M.A. PREVIOUS ECONOMICS

PAPER I

MICRO ECONOMIC ANALYSIS

BLOCK 1

PARTIAL AND GENERAL EQUILIBRIUM,
LAW OF DEMAND AND DEMAND ANALYSIS
PAPER I

MICRO ECONOMIC ANALYSIS

BLOCK 1

PARTIAL AND GENERAL EQUILIBRIUM, LAW OF DEMAND AND DEMAND ANALYSIS

CONTENTS

<table>
<thead>
<tr>
<th>Unit</th>
<th>Title</th>
<th>Page number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction to Demand Theory</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>Concepts of Demand and Supply</td>
<td>22</td>
</tr>
<tr>
<td>3</td>
<td>Theories of Demand</td>
<td>42</td>
</tr>
</tbody>
</table>
The block opens with introduction to demand theory. Basic concepts of Demand are explained with Concept of Elasticity of Demand, Price Elasticity of Demand, Income Elasticity of Demand and Cross Price Elasticity. The unit also gives you the insight of various market forms.

The second unit covers different concepts of demand and supply. Models of Demand and Supply are discussed along with Demand and Supply Curves. The general and partial equilibrium approaches are also discussed in depth. The figurative representation of the approaches is taken to give readers an easy way to understand the concepts.

The third unit takes us into the domain of theories of demand. The Utility theory; Income and substitution effect; Indifference Curve; Revealed Preference; The Slutsky theorem and the Hicks Theory are discussed with price formation and discovery.
UNIT 1

INTRODUCTION TO DEMAND THEORY

Objectives

After studying this unit, you should be able to understand and appreciate:

- The concept of microeconomics and relevance of Demand
- The need to identify or define the concept of Demand.
- How to define elasticity of Demand
- Relevance of Price Elasticity of Demand
- Understand the approach to Income Elasticity of Demand
- The concept of Cross Price Elasticity
- Know the other forms of Markets in context of Microeconomics

Structure

1.1 Introduction
1.2 Basic concepts of Demand
1.3 Concept of Elasticity of Demand
1.4 Price Elasticity of Demand
1.5 Income Elasticity of Demand
1.6 Cross Price Elasticity
1.7 Other Market Forms
1.8 Summary
1.9 Further readings

1.1 INTRODUCTION

Besides Macroeconomics, the other basic way to view economics is the “Microeconomic” view. This view concerns itself with the particulars of a specific segment of the population or a specific industry within the larger population of good and service providers. More importantly, from a financial standpoint microeconomics concerns itself with the distribution of products, income, goods and services. Of course it is this distribution, which directly affects financial markets and the overall value of any particular resource at a specific point in time. If there is one concept integral to an understanding of microeconomics it is the law of supply and demand. A more detailed look at supply and demand as well as how they affect price will be helpful in understanding microeconomics.

Before discussing supply and demand it is helpful to understand what price is as a concept and how it relates to supply and demand. Price is essentially the feedback both the buyer and seller receive about the relative demand of a product, good or service.
When the price is high demand will be low and when the price is low demand will be high.

There are two laws intrinsically related to microeconomics. These two laws are the Law of Supply and the Law of Demand. A closer look at each will illustrate how they relate to pricing and the distribution of goods and services.

According to the LAW OF DEMAND, as price goes up; the quantity demanded by consumers goes down. As the price falls, the quantity demanded by consumers goes up. This law concerns itself with the consumer side of microeconomics. It tells us the quantity desired of a given product or service at a given price.

The LAW OF SUPPLY concerns itself with the entrepreneur or business, which supplies the products and services. This law tells us the amount of a product or service businesses will provide at a given price. Essentially, if everything else remains the same, businesses will supply more of a product or service at a higher price than they will at a lower price. This is because the higher price will attract more providers who seek to make a profit on the good or service. By the same token a low price will not attract additional suppliers and as a result the overall supply will remain low.

