again. All the allocations represented by the infinite number of points on the efficiency locus satisfy the efficient production condition. However, whereas there were an infinite number of possible Edgeworth boxes in the efficient consumption case (each corresponding to different amounts of x and y in the economy), there is only one Edgeworth box in this case, because the total endowments of capital $\overline{K}$ and labour $\overline{L}$ are assumed to be fixed at the outset. So in the case of the efficient production condition, we know which Edgeworth box we should be looking at, but the efficient production condition alone cannot tell us which point on the efficiency locus is best.

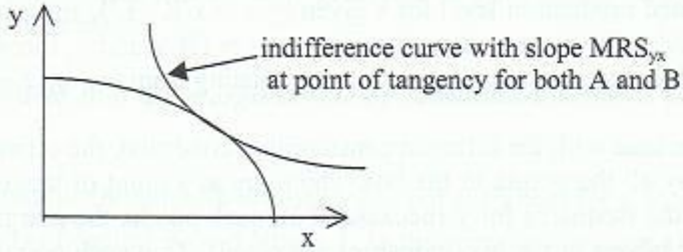
As before, we can make the trade-off between output levels of x and y along the efficiency locus clearer by plotting the combinations of outputs along the efficiency locus in 'product space' (in the above Edgeworth box, they are plotted in 'factor space'). The result is called a transformation curve (for given $\overline{K}$ and $\overline{L}$):



The transformation curve shows the maximum output level for y along the efficiency locus for any given output level for x (and for given $\overline{K}$ and $\overline{L}$). Its slope is called the marginal rate of transformation of y into x, denoted by MRT_{yx}, and can be interpreted as measuring the marginal cost of x in terms of y. This is because MRT_{yx} measures how much y we would have to sacrifice in order to increase the amount of x by one unit along the effiency locus (which corresponds to a movement along the transformation curve from left to right).

4.3. Efficient product-mix

Essentially, this says that for a social optimum we need to choose from the transformation curve the combination of x and y which maximises the individuals' utilities. This will be a point of tangency between one of each individual's indifference curves and the transformation curve:



Note that because of the efficient consumption requirement (which says that both individuals must have the same MRS_{yx} at the optimum), both individuals will achieve a utility maximum at the same point on the transformation curve i.e. both will agree about the total amount of x and the total amount of y which should be produced. Focusing on A, recall that the marginal rate of substitution of y for x can be written in terms of marginal utilities as $-(MU_x/MU_y)^A$. At the point of tangency with the transformation curve, this has to be equal to the marginal rate of transformation of y into x, MRT_{yx}. It can be shown that $MRT_{yx} = -(MP^y{}_K/MP^x{}_K)$. (This is because one way to increase x at the expense of y is to transfer 1 unit of capital from the y industry to the x industry. The amount of y lost per additional unit of x is $MP^y{}_K/MP^x{}_K$, which is the MRT_{yx}. Although it is not obvious by inspection, it can be shown that we get the same result when we consider shifting some of both of the factors between the industries, rather than just capital). Equating the two gives us the product-mix efficiency condition:

$$\left(\frac{MU_x}{MU_y}\right)^A = \left(\frac{MP_K^y}{MP_K^x}\right)$$

Because the efficient production condition must also be satisfied at the optimum, the above condition can be expressed equivalently in terms of labour rather than capital i.e.

$$\left(\frac{MU_x}{MU_y}\right)^A = \left(\frac{MP_L^y}{MP_L^x}\right)$$

and, as stated above, the efficient consumption condition implies that we could equivalently have expressed the condition in terms of individual B's marginal rate of substitution of y into x. Recall that the slope of the transformation curve (i.e. the marginal rate of transformation of y into x, MRT_{yx}) has the interpretation of being the marginal cost of x in terms of y. We can therefore summarise the product-mix efficiency condition as follows:

Product-mix efficiency requires that the subjective value of x in terms of y (as measured by the marginal rate of substitution of y for x) should equal the marginal cost of x in terms of y.

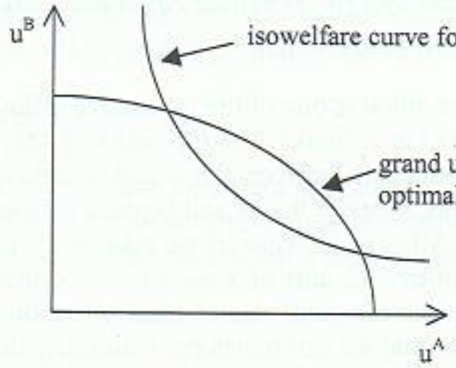
This product-mix efficiency condition is one of the 'allocative efficiency' conditions because, if it is not satisfied, it is possible to make one (or both) of the individuals better off by reallocating the factors of production between the x and y industries, without making the other worse off. To illustrate this, suppose that A and B place the same relative value on x and y, given by $MRS_{yx} = -(MU_x/MU_y) = -2$. Thus, if we take away 1 unit of y from each of A and B, we only have to give them 1/2 unit of x each to leave them as happy as they were before. However, suppose that the marginal rate of transformation of y into x is $MRT_{yx} = -(MP^y_K/MP^x_K) = -1$. This implies that we can transform 1 unit of y into 1 unit of x (i.e. to get 1 more unit of x we only have to give up 1 unit of y). The product-mix efficiency condition is violated since $(MU_x/MU_y) > (MP^y_K/MP^x_K)$. Suppose we take away 1 unit of y from each of A and B, and transform them into 2 units of x, giving 1 of the extra units of x to A and one of the extra units of x to B. This leaves them both better off than they were before, because they only needed 1/2 unit of x each to compensate them for the 1 unit loss of y they each suffered. Only when $(MU_x/MU_y) = (MP^y_K/MP^x_K)$, i.e. when the product-mix efficiency condition is satisfied, will it be impossible to achieve gains like this.

5. The fourth optimality condition: social justice

So far, we have only worried about allocative or Pareto efficiency. However, we have seen using the concept of a utility frontier that there are many Pareto efficient allocations that are not 'equitable' or 'socially just'. To find the Pareto efficient allocation that *does* yield a social optimum, we proceed as follows. First, recall that for each (x, y) combination on the transformation curve, we have an Edgeworth box with an associated efficiency locus and utility frontier. Plot on the same graph all the utility frontiers for all the (x, y) combinations on the transformation curve, to give a 'family' of utility frontiers. The utility frontier for the economy as a whole (called the 'grand utility frontier' of Pareto optimal points) is the outer envelope of all these frontiers. This is a curved frontier which looks just like the utility frontier we met earlier, but it has to be interpreted as the 'outer envelope' of all possible utility frontiers in the economy. It can be shown that the slope of this grand utility frontier at any point is