These two laws help to determine the overall price level of a product with a defined market. When evaluating the prices of an undefined market then another factor must be considered. This additional factor is called OPPORTUNITY COST. Opportunity cost is the relative loss of opportunity one must deal with in making a decision to invest time and money in something else. Needless to say, determining opportunity cost is very complicated and hard to evaluate in terms of economics.

Opportunity Cost is also used in evaluating the net cost of any good or service currently being utilized by an individual or the market as a whole. This can be illustrated by the decision a city makes to allocate a zone of land toward public recreation in the form of a park. The opportunity cost in this situation would be the loss of revenue the city would suffer by allocating the park instead of zoning the land for industrial use. Most situations involving opportunity cost are not so clear though.

The important concept to take away from opportunity costs is that for every purchasing or investing decision made there are other alternatives, which one is giving up. Therefore one is not just investing $5000 in government bonds but one is choosing to invest in bonds over funding the education of a child or of taking a vacation to the Bahamas for the entire family. Whether the investment is good or not depends on the value the family and the individual places on the alternative. These are the type of insights a microeconomic view can give the individual investor when applied correctly.
Supply and demand is an economic model based on price, utility and quantity in a market. It concludes that in a competitive market, price will function to equalize the quantity demanded by consumers, and the quantity supplied by producers, resulting in an economic equilibrium of price and quantity. An increase in the quantity produced or supplied will typically result in a reduction in price and vice-versa. Similarly, an increase in the number of workers tends to result in lower wages and vice-versa. The model incorporates other factors changing equilibrium as a shift of demand and/or supply.

1.2.1 Law of Demand

The Law of Demand states that other things held constant, as the price of a good increases, the quantity demanded will fall. Other factors that can influence demand include:

1. **Income** - Generally, as income increases, we are able to buy more of most goods. When demand for a good increases when incomes increase, we call that good a "normal good". When demand for a good decreases when incomes increase, then that good is called an inferior good.

2. **Price of related products** - Related goods come in two types, the first of which are "substitutes". Substitutes are similar products that can be used as alternatives. Examples of substitute goods are Coke/Pepsi, and butter/margarine. Usually, people substitute away to the less expensive good. Other related products are classified as "complements". Complements are products that are used in conjunction with each other. Examples of complements are pencil/eraser, left/right shoes, and coffee/sugar.

3. **Tastes and preferences** - Tastes are a major determinant of the demand for products, but usually does not change much in the short run.

4. **Expectations** - When you expect the price of a good to go up in the future, you tend to increase your demand today. This is another example of the rule of substitution, since you are substituting away from the expected relatively more expensive future consumption.

1.2.2 Demand Curves

Demand curves isolate the relationship between quantity demanded and the price of the product, while holding all other influences constant (in Latin: ceteris paribus). These curves show how many of a product will be purchased at different prices. Note that demand is represented by the entire curve, not just one point on the curve, and represents all the possible price-quantity choices given the ceteris paribus assumptions. When the
price of the product changes, quantity demanded changes, but demand does not change. Price changes involve a movement along the existing demand curve.

Market demand is the summation of all the individual demand curves of those in the market. It is the horizontal sum of individual curves and add up all the quantities demanded at each price. The main interest is in market demand curves, because they are averages of individual behaviour tend to be well-behaved.

When any influence other than the price of the product changes, such as income or tastes, demand changes, and the entire demand curve will shift (either upward or downward). A shift to the right (and up) is called an increase in demand, while a shift to the left (and down) is called a decrease in demand. In example, there are two ways to discourage smoking: raise the price through taxes or; make the taste less desirable.

1.2.3 Law of Supply

As the price of a product rises, ceteris paribus, suppliers will offer more for sale. This implies that price and quantity supplied are positively related. The major factor that influences supply is the "cost of production", and includes:

1. **Input prices** - As the prices of inputs such as labour, raw materials, and capital increase, production tends to be less profitable, and less will be produced. This leads to a decrease in supply.

2. **Technology** - Technology relates to methods of transforming inputs into outputs. Improvements in technology will reduce the costs of production and make sales more profitable so it tends to increase the supply.

3. **Expectations** - If firms expect prices to rise in the future, may try to product less now and more later.

1.2.4 Supply Curves and Schedules

The relationship between the price of a product and the quantity supplied, holding all other things constant is generally sloping upwards. Supply is represented by the entire curve and not just one point on the curve. When the price of the product changes, the quantity supplied changes, but supply does not change. When cost of production changes, supply changes, and the entire supply curve will shift.

Market Supply is the summation of all the individual supply curves, and is the horizontal sum of individual supply curves. It is influenced by the factors that determine individual supply curves, such as cost of production, plus the number of suppliers in the market. In general, the more firms producing a product, the greater the market supply.

When quantity supplied at a given price decreases, the whole curve shifts to the left as there is a decrease in supply. This is generally caused by an increase in the cost of
production or decrease in the number of sellers. An increase in wages, cost of raw materials, cost of capital, ceteris paribus, will decrease supply. Sometimes weather may also affect supply, if the raw materials are perishable or unattainable due to transportation problems.

1.2.5 Reaching Equilibrium

We can analyze how markets behave by matching (or combining) the supply and demand curves. Equilibrium is defined as the intersection of supply and demand curves. The equilibrium price is the price where the quantity demanded matches the quantity supplied. The equilibrium quantity is the quantity where price has adjusted so that \( Q_D = Q_S \). At the equilibrium price, the quantity that buyers are willing to purchase exactly equals the quantity the producers are willing to sell. Actions of buyers and sellers naturally tend to move a market towards the equilibrium. The concept and relationship between demand and supply and equilibrium will be discussed in depth in later units.

1.2.6 Excess Supply/Demand

Excess Supply is where Quantity supplied > Quantity demanded, and results in surpluses at the current price. A large surplus is known as a "glut". In cases of excess supply:

- price is too high to be at equilibrium
- suppliers find that inventories increase
- suppliers react by lowering prices
- this continues until price falls to equilibrium

Excess Demand occurs when Quantity demanded > Quantity supplied, and results in shortages at current prices. In cases of excess demand:

- buyers cannot buy all they want at the going price
- sellers find that their inventories are decreasing
- sellers can raise prices without losing sales
- prices increase until market reaches equilibrium

1.2.7 Demand schedule

In microeconomic theory, demand is defined as the \textit{willingness} and \textit{ability} of a consumer to purchase a given product in a given frame of time.

The demand schedule, depicted graphically as the demand curve, represents the amount of goods that buyers are willing and able to purchase at various prices, assuming all other non-price factors remain the same. The demand curve is almost always represented as downwards-sloping, meaning that as price decreases, consumers will buy more of the good.
Just as the supply curves reflect marginal cost curves, demand curves can be described as marginal utility curves.

The main determinants of individual demand are: the price of the good, level of income, personal tastes, the population (number of people), the government policies, the price of substitute goods, and the price of complementary goods.

The shape of the aggregate demand curve can be convex or concave, possibly depending on income distribution. In fact, an aggregate demand function cannot be derived except under restrictive and unrealistic assumptions.

As described above, the demand curve is generally downward sloping. There may be rare examples of goods that have upward sloping demand curves. Two different hypothetical types of goods with upward-sloping demand curves are a Giffen good (an inferior, but staple, good) and a Veblen good (a good made more fashionable by a higher price).

Similar to the supply curve, movements along it are also named expansions and contractions. A move downward on the demand curve is called an expansion of demand, since the willingness and ability of consumers to buy a given good has increased, in tandem with a fall in its price. Conversely, a move up the demand curve is called a contraction of demand, since consumers are less willing and able to purchase quantities of the product in question.