$$- \frac{(MU_x)^B}{(MU_x)^A}$$

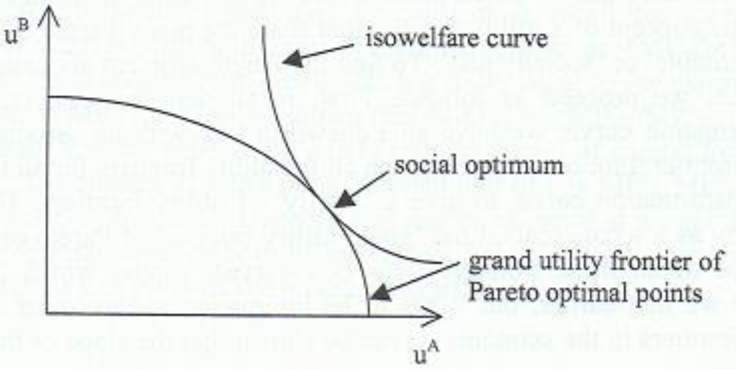
(This is because every point on the grand utility frontier is actually a point on an underlying utility frontier corresponding to some particular (x, y) output mix. Suppose that, with these given total values of x and y, we transfer one unit of x from A to B. The utility of B goes up by $(MU_x)^B$ and that of A goes down by $(MU_x)^A$, so the rate at which utility is being transferred between the individuals i.e. the slope of the grand utility frontier is $-(MU_x)^B/(MU_x)^A$). We can also plot a 'social indifference curve' (called an isowelfare curve) in the same diagram, by fixing the social welfare function $W = W(u^A, u^B)$ at a particular value, say $W = W^*$, and plotting all the combinations of u^A and u^B which yield a level of social welfare equal to W*:



By totally differentiating $W = W^*$, it can be shown that the slope of the isowelfare curve is

$$- \frac{MW_u^A}{MW_u^B}$$

Finding the socially optimal combination of u^A and u^B amounts to finding the highest isowelfare curve that is still 'touching' the grand utility frontier:



At this optimum, the slope of the isowelfare curve equals the slope of the utility frontier. This gives us the fourth optimality condition (social justice):

$$\frac{(MU_x)^B}{(MU_x)^A} = \frac{MW_u^A}{MW_u^B}$$

Rearranging this equation gives

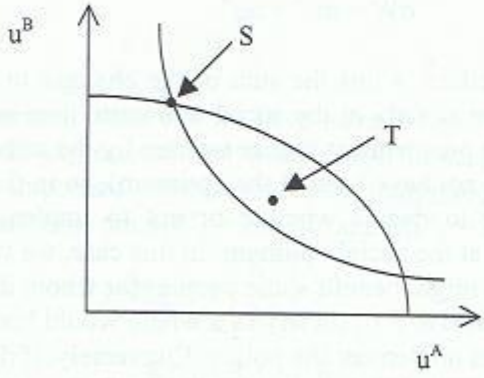
$$(MU_x)^A \cdot MW^A{}_u = (MU_x)^B \cdot MW^B{}_u$$

Essentially this says that the social value of giving an extra unit of x to A must be the same as the social value of giving it to B i.e. we should not be able to increase social welfare by reallocating x between A and B. Giving an extra unit of x to A increases A's utility level by an amount equal to $(MU_x)^A$. Each unit increase in A's utility increases the level of social welfare by an amount $MW^A{}_u$. Thus, $(MU_x)^A \cdot MW^A{}_u$ is the total increase in social welfare produced by giving an extra unit of x to A. Similarly, $(MU_x)^B \cdot MW^B{}_u$ is the total increase in social welfare produced by giving an extra unit of x to B. If we take one unit of x away from B and give it to A, A's utility level will rise, producing an increase in social welfare equal to $(MU_x)^A \cdot MW^A{}_u$, but B's utility will fall, producing a fall in social welfare whose magnitude is $(MU_x)^B \cdot MW^B{}_u$. As long as the above equation holds, society as a whole will have gained nothing by such a reallocation. Note that the condition also applies to y i.e.

$$(MU_y)^A \cdot MW^A{}_u = (MU_y)^B \cdot MW^B{}_u$$

Once we have found the socially optimal point, we have solved our economic problem. We have a unique allocation of factors between goods (K^x, K^y, L^x, L^y) - this solves the problem of what shall be produced and how. We also have a unique allocation of goods between people (x^A, x^B, y^A, y^B). This solves the problem of for whom.

There is one other crucial point. Consider the allocations S and T in the following diagram:



Suppose we are for some reason prevented from producing the ideal output (at a point of tangency between an isowelfare curve and the grand utility frontier of Pareto optimal points). There are many allocatively inefficient points (like T) which are preferable to allocatively efficient points (like S) in the sense that they are on a higher isowelfare curve. Thus, allocative efficiency is only unambiguously 'desirable' if equity is also fully satisfied, in other words, if the distribution of income is right. It is important to emphasise this point because the terminology 'allocative efficiency' seems to carry with it the implication that it is inherently better than 'allocative inefficiency'. The above clearly shows that this is not true.

Now, assessing the desirability of particular policies typically involves comparing different economic states in terms of the social welfare gains and costs associated with those states. For example, if one is doing a cost benefit analysis of a public motorway project, one wants to compare social welfare in two 'states of the world': state 0, where the motorway is not built, and state 1, where the motorway is built and the money to build it is raised in some specified way. This is called *cost benefit analysis* because in comparing state 0 with state 1, we want to assess the additional benefits generated by the project, minus the alternative benefits which would have been available but have been lost due to the project (i.e. the 'opportunity cost' of the project). The action of shifting from state 0 to state 1 should be judged by its effects on the happiness of all those affected. One can imagine two main possible types of outcome:

1. Someone gains and no one loses (or no one gains and someone loses).
2. Someone gains but someone else loses.

We do not really need a social welfare function in case 1, because we can tell straight away that we should implement a policy if it benefits some people, while harming no one. Such a policy is said to yield a Pareto improvement. Conversely, we do not need a social welfare function to tell us that it would be silly to implement a policy that does not do anyone any good, and actually harms some people ! Unfortunately, in the real world we are most often confronted with case 2. Then we need to invoke our social welfare function to tell us whether or not society as a whole would benefit from a particular policy, even if it would involve hurting some people.

When economists talk about 'welfare gains and costs' associated with a particular policy, they are more often than not making the assumption

$$W = W(u^A, u^B) = u^A + u^B$$

i.e. that total welfare is just the sum of individual utilities. Then we get a simple formula which we can use in cost-benefit analysis: letting dW denote the change in social welfare as a result of a policy, and du^A and du^B the changes in individual utility, we get

$$dW = du^A + du^B$$

i.e. the change in social welfare is just the sum of the changes in individual utilities across society. Of course, if we are already at the social optimum, then any policy which increases welfare by benefitting some people must reduce welfare by the same amount by hurting other people (otherwise we could not have been at the optimum), so in this case $dW = 0$. But in the real world, we often have to decide whether or not to implement particular policies in situations where we are not at the social optimum. In this case, we would get either $dW > 0$ or $dW < 0$. A particular policy might benefit some people (for whom $du > 0$) and hurt others (for whom $du < 0$), but as long as $dW > 0$, society as a whole would benefit, and we might decide on this basis to go ahead and implement the policy. Conversely, if $dW < 0$, we might abandon the policy, even if some individuals in society would benefit from it. (Note that this process of adding up does require us to be able to express utility levels across society in terms of a single objective unit of measurement, say £s. For simplicity, we will assume that we can do this, but it must be borne in mind that this can be problematic e.g. if not everyone values a £ equally).

6. The market system

So far, we have said nothing about the system of economic organisation. We have simply described the technical conditions (three for efficiency and one for equity) that must hold at the optimum. In this section, we extend our mathematical model to consider the circumstances under which a freely functioning capitalist economy (i.e. an economy with no government intervention whatsoever) will yield the socially optimal state. We will find that the 'free market' performs extremely well, provided that we make four sweeping assumptions. It is the failure of one or more of these assumptions that makes certain forms of government intervention necessary (we shall discuss this in more detail in Lecture 2). The key assumption is that in every market there are so many buyers and sellers, that all of them behave as though their individual actions cannot affect prices. In other words, the key assumption is that we have perfect competition. To investigate this, we extend our model by assuming that we have many consumers and many producers. First, we briefly discuss the concept of 'constant returns to scale', which will play an important role.

6.1. Returns to scale

Consider the production function for good x:

$$x = x(K^x, L^x)$$

This production function is said to exhibit constant returns to scale if multiplying each input K^x and L^x by a factor λ causes the output of x to increase by the same factor i.e.

$$x(\lambda K^x, \lambda L^x) = \lambda x$$

Thus, for example, doubling each input causes output to be doubled. If we set $\lambda = 1/L^x$, constant returns to scale therefore implies that

$$\frac{x}{L^x} = x\left(\frac{K^x}{L^x}, 1\right)$$

The right hand side is a function of the ratio of K^x to L^x only, because the number 1 is a constant. Thus, we can forget about the number 1 and rewrite the above equation as

$$x = L^x f\left(\frac{K^x}{L^x}\right)$$

where f(·) denotes a function. Note that we have taken L^x over to the right hand side. If we partially differentiate this expression, first with respect to capital, and then with respect to labour, we get two very important expressions for the marginal products of capital and labour in producing x (recall that these are $MP^x_K = \partial x/\partial K^x$ and $MP^x_L = \partial x/\partial L^x$):

$$MP^x_K = f'\left(\frac{K^x}{L^x}\right)$$

$$MP^x_L = f\left(\frac{K^x}{L^x}\right) - \frac{K^x}{L^x} f'\left(\frac{K^x}{L^x}\right)$$

(The symbol $f'(\,)$ means the derivative of the function f with respect to the ratio K^x/L^x). These expressions are very important because they tell us that, with constant returns to scale, the marginal products of capital and labour depend only on the capital-labour ratio, not on the levels of K^x and L^x individually. In our model with many producers, this will allow us to express the required optimality conditions in terms of the marginal products for any single producer, which is a great simplification.

As we shall see, problems arise for free markets when there are increasing returns to scale. In this case, multiplying the inputs by λ raises output by more than λx i.e.

$$x(\lambda K^x, \lambda L^x) > \lambda x$$

Increasing returns to scale give rise to monopoly producers, and therefore to allocations which are not efficient. Decreasing returns to scale occur when multiplying each input by λ causes output to be multiplied by a factor which is less than λ i.e.

$$x(\lambda K^x, \lambda L^x) < \lambda x$$

1.2. A freely functioning capitalist economy

To see how a freely functioning capitalist economy works, let us start with consumption. Recall that the optimality condition for efficient consumption says that all consumers must place the same relative values on x and y. We expressed this as

(1). Efficient consumption: $$\left(\frac{MU_x}{MU_y}\right)^A = \left(\frac{MU_x}{MU_y}\right)^B$$

which says that $(MRS_{yx})^A = (MRS_{yx})^B$. Suppose in our new model that there are very many consumers of type A, each with identical tastes and the same factor endowments, and the same number of consumers of type B, likewise identical to each other. As we saw at the beginning of this lecture, each consumer i is maximising utility $u^i(x^i, y^i)$ subject to a budget constraint $p_x x^i + p_y y^i = m^i$, where we now interpret m^i as the money income derived from individual i's factor endowment, and p_x and p_y are the respective product prices. (Note: each consumer i is assumed to own a quantity K^i of capital, and a quantity L^i of labour. This is his/her 'factor endowment'. Money income is then computed as $m^i = w_K K^i + w_L L^i$, so that consumers can be viewed as deriving their income from selling their factor endowments to producers. This is discussed further below). For each consumer i, we saw that the optimality condition was

$$(MRS_{yx})^i = -\frac{p_x}{p_y}$$

But all the consumers in our economy face the same prices. Thus they will all set their consumption levels of x and y so that their marginal rate of substitution of y for x equals the price ratio $-p_x/p_y$. Thus, the price mechanism ensures that the relative values of x and y get equated among all citizens. Accordingly, the price mechanism ensures efficient consumption, as we have expressed it above.

Next, let us consider production. Recall that the optimality condition for efficient production says that the relative productivities of capital and labour must be the same in all industries. We expressed this as

(2). Efficient production: $\left(\frac{MP_L^x}{MP_K^x}\right) = \left(\frac{MP_L^y}{MP_K^y}\right)$

which says that $(MRS^x_{KL}) = (MRS^y_{KL})$. Suppose in our new model that we have very many producers, some of them in the x industry, and the rest in the y industry. To keep the model simple, assume that each firm i in the x industry has a production function of the form $x(K^{ix}, L^{ix})$, and each firm i in the y industry has a production function of the form $y(K^{iy}, L^{iy})$, where K^{ix} denotes the amount of capital used to produce x by firm i, etc. Thus, within each industry, all firms have the same basic production function. As we saw earlier in this lecture, a firm i having the production function $x(K^{ix}, L^{ix})$ will always want to produce as much output as it can for any given money cost C^{ix}. If w_K and w_L are the respective factor prices (taken as given by all producers), then $w_K K^{ix} + w_L L^{ix} = C^{ix}$. We saw earlier that the optimality condition for each producer i in the x industry is of the form

$$(MRS^x_{KL})^i = -\frac{w_L}{w_K}$$

A similar condition applies to each producer i in the y industry:

$$(MRS^y_{KL})^i = -\frac{w_L}{w_K}$$

Thus, the price mechanism has in this case ensured that marginal rates of substitution are equated among all the individual firms in the two industries. However, we have not yet proved that the equation for efficient production above holds, because this equation is written in terms of marginal products as functions of total industry inputs K^x, L^x, K^y and L^y, whereas our expressions $(MRS^x_{KL})^i$ and $(MRS^y_{KL})^i$ are in terms of inputs for each individual firm i, namely K^{ix}, L^{ix}, K^{iy}, L^{iy}. We complete the proof by invoking constant returns to scale. Since in each industry all firms have the same basic production function (and therefore use the same technology), each firm will employ the same factor proportions i.e. the ratio K^{ix}/L^{ix} will be the same for all firms in the x industry, and K^{iy}/L^{iy} will be the same for all firms in the y industry. (Think of 'technology' as a recipe for baking a particular type of cake, which tells you in what proportions the ingredients must be combined in order to get that type of cake. Different firms in a particular industry may bake cakes of different sizes, and may therefore use different quantities of ingredients, but since they are all following the same recipe, they will all be combining the ingredients in the same proportions). Now, the total amount of capital used to produce x in the x industry is the sum of the quantities of capital used by all the individual firms in the x industry i.e. $K^x = \sum_i K^{ix}$. Similarly, $L^x = \sum_i L^{ix}$, $K^y = \sum_i K^{iy}$, and $L^y = \sum_i L^{iy}$. The crucial point is that if K^{ix}/L^{ix} is the same for all firms in the x industry, and K^{iy}/L^{iy} is the same for all firms in the y industry, then we must have

$$\frac{K^{ix}}{L^{ix}} = \frac{\sum_i K^{ix}}{\sum_i L^{ix}} = \frac{K^x}{L^x} \quad \text{and} \quad \frac{K^{iy}}{L^{iy}} = \frac{\sum_i K^{iy}}{\sum_i L^{iy}} = \frac{K^y}{L^y}$$

i.e. the capital-labour ratio for an individual firm in a particular industry must equal the capital-labour ratio for the industry as a whole. Using our formulas for the marginal products of capital and labour from the previous sub-section (which we derived under the assumption of constant returns to scale) we must then have

$$MP^{ix}{}_K = f'\left(\frac{K^{ix}}{L^{ix}}\right) = f'\left(\frac{K^x}{L^x}\right) = MP^x{}_K$$

$$\text{and} \quad MP^{ix}{}_L = f\left(\frac{K^{ix}}{L^{ix}}\right) - \frac{K^{ix}}{L^{ix}} f'\left(\frac{K^{ix}}{L^{ix}}\right) = f\left(\frac{K^x}{L^x}\right) - \frac{K^x}{L^x} f'\left(\frac{K^x}{L^x}\right) = MP^x{}_L$$

i.e. the marginal products of capital and labour for each individual firm must equal the marginal products of capital and labour for the industry as a whole. Similarly, in the y industry we have

$$MP^{iy}{}_K = MP^y{}_K \qquad \text{and} \qquad MP^{iy}{}_L = MP^y{}_L$$

Since the marginal rate of substitution of capital for labour in each industry is the ratio of the marginal products, we must have

$$(MRS^x{}_{KL})^i = -\left(\frac{MP^{ix}_L}{MP^{ix}_K}\right) = -\left(\frac{MP^x_L}{MP^x_K}\right)$$

for each firm i in the x industry, and

$$(MRS^y{}_{KL})^i = -\left(\frac{MP^{iy}_L}{MP^{iy}_K}\right) = -\left(\frac{MP^y_L}{MP^y_K}\right)$$

for each firm i in the y industry. Therefore the equalisation of marginal rates of substitution across all individual firms in the two industries (they all set the marginal rate of substitution equal to the factor price ratio $-w_L/w_K$) must mean that the relative productivities of labour and capital are being equalised across the two industries as well i.e. the price mechanism has ensured that the condition for efficient production above is satisfied.

Next, let us consider the efficient product-mix condition. The optimality condition for efficient product-mix says:

(3). <u>Efficient product-mix:</u> $\left(\frac{MU_x}{MU_y}\right)^A = \left(\frac{MP^y_K}{MP^x_K}\right)$

In our multi-person model, this requires that each individual's marginal rate of substitution of y for x should equal the relative productivity of capital in the y and x industries. Now, under the assumption of perfect competition, each profit-maximising firm in our model expands output until the <u>marginal cost</u> of output has risen to equal the price. The price of good x is p_x. But what is the marginal cost of x? Suppose a firm with a production function $x(K^{ix}, L^{ix})$ employs one more unit of capital. Its costs rise by w_K, and its output of x by $MP^{ix}{}_K$. Therefore the increase in the cost per unit of extra output (i.e. the marginal cost) is just $w_K/MP^{ix}{}_K$. So each individual firm will produce x until $p_x = w_K/MP^{ix}_K$. But under the assumption of constant returns to scale, we have $MP^{ix}_K = MP^x_K$ (see above). Therefore for the x industry as a whole we have

$$p_x = \left(\frac{w_K}{MP^x_K}\right)$$

Using exactly the same argument, we get the following equation for the y industry:

$$p_y = \left(\frac{w_K}{MP^y_K}\right)$$

Dividing the first equation by the second equation gives us

$$\frac{p_x}{p_y} = \left(\frac{MP^y_K}{MP^x_K}\right)$$

but this is the same as the efficient product-mix condition above, because p_x/p_y is just the negative of the marginal rate of substitution of y for x, which is the same for all individuals under the assumption of efficient consumption. Thus the price mechanism ensures an efficient product-mix.

We have seen so far, then, that the freely functioning market mechanism ensures that the three conditions for Pareto efficiency in the economy are satisfied. But what about social justice? It turns out that the market mechanism will also 'automatically' lead to a maximum of the social welfare function if the ownership of the factors of production is correctly distributed across society. Let us consider this in detail. Each consumer i must be allocated the correct quantity K^i of capital, and the correct quantity L^i of labour, to ensure that he/she will be able to buy the consumption bundle which corresponds to the welfare-maximising configuration of the economy. This is his/her 'factor endowment'. Money income is then computed as $m^i = w_K K^i + w_L L^i$ so that in the market model each consumer i is maximising utility $u^i(x^i, y^i)$ subject to a budget constraint $p_x x^i + p_y y^i = m^i$, where we now interpret m^i as the money income derived from individual i's factor endowment. To start with, consider all individuals of type A as a single group, and all individuals of type B as a single group. From the solution of society's economic problem in Section 5, we know that the optimal allocation of goods to group A is (x^A, y^A), and the optimal allocation of goods to group B is (x^B, y^B). In the market system, group A's socially optimal expenditure on x and y will therefore be $p_x x^A + p_y y^A = m^A$. If, by choosing A's endowment of capital K^A and labour L^A, we correctly set $m^A = w_K K^A + w_L L^A$, the market will ensure that group A will consume x^A and y^A. Similarly, if we set group B's endowment of capital equal to $K^B = \overline{K} - K^A$ and group B's endowment of labour equal to $L^B = \overline{L} - L^A$, then B's income will be $m^B = w_K K^B + w_L L^B$ and the market will ensure that group B will consume the socially optimal allocation (x^B, y^B) at a cost of $p_x x^B + p_y y^B = m^B$. Since all individuals of type A are the same, we can then set each individual's endowment of capital and labour in group A equal to $K^{iA} = K^A/N^A$ and $L^{iA} = L^A/N^A$ respectively, where N^A is the total number of individuals in group A. Similarly, we can set each individual's endowment in group B equal to $K^{iB} = K^B/N^B$ and $L^{iB} = L^B/N^B$, where N^B is the total number of individuals in group B. If we could somehow 'arrange' for these to be the factor endowments across society, then the 'free market' would automatically lead to the same socially optimal state as the one we achieved in Section 5. The concept of 'distributional justice' is characterised in our model as the 'correct' distribution of factor endowments, in the sense that this is the distribution that will take the free market to a social welfare maximum.