### 1.3 CONCEPT OF ELASTICITY OF DEMAND

Elasticity is a central concept in the theory of demand. In this context, elasticity refers to how demand respond to various factors, including price as well as other stochastic principles. One way to define elasticity is the percentage change in one variable divided by the percentage change in another variable (known as arc elasticity, which calculates the elasticity over a range of values, in contrast with point elasticity, which uses differential calculus to determine the elasticity at a specific point). It is a measure of relative changes.

Often, it is useful to know how the quantity demanded or supplied will change when the price changes. This is known as the price elasticity of demand and the price elasticity of supply. If a monopolist decides to increase the price of their product, how will this affect their sales revenue? Will the increased unit price offset the likely decrease in sales volume? If a government imposes a tax on a good, thereby increasing the effective price, how will this affect the quantity demanded?

Elasticity corresponds to the slope of the line and is often expressed as a percentage. In other words, the units of measure (such as gallons vs. quarts, say for the response of quantity demanded of milk to a change in price) do not matter, only the slope. Since supply and demand can be curves as well as simple lines the slope, and hence the elasticity, can be different at different points on the line.
Elasticity is calculated as the percentage change in quantity over the associated percentage change in price. For example, if the price moves from $1.00 to $1.05, and the quantity supplied goes from 100 pens to 102 pens, the slope is \( \frac{2}{0.05} \) or 40 pens per dollar. Since the elasticity depends on the percentages, the quantity of pens increased by 2%, and the price increased by 5%, so the price elasticity of supply is 2/5 or 0.4.

Since the changes are in percentages, changing the unit of measurement or the currency will not affect the elasticity. If the quantity demanded or supplied changes a lot when the price changes a little, it is said to be elastic. If the quantity changes little when the prices changes a lot, it is said to be inelastic. An example of perfectly inelastic supply, or zero elasticity, is represented as a vertical supply curve. (See that section below)

Elasticity in relation to variables other than price can also be considered. One of the most common to consider is income. How would the demand for a good change if income increased or decreased? This is known as the income elasticity of demand. For example, how much would the demand for a luxury car increase, if average income increased by 10%? If it is positive, this increase in demand would be represented on a graph by a positive shift in the demand curve. At all price levels, more luxury cars would be demanded.

Another elasticity sometimes considered is the cross elasticity of demand, which measures the responsiveness of the quantity demanded of a good to a change in the price of another good. This is often considered when looking at the relative changes in demand when studying complement and substitute goods. Complement goods are goods that are typically utilized together, where if one is consumed, usually the other is also. Substitute goods are those where one can be substituted for the other, and if the price of one good rises, one may purchase less of it and instead purchase its substitute.

Cross elasticity of demand is measured as the percentage change in demand for the first good that occurs in response to a percentage change in price of the second good. For an example with a complement good, if, in response to a 10% increase in the price of fuel, the quantity of new cars demanded decreased by 20%, the cross elasticity of demand would be -2.0.

In a perfect economy, any market should be able to move to the equilibrium position instantly without travelling along the curve. Any change in market conditions would cause a jump from one equilibrium position to another at once. So the perfect economy is actually analogous to the quantum economy. Unfortunately in real economic systems, markets don't behave in this way, and both producers and consumers spend some time travelling along the curve before they reach equilibrium position. This is due to asymmetric, or at least imperfect, information, where no one economic agent could ever be expected to know every relevant condition in every market. Ultimately both producers and consumers must rely on trial and error as well as prediction and calculation to find an the true equilibrium of a market.
When demand $D_1$ is in effect, the price will be $P_1$. When $D_2$ is occurring, the price will be $P_2$. The quantity is always $Q$, any shifts in demand will only affect price.

It is sometimes the case that a supply curve is vertical: that is the quantity supplied is fixed, no matter what the market price. For example, the surface area or land of the world is fixed. No matter how much someone would be willing to pay for an additional piece, the extra cannot be created. Also, even if no one wanted all the land, it still would exist. Land therefore has a vertical supply curve, giving it zero elasticity (i.e., no matter how large the change in price, the quantity supplied will not change).