There is one more loose end that we need to tie up: where do the prices p_x, p_y, w_K and w_L come from ? We made no mention of prices in solving society's economic problem in Section 5, but they played a crucial role in the functioning of the 'free market' economy in this section. In the context of the free market model, we imagine that the interactions of many buyers and many sellers of goods give rise to 'market forces' which push prices towards their equlibrium values. For example, if p_x is too high, the market supply of good x will exceed the market demand for good x, putting pressure on firms to reduce the price in order to clear their stocks of good x. If p_x is too low, there will be excess demand for good x, and consumers will tend to 'bid up' the price. Only when the price is such that demand = supply will there be no pressure for p_x to change. A similar story applies to p_y, w_K and w_L. However, we can answer the question 'where do prices come from' at a more fundamental level by referring to our model of society's economic problem in Section 5. Although we did not mention prices when solving this problem, the prices were 'implied' by the constraints the economy faced (factor endowments, technology and tastes). (In mathematical jargon, the 'free market prices' correspond to the 'shadow prices' given by the values of the Lagrange multipliers in the solution of society's constrained optimisation problem). In fact, once we have solved society's constrained optimisation problem in Section 5, we can always calculate a set of prices p_x, p_y, w_K and w_L that would be 'market clearing' or 'equilibrium' prices in the context of the corresponding free market model. All we need to know are the marginal rate of substitution of y for x at the optimum, and the marginal products of capital and labour at the optimum. If we let p_y be any number, and then define $p_x = -(MRS_{yx}) \times p_y$, $w_K = (MP^y{}_K) \times p_y$ and $w_L = (MP^y{}_L) \times p_y$, the resulting set of prices p_x, p_y, w_K and w_L will be 'equilibrium' prices in the corresponding free market system (i.e. they will be such that demand = supply in each market). These equations are derived directly from the utility maximising behaviour of consumers and the profit maximising behaviour of producers. Utility maximisation implies $-p_x/p_y = MRS_{yx}$ which rearranges to give the equation for p_x above. The price = marginal cost condition for profit maximisation implies $p_y = w_K/MP^y{}_K$ and $p_y = w_L/MP^y{}_L$, which can be rearranged to give the equations for w_K and w_L above. Thus, we might say that market prices reflect the resource constraints our society faces.

1.3. A note on the relationship between market price and opportunity cost

When analysing public policies relating to the provision of goods and services, it is typically assumed that utility increases as the quantity of goods and services made available for consumption increases. Thus, additional housing, better transport, improvements in the quality of education or health-care etc. will all increase utility. But consumption in this context is limited by the fact that the resources of land, labour and capital which are used to produce these commodities are not available in unlimited supply, so we cannot hope to produce a sufficient quantity to satisfy all our wants. The scarcity of resources in relation to the demands made on them leads inevitably to the need to make a series of choices about the quantities of different commodities that are to be produced.

Production of a particular commodity will therefore have costs in the form of other goods that could have been produced with the resources it uses up. For example, labour time or machinery that is devoted to car production is obviously not available for providing hospital facilities or schools. Because it is possible to look at the costs of production in terms of alternative goods and services that could have been produced, economists have devised the term *opportunity cost*. Thus, the opportunity cost of producing cars is the forgone opportunity of providing other commodities such as hospital facilities.

In many situations, the market price that is paid for the use of a resource reflects its opportunity cost. Thus, we may expect the price a hospital has to pay for the services of plumbers or carpenters to be the same as they would receive when employed in the construction of new housing. Therefore the price paid by the hospital measures the value of plumbers' or carpenters' services, not only to the hospital, but also in their alternative use in the housebuilding industry. But market prices do not always measure opportunity costs. In situations where society's resources are not being utilised fully, the decision to employ an additional resource in, say, the provision of hospital services will not necessarily imply a reduction in the availability of resources for alternative uses. So if there is unemployment of labour, or if machines are standing idle, the decision to put them to work may incur a zero opportunity cost even though the market price for their services will almost certainly be positive. In any discussion of the costs of resources used in production, the concept of opportunity cost should always be borne in mind.

6.4. Marginal social benefit and marginal social cost

It will be useful in future discussions to be able to characterise the 'allocatively efficient' level of production of each good as the level <u>which equates the marginal social benefit with the marginal social cost</u>. We will now show that this is equivalent to the three conditions for efficiency discussed above in the context of a simple social welfare function which is just the sum of individual utilities in society. With many individuals of type A, and many of type B, we then have

$$W = \sum_i u^{iA} + \sum_i u^{iB}$$

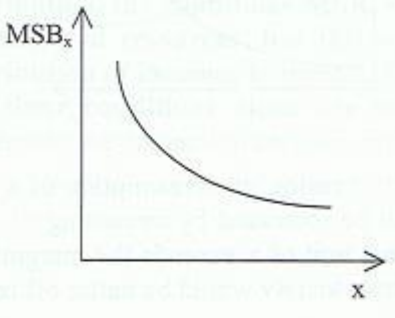
The <u>marginal social benefit</u> of an additional unit of good x, denoted by MSB_x (i.e. $MSB_x = \partial W/\partial x$) is then

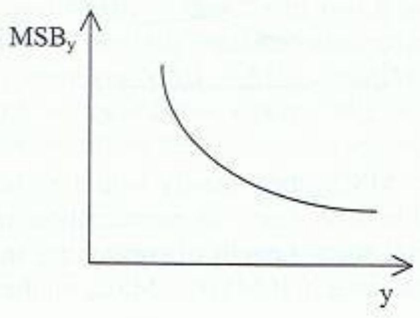
$$MSB_x = \sum_i (MU_x)^{iA} + \sum_i (MU_x)^{iB}$$

i.e. just the sum of the individual marginal utilities in society. Similarly, the marginal social benefit of an additional unit of good y is

$$MSB_y = \sum_i (MU_y)^{iA} + \sum_i (MU_y)^{iB}$$

Since each individual's marginal utility of consumption of a good decreases as the quantity consumed increases, it must be the case that MSB_x decreases as the quantity of x increases, and similarly MSB_y decreases as the quantity of y increases. In other words, the social welfare function exhibits a diminishing marginal social benefit (just as an individual utility function exhibits a diminishing marginal utility), and MSB_x and MSB_y must be downward-sloping curves:





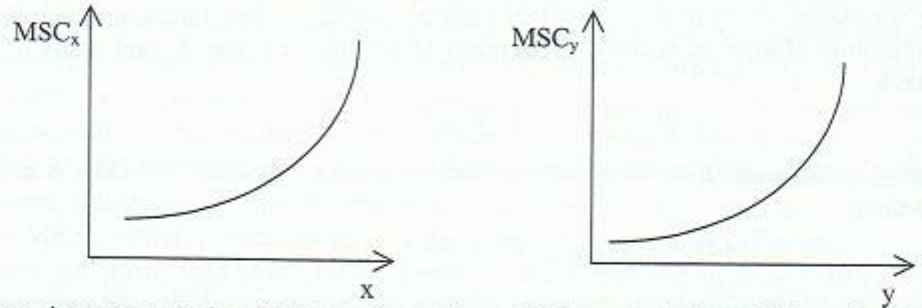
Now let us turn our attention to production. We said earlier that, in the x industry, the marginal cost is given by w_K/MP_K^x i.e. this is how much it costs to produce one more unit of x in the x industry as a whole (note that this can be expressed equivalently in terms of labour: w_L/MP_L^x. The two must be equal because of the first-order condition for profit maximisation). The term w_K/MP_K^x is now interpreted as the <u>marginal social cost</u> of producing an additional unit of x, which we shall denote by MSC_x. Thus,

$$MSC_x = \left(\frac{w_K}{MP_K^x}\right)$$

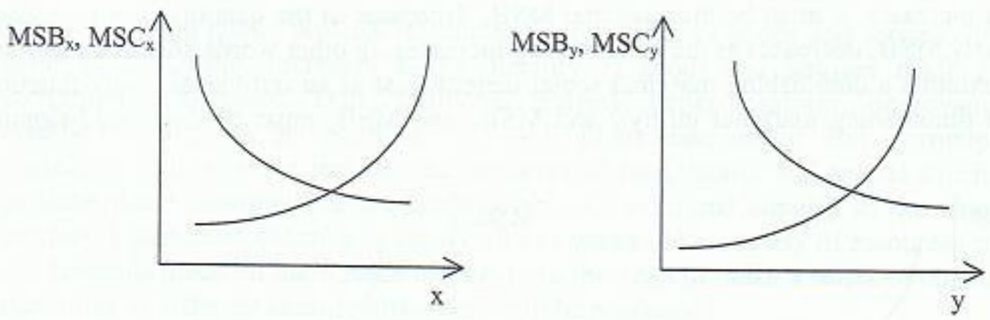
Similarly, the term w_K/MP_K^y is the marginal social cost of producing y, denoted by MSC_y. Thus,

$$MSC_y = \left(\frac{w_K}{MP_K^y}\right)$$

These curves must be upward-sloping because of the diminishing marginal productivity property of production functions. As the amount of capital rises in the x industry (to increase the level of production of x) the marginal product MP_K^x falls, and with constant w_K, MSC_x must rise. Similarly, as more y is produced, MSC_y must rise. Thus, we get the following upward-sloping marginal social cost curves:



Now, society as a whole will demand good x until the marginal social benefit of an additional unit of good x equals the marginal social cost. This occurs where the MSB_x and MSC_x curves intersect:



If $MSB_x > MSC_x$, then society would be better off increasing its consumption of x (i.e. the 'net social benefit' from the consumption of x would be increased by increasing x), because the marginal social benefit of consuming an additional unit of x exceeds the marginal social cost of producing it. If $MSB_x < MSC_x$ on the other hand, society would be better off reducing its consumption of x (i.e. the 'net social benefit' from consuming x would be increased by reducing x). Only when $MSB_x = MSC_x$ is the 'net social benefit' from consuming x maximised: it is impossible to increase it any further by adjusting the amount of x up or down. At this point, 'equilibrium' is reached in the market for x. By the same argument, society will consume good y until $MSB_y = MSC_y$, and this is the level of y which maximises the 'net social benefit' from consuming y. It is the point at which equilibrium is reached in the market for y.

We have said that if $MSB_x = MSC_x$ and $MSB_y = MSC_y$ then net social benefit is maximised in every market. But then it must be the case that the efficient consumption condition of welfare maximisation is satisfied (i.e. everyone's marginal rate of substitution of y for x must be the same), otherwise we would be able to make someone better off by adjusting x and y without making anyone else worse off (see the discussion of the efficient consumption condition above). This would contradict the assertion that net social benefit has been *maximised*. It must also be the case that the efficient production condition is satisfied, otherwise we could increase the amount of one of the goods without decreasing the amount of the other, again contradicting the assertion that net social benefit has been maximised. Finally, it must be the case that the efficient product-mix condition is satisfied. This is because if $MSB_x = MSC_x$ and $MSB_y = MSC_y$, then it is also true that $MSB_x/MSB_y = MSC_x/MSC_y = MP_K^y/MP_K^x$. But because of the efficient consumption condition, each individual's marginal rate of substitution of y for x is equal to MSB_x/MSB_y, so each individual's marginal rate of substitution of y for x must be equal to MP_K^y/MP_K^x. This is the condition for product-mix efficiency which we discussed earlier.

We have just shown that the conditions $MSB_x = MSC_x$ and $MSB_y = MSC_y$ are equivalent to the three efficiency conditions we discussed earlier. From now on, we will say that an allocatively efficient level of production of a commodity is one where the marginal social benefit from producing the commodity equals the marginal social cost.

Note that the above diagrams look very similar to the demand-supply diagrams for goods x and y that we 'derived' from individual behaviour at the beginning of this lecture. This is no accident: THEY ARE THE SAME ! We have just 'derived' them in a different way here. An essential part of the theory underlying the market system is that, in certain circumstances (i.e. the four sweeping assumptions mentioned at the start of this section), the demand curve is an alternative representation of the marginal social benefit curve, and the supply curve is an alternative form of the marginal social cost curve. Under those circumstances, the equilibrium output obtained through the market system by the interaction of demand and supply will be an efficient level of output. Note, however, that it will not be the socially optimal allocation (i.e. the one that maximises social welfare) unless the distribution of incomes is correct. We still need this fourth 'social justice' condition to be satisfied in order to achieve a social welfare maximum. **This is a very important point to keep in mind: for a given income distribution, the conditions $MSB_x = MSC_x$ and $MSB_y = MSC_y$ will result in an efficient allocation of resources, but this will not be the socially optimal allocation unless the distribution of incomes is correct. Efficiency requires $MSB_x = MSC_x$ and $MSB_y = MSC_y$, but these conditions alone are not enough to guarantee equity**. We will constantly emphasise the distinction between efficiency and equity in the discussions which follow.

2. Society's objectives: efficiency and social justice in the context of health-care

We can now clearly see that two important aims have to be considered in any question of resource allocation: the attainment of efficiency and the promotion of social justice or equity. Let us see how these can be interpreted in the context of health-care.

7.1. Efficiency

It might seem a bit unethical to discuss costs and benefits in relation to people's health. Usually, one wants to consider only the benefits, and forget about the costs. But costs cannot be ignored in making policy choices. The building of hospitals, the training of doctors and nurses, the manufacture of drugs and technical equipment, all consume scarce resources. That is, they use up land, labour and capital that could have been put to other uses, such as building schools, training teachers or building cars. The 'best possible' standard of health-care could only be achieved by devoting all of the economy's resources to it. This can hardly be a wise course of action, since no other commodities (including those vital to health, such as food) could be produced. Doctors who give their patients 'Rolls-Royce' treatment are likely to be taking resources from other patients. Keeping patients alive as long as possible cannot be a target for the allocation of resources, even within the health-care sector, because the survival prospects of one patient can usually only be increased at the expense of the survival prospects of others.

What is needed is a definition of *efficiency* that takes account of the costs of, as well as the benefits from, health-care. An obvious starting point is the definition suggested earlier: an efficient level of production of a commodity was defined as one where the marginal social benefit (MSB) equals the marginal social cost (MSC). A simple example shows how this might be applied in the field of health-care.

Suppose we are trying to decide what is the efficient number of hospital beds to provide in a particular town. Imagine that we can measure the benefits and costs of providing hospital beds in terms of money (how this might be done is discussed below). Looking at these benefits and costs, we find that to provide 1000 hospital beds would yield a social benefit of £20 million, and would cost only £10 million. To provide a second 1000 beds would benefit the town rather less than the first: say, by £12 million. It would cost slightly more: say, also £12 million. To provide a third 1000 beds would cost yet more - £16 million - and create benefits of only £8 million. Thus, the MSB from providing each 1000 beds declines, while the MSC increases. This is as we would expect, given dimishing marginal utilities and marginal products. Once the major needs of an area have been satisfied, the benefits from providing more and more hospital beds are likely to diminish, while with each new hospital built, the resources available for building yet another one become increasingly scarce and, therefore, more expensive.

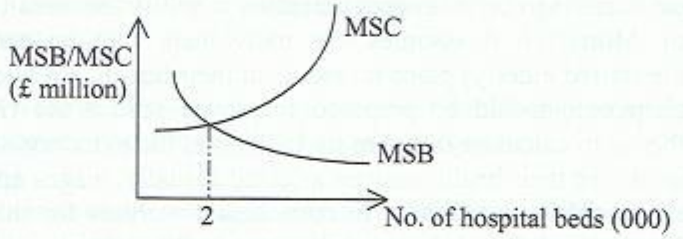
These figures, together with their equivalent for fourth and fifth thousand beds, are summarised in Table 1 below:

Table 1. Social costs and benefits of hospital beds.

Number of hospital beds (000)	Marginal social benefit (£million)	Marginal social cost (£million)
1	20	10
2	12	12
3	8	16
4	6	22
5	5	30

From this table, we can deduce what would be the efficient number of hospital beds to provide. The gain (the MSB) from providing the first 1000 beds is £10 million greater than the cost (the MSC). So those beds are worth providing. The MSB from the second 1000 beds equals the MSC. Hence, they are also (just about) worth providing. However, the costs of the third, fourth and fifth thousand beds are greater than the benefits, and so it would not be efficient to provide them. Therefore the efficient number of beds is two thousand, the point at which the MSC of providing more beds begins to exceed the MSB.

We can make the same point by use of a diagram. In the diagram below, the downward sloping curve shows how the MSB declines as the number of beds provided increases, while the upward sloping curve shows how the MSC increases. The point at which the two curves intersect is the efficient number of hospital beds.



This is an example of how the concept of efficiency outlined in the previous section could be applied in the field of health-care. It is, in fact, perfectly general. In principle, it could be applied to any problem of allocating resources in the area (such as that of determining the efficient number of doctors to train, or of kidney machines to provide). But to use the concept in practice is not so easy. Difficulties arise in measuring the quantities involved, particularly the benefits. As this is a particularly active and interesting area of current research in health economics, I will expand on it in Section 8.

6.2. Social justice

Most people agree that the health care system should be fair or equitable. However, there is much less agreement about what 'fair' or 'equitable' means ! This is because in real life we rarely know for sure what the 'true' socal welfare function looks like. In order to operationalise the welfare economic theory we have spent so much time discussing in this lecture, we need to adopt some view about what we mean by the term 'socially just'.

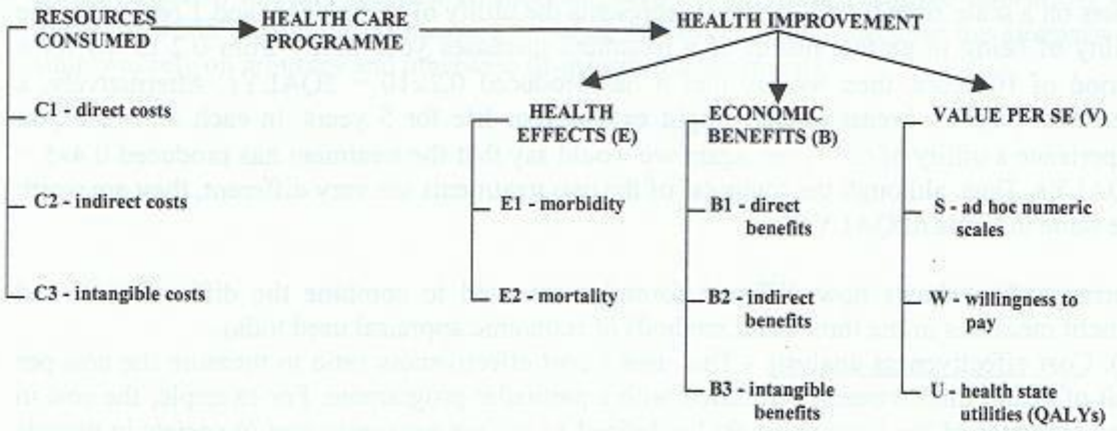
Two particular definitions of social justice have figured prominently in policy discussions about the equitable allocation of health care: one is in terms of a *minimum standard* i.e. there should always be at least a minimum standard of treatment for those who need it; the other is in terms of *full equality* i.e. there should be an equal standard of treatment for all those who are equally in need.

A third interpretation of equity found in discussions concerning the organisation of health care is that it should promote *equality of access*. What precisely is meant by this is rarely specified, but one interpretation is to define it in terms of the costs or sacrifices that people have had to make to get medical care. These include any fees or charges that may be levied, any money lost through having to take time off work (the wage rate is often assumed to be the 'opportunity cost of time' in theoretical models of health care use), and the costs of travelling to the medical facility concerned. If these differ between people, e.g. if some individuals have had further to travel than others, then there would seem to be inequality of access. Hence, this interpretation of access implies equality of personal (or 'private') cost.