Supply-side economics argues that the aggregate supply function – the total supply function of the entire economy of a country – is relatively vertical. Thus, supply-siders argue against government stimulation of demand, which would only lead to inflation with a vertical supply curve.

### 1.4 PRICE ELASTICITY OF DEMAND

**Price elasticity of demand (PED)** is defined as the measure of responsiveness in the quantity demanded for a commodity as a result of change in price of the same commodity. It is a measure of how consumers react to a change in price. In other words, it is percentage change in quantity demanded by the percentage change in price of the same commodity. In economics and business studies, the **price elasticity of demand** is a measure of the sensitivity of quantity demanded to changes in price. It is measured as elasticity, that is it measures the relationship as the ratio of percentage changes between quantity demanded of a good and changes in its price. In simpler words, demand for a product can be said to be very inelastic if consumers will pay almost any price for the
product, and very elastic if consumers will only pay a certain price, or a narrow range of prices, for the product.

Inelastic demand means a producer can raise prices without much hurting demand for its product, and elastic demand means that consumers are sensitive to the price at which a product is sold and will not buy it if the price rises by what they consider too much. Drinking water is a good example of a good that has inelastic characteristics in that people will pay anything for it (high or low prices with relatively equivalent quantity demanded), so it is not elastic. On the other hand, demand for sugar is very elastic because as the price of sugar increases, there are many substitutions which consumers may switch to.

**Interpretation of elasticity**

A price fall usually results in an increase in the quantity demanded by consumers (see Giffen good for an exception). The demand for a good is relatively inelastic when the change in quantity demanded is less than change in price. Goods and services for which no substitutes exist are generally inelastic. Demand for an antibiotic, for example, becomes highly inelastic when it alone can kill an infection resistant to all other antibiotics. Rather than die of an infection, patients will generally be willing to pay whatever is necessary to acquire enough of the antibiotic to kill the infection.

![Figure 2 Perfectly inelastic demand](image-url)
Various research methods are used to calculate price elasticity:

- Test markets
- Analysis of historical sales data
- Conjoint analysis

**Determinants**

A number of factors determine the elasticity:
• **Substitutes:** The more substitutes, the higher the elasticity, as people can easily switch from one good to another if a minor price change is made.

• **Percentage of income:** The higher the percentage that the product’s price is of the consumer’s income, the higher the elasticity, as people will be careful with purchasing the good because of its cost.

• **Necessity:** The more necessary a good is, the lower the elasticity, as people will attempt to buy it no matter the price, such as the case of insulin for those that need it.

• **Duration:** The longer a price change holds, the higher the elasticity, as more and more people will stop demanding the goods (i.e. if you go to the supermarket and find that blueberries have doubled in price, you’ll buy it because you need it this time, but next time you won’t, unless the price drops back down again).

• **Breadth of definition:** The broader the definition, the lower the elasticity. For example, Company X's fried dumplings will have a relatively high elasticity, whereas food in general will have an extremely low elasticity (see Substitutes, Necessity above).

### Elasticity and revenue

![Figure 4 Elasticity and Revenue Relationship](image)

A set of graphs shows the relationship between demand and total revenue. As price decreases in the elastic range, revenue increases, but in the inelastic range, revenue decreases.
When the price elasticity of demand for a good is inelastic ($|E_d| < 1$), the percentage change in quantity demanded is smaller than that in price. Hence, when the price is raised, the total revenue of producers rises, and vice versa.

When the price elasticity of demand for a good is elastic ($|E_d| > 1$), the percentage change in quantity demanded is greater than that in price. Hence, when the price is raised, the total revenue of producers falls, and vice versa.

When the price elasticity of demand for a good is unit elastic (or unitary elastic) ($|E_d| = 1$), the percentage change in quantity is equal to that in price.

When the price elasticity of demand for a good is perfectly elastic ($E_d$ is undefined), any increase in the price, no matter how small, will cause demand for the good to drop to zero. Hence, when the price is raised, the total revenue of producers falls to zero. The demand curve is a horizontal straight line. A banknote is the classic example of a perfectly elastic good; nobody would pay £10.01 for a £10 note, yet everyone will pay £9.99 for it.

When the price elasticity of demand for a good is perfectly inelastic ($E_d = 0$), changes in the price do not affect the quantity demanded for the good. The demand curve is a vertical straight line; this violates the law of demand. An example of a perfectly inelastic good is a human heart for someone who needs a transplant; neither increases nor decreases in price affect the quantity demanded (no matter what the price, a person will pay for one heart but only one; nobody would buy more than the exact amount of hearts demanded, no matter how low the price is).

### 1.5 INCOME ELASTICITY OF DEMAND

In economics, the **income elasticity of demand** measures the responsiveness of the demand of a good to the change in the income of the people demanding the good. It is calculated as the ratio of the percent change in demand to the percent change in income. For example, if, in response to a 10% increase in income, the demand of a good increased by 20%, the income elasticity of demand would be $20%/10% = 2$.

Thus far, we have dealt with the effect of a change in the price of a good on the same good's quantity demanded or supplied. Now we turn our attention to the impact on the demand for a good when consumer incomes change, holding prices constant. The business cycle describes alternating periods of economic growth, when incomes generally increase, and contraction (recession) which lead to a decrease in consumer incomes. A firm needs to understand income elasticity to see how changes in the macroeconomy translates into the demand for the good or service produced by the firm. Our consumption of some goods, such as luxuries, is very sensitive to changes in economic growth and consumer incomes. In contrast, necessities such as food and housing tend to be less affected by economic swings and the corresponding changes in consumer incomes.

There are three possibilities for a good's income elasticity:
1. A good is **income elastic** if the income elasticity of demand is greater than 1. This implies that for a 1% change in income, demand for the good changes by more than 1%.

2. A good is **income inelastic** if the income elasticity of demand is greater than 0 but less than 1. This implies that for a 1% change in income, demand for the good changes by less than 1%.

3. A good is considered **inferior** if the associated income elasticity of demand is a negative number. In this case, if income increases, consumers actually buy less of the good.

**Normal Goods**

A positive income elasticity of demand is associated with normal goods; an increase in income will lead to a rise in demand. If income elasticity of demand of a commodity is less than 1, it is a necessity good. If the elasticity of demand is greater than 1, it is a luxury good or a superior good.

Since Normal goods have a positive income elasticity of demand so as income rise more is demand at each price level. We make a distinction between normal necessities and normal luxuries (both have a positive coefficient of income elasticity). Necessities have an income elasticity of demand of between 0 and +1. Demand rises with income, but less than proportionately. Often this is because we have a limited need to consume additional quantities of necessary goods as our real living standards rise. The class examples of this would be the demand for fresh vegetables, toothpaste and newspapers. Demand is not very sensitive at all to fluctuations in income in this sense total market demand is relatively stable following changes in the wider economic (business) cycle.

**Luxuries**

Luxuries are said to have an income elasticity of demand > +1. (Demand rises more than proportionate to a change in income). Luxuries are items we can (and often do) manage to do without during periods of below average income and falling consumer confidence. When incomes are rising strongly and consumers have the confidence to go ahead with “big-ticket” items of spending, so the demand for luxury goods will grow. Conversely in a recession or economic slowdown, these items of discretionary spending might be the first victims of decisions by consumers to rein in their spending and rebuild savings and household financial balance sheets.

Many luxury goods also deserve the sobriquet of “positional goods”. These are products where the consumer derives satisfaction (and utility) not just from consuming the good or service itself, but also from being seen to be a consumer by others.