3. The economic appraisal of health-care programmes

The benefits from a course of medical treatment are the value of the improvements in health that result from the treatment. Any attempt to measure these benefits therefore requires information concerning the value people place on improvements in their health. But this is not easy to obtain. Ill health usually involves a loss of earnings or earnings-potential, and one method of measuring the benefits from a particular health improvement is to calculate the associated reduction in lost earnings. However, this takes no account of the value of any reduction in pain and suffering (i.e. it ignores increases in utility that result from better health due to health care). Moreover, it assumes that individuals with no earnings or earnings potential (such as the retired elderly) place no value on their health. An alternative is to try to find out how much people would be prepared to pay to reduce the risk of their health deteriorating. Another is to calculate this sum by looking at the extra amount people are paid to work in industries where their health may be affected (usually, wages are higher in higher-risk occupations, such as deep sea diving, to compensate workers for the health risks they face). Yet another is to ask individuals to value different health states.

Research on this is currently under way in several European countries and in North America. While some of it is controversial, it is intended that the results can be used to establish the benefits of particular types of medical intervention. A good (but somewhat dated) article from the health economics literature on this topic is a review paper by Torrance, W., 1986, *Measurement of health state utilities for economic appraisal*, Journal of Health Economics 5, pages 1-30. I will briefly outline its main points. As we have already seen, the economic appraisal of a health care programme typically involves comparing the resources consumed with the health improvement generated by the programme. Torrance provides the following diagrammatic representation of the possible ways in which one can measure these costs and benefits:



Torrance identifies three types of costs, reflecting the resources consumed by a particular health programme:

C1 - these are direct costs, such as the cost of hiring the labour time of doctors, the cost of drugs and equipment used, etc.

C2 - these are indirect costs, such as the cost to the economy of individuals taking time off work in order to undergo the treatment.

C3 - these are intangible costs, expressed as the monetary value of the disutility (e.g. pain, discomfort, etc.) associated with a particular type of treatment, such as chemotherapy in the treatment of cancer.

He then sets out a range of possible measures of the health improvement (i.e. the 'output') of the health programme. These fall into three basic categories:

Health effects (E)

E1 - these are measures of output in terms of reduced morbidity (in the health economics jargon, 'morbidity' is just another word for 'illness'), expressed in terms of the illness itself. For example, cases of flu prevented, disability-days prevented, hospitalisation-days prevented.

E2 - these are measures of output in terms of reduced mortality e.g. number of lives saved by a particular programme, or the number of extra life-years produced by a programme.

Economic benefits (B)

B1 - these are the direct economic benefits arising from the health improvement produced by a health programme, such as the reduced health care costs resulting from making people healthier and therefore less dependent on the health care system in the future.

B2 - these are the indirect economic benefits, such as the increased output in the economy as a result of reducing days taken off work due to illness, or making workers healthier and therefore more productive at work.

B3 - these are the intangible economic benefits, expressed as the monetary value of reduced disutility due to illness (e.g. pain, discomfort etc.) as a result of implementing the programme.

Value of health improvement per se, regardless of economic consequences (V)

S - these are measures on ad hoc numeric scales (e.g. 'on a scale from 1 to 10') of the value of the health improvement itself. For example, on a scale from 1 to 10 (where 1 is 'worthless' and 10 is 'indispensable') how highly would you value your health improvement as a result of having an operation on your back to reduce back ache?

W - these are measures of willingness to pay for the health improvement that results from a programme e.g. if you had to, how much would you be willing to pay for the health improvement you get from an operation on your back to reduce back ache?

U - these are measures of <u>quality-adjusted life years</u> (QALYs) produced by the health programme. They are based on measurements of the utility associated with particular health states on a scale from 0 to 1, where 0 represents the utility of being dead, and 1 represents the utility of being in perfect health. If a treatment increases your utility from 0.2 to 0.4 for a period of 10 years, then we say that it has produced 0.2×10 = 2QALYs. Alternatively, a treatment (say for breast cancer) might extend your life for 5 years, in each of which you experience a utility of 0.4. Then again we would say that the treatment has produced 0.4×5 = 2QALYs. Thus, although the 'outputs' of the two treatments are very different, they are worth the same in terms of QALYs.

Torrance then shows how different formulas are used to combine the different cost and benefit measures in the three main methods of economic appraisal used today:
(1). <u>Cost effectiveness analysis</u> - This uses a cost/effectiveness ratio to measure the cost per unit of health improvement associated with a particular programme. For example, the cost in the numerator of the formula might be defined as the *net economic cost to society* in pounds sterling (C1 + C2 - B1 - B2), and the denominator might be defined as *reduced mortality* in terms of extra life-years (E2). Then the cost/effectiveness ratio would measure the cost in pounds sterling per extra life-year produced by the programme. The main disadvantage of this method is that it cannot be used to compare two health programmes whose effectiveness measures are expressed in different units. For example, it cannot be used to compare a health programme whose effectiveness is measured in terms of extra life-years, against a health programme whose effectiveness is measured in terms of cases of flu prevented.
(2). <u>Cost benefit analysis</u> - This measures the *net social benefit* (NSB) of a health-programme as

$$NSB = B1 + B2 + W - C1 - C2$$

From a theoretical viewpoint, it would be nice to be able to include the intangibles B3 and C3 in this measure, but Torrance points out that these are usually too difficult to estimate in practice. The decision rule is simply 'adopt the programme if its NSB is positive, do not implement it if its NSB is negative'. Cost benefit analysis based on the NSB overcomes some of the disadvantages of cost effectiveness analysis based on the cost/effectiveness ratio. However, Torrance argues that it places too much emphasis on labour market activity (via the term B2), and is therefore likely to underestimate the true production gain to society through effects on people such as housewives (and househusbands !), who do not earn wages, but who nevertheless do a valuable job for society in looking after children etc.
(3). <u>Cost utility analysis</u> - This is just another form of cost-effectiveness analysis (i.e. it is based on a cost/effectiveness ratio), but this time *effectiveness* is measured in QALYs. This overcomes the units-of-measurement problem associated with conventional cost effectiveness analysis, because QALYs provide a 'standard' measurement unit for the effectiveness of different health programmes.

In the rest of his paper, Torrance considers in some detail the use of utilities and QALYs in cost utility analysis. The fact that measuring the benefits from health care is difficult should not obscure the essential truth that it is impossible to make any reasonable decision concerning resource allocation in the health field without taking some view (however crude !) of the benefits involved. The hospital administrator who decides to allocate money to heart transplants rather than to kidney machines is making an implicit judgement about the relative benefits to be derived from each. The decision to use labour and capital to build a clinic rather than, say, a school or a mile of motorway implies that the benefits from an extra clinic are greater than those from an extra school or more motorway. **The difficulties in measuring benefits do not invalidate the definition of an efficient level of health care as one that**

equates the MSB and the MSC, because that definition is just a formal way of stating a social objective already implicit in most decisions concerning resource allocation. It is therefore worth attempting to measure benefits, however crudely, because the alternative is simply to rely on arbitrary and otherwise ill-informed assessments.