**Inferior Goods**

Inferior goods have a negative income elasticity of demand. Demand falls as income rises. In a recession the demand for inferior products might actually grow (depending on the severity of any change in income and also the absolute co-efficient of income
For example if we find that the income elasticity of demand for cigarettes is -0.3, then a 5% fall in the average real incomes of consumers might lead to a 1.5% fall in the total demand for cigarettes (ceteris paribus).

<table>
<thead>
<tr>
<th>Normal Luxury</th>
<th>Normal Necessity</th>
<th>Inferior Good</th>
</tr>
</thead>
<tbody>
<tr>
<td>International air travel</td>
<td>Fresh vegetables</td>
<td>Frozen vegetables</td>
</tr>
<tr>
<td>Fine wines</td>
<td>Instant coffee</td>
<td>Cigarettes</td>
</tr>
<tr>
<td>Luxury chocolates</td>
<td>Natural cheese</td>
<td>Processed cheese</td>
</tr>
<tr>
<td>Private education</td>
<td>Fruit juice</td>
<td>Margarine</td>
</tr>
<tr>
<td>Private health care</td>
<td>Spending on utilities</td>
<td>Tinned meat</td>
</tr>
<tr>
<td>Antique furniture</td>
<td>Shampoo / toothpaste / detergents</td>
<td>Value “own-brand” bread</td>
</tr>
<tr>
<td>Designer clothes</td>
<td>Rail travel</td>
<td>Bus travel</td>
</tr>
</tbody>
</table>

Table 1

Within a given market, the income elasticity of demand for various products can vary and of course the perception of a product must differ from consumer to consumer. The hugely important market for overseas holidays is a great example to develop further in this respect.

What to some people is a necessity might be a luxury to others. For many products, the final income elasticity of demand might be close to zero, in other words there is a very weak link at best between fluctuations in income and spending decisions. In this case the “real income effect” arising from a fall in prices is likely to be relatively small. Most of the impact on demand following a change in price will be due to changes in the relative prices of substitute goods and services.
The income elasticity of demand for a product will also change over time – the vast majority of products have a finite life-cycle. Consumer perceptions of the value and desirability of a good or service will be influenced not just by their own experiences of consuming it (and the feedback from other purchasers) but also the appearance of new products onto the market. Consider the income elasticity of demand for flat-screen colour televisions as the market for plasma screens develops and the income elasticity of demand for TV services provided through satellite dishes set against the growing availability and falling cost (in nominal and real terms) and integrated digital televisions.

A zero income elasticity (or inelastic) demand occurs when an increase in income is not associated with a change in the demand of a good. These would be sticky goods.

1.6 CROSS-PRICE ELASTICITY

The final type of elasticity is known as the cross-price elasticity and measures the responsiveness of our consumption of one good when the price of another good changes. The cross-price elasticity of two goods, say good A and good B, measures the percentage change in the quantity demanded of good A, when the price of good B changes by 1%. Cross-price elasticity's are given two categories: complements and substitutes.

- **Complements** - Two goods that have a **negative value for their cross-price elasticity** are considered complementary goods such as compact disk (CD) players and compact disks. If the price of CD players increases then our consumption of CD's decreases, leading to a negative relationship between the two. Conversely, if the price of CD players falls (a negative coefficient), our consumption of CD's rises (a positive coefficient).
Substitutes - Two goods that have a positive value for their cross-price elasticity are considered substitutes such as gasoline prices and the demand for public transportation. If the price of gasoline rises, so does consumer demand for less expensive transportation alternatives such as public transportation (buses, subways).

Cross-price elasticity is important for producers to recognize. Demand for goods in industries such as autos are significantly impacted by changes in price of complements and substitutes, most noticeably gasoline and the prices of cars produced by a competing firm. Individual firms will carefully judge the impact of competitor pricing and the responsiveness of consumers to those price changes. Goods with a higher degree of substitution will have a greater positive value for their cross-price elasticity. Likewise, goods that show a more complementary relationship have an increasing negative value for their cross-price elasticity.

Take the example of the airline industry and consider goods that are close substitutes. For example one good is the price of seat on American Airlines, the other good is the demand for seat on United Airlines, each on an identical flight route - say Boston to Washington DC. In the case of the airline industry, the cross-price elasticity of demand for airline tickets is very high, and firms respond immediately to fare changes. If one airline such as American initiates a fare war, competitors such as United quickly follow in reducing prices to prevent a loss of market share. Since there is a high cross-price elasticity, if American lowers its fare from Boston to Washington DC, and United keeps its fares constant, consumers quickly shift consumption towards the lower priced American tickets. The resulting decrease in the demand for United Airlines tickets is large.

1.7 OTHER MARKET FORMS

The supply and demand model is used to explain the behavior of perfectly competitive markets, but its usefulness as a standard of performance extends to other types of markets. In such markets, there may be no supply curve, such as above, except by analogy. Rather, the supplier or suppliers are modeled as interacting with demand to determine price and quantity. In particular, the decisions of the buyers and sellers are interdependent in a way different from a perfectly competitive market.

A monopoly is the case of a single supplier that can adjust the supply or price of a good at will. The profit-maximizing monopolist is modeled as adjusting the price so that its profit is maximized given the amount that is demanded at that price. This price will be higher than in a competitive market. A similar analysis can be applied when a good has a single buyer, a monopsony, but many sellers. Oligopoly is a market with so few suppliers that they must take account of their actions on the market price or each other. Game theory may be used to analyze such a market.

The supply curve does not have to be linear. However, if the supply is from a profit-maximizing firm, it can be proven that curves-downward sloping supply curves (i.e., a price decrease increasing the quantity supplied) are inconsistent with perfect competition.
in equilibrium. Then supply curves from profit-maximizing firms can be vertical, horizontal or upward sloping.

**Other markets**

The model of supply and demand also applies to various specialty markets.

The model applies to wages, which are determined by the market for labor. The typical roles of supplier and consumer are reversed. The suppliers are individuals, who try to sell their labor for the highest price. The consumers of labors are businesses, which try to buy the type of labor they need at the lowest price. The equilibrium price for a certain type of labor is the wage.

The model applies to interest rates, which are determined by the money market. In the short term, the money supply is a vertical supply curve, which the central bank of a country can influence through monetary policy. The demand for money intersects with the money supply to determine the interest rate.

**Activity 1**

1. Discuss in brief basic concepts of Demand.
2. What do you understand by the concept of elasticity? Discuss price elasticity of demand.
3. Explain the relevance of income elasticity of demand giving suitable examples.
4. Give the brief note on cross-price elasticity of demand. Why it is important for producers to focus on this kind of elasticity?

**1.8 SUMMARY**

The demand for a product is the amount that buyers are willing and able to purchase. Quantity demanded is the demand at a particular price, and is represented as the demand curve. The supply of a product is the amount that producers are willing and able to bring to the market for sale. Quantity supplied is the amount offered for sale at a particular price. The main determinant of supply/demand is the price of the product. If there is an increase in the price of a good, the quantity demanded will fall. We use the concept of elasticity to determine how much the quantity demanded of a good responds to a change in the price of that good. The price elasticity of demand measures the change in the quantity demanded for a good in response to a change in price. Similarly income elasticity measures the responsiveness of the demand of a good to the change in the income of the people demanding. Cross price elasticity measures the responsiveness of our consumption of one good when the price of another good changes. Other market forms in context of theory of demand have also discussed in brief.
1.9 FURTHER READINGS

UNIT 2

CONCEPTS OF DEMAND AND SUPPLY

Objectives

After studying this unit, you should be able to understand:

- Concepts of Demand, Supply and equilibrium
- The Model of Demand and Supply
- The contributions of Aggregate Demand and Aggregate Supply to Demand analysis
- The methods of developing Demand and Supply curves
- The approaches toward General and Partial Equilibrium
- Demand Analysis and recent developments.

Structure

2.1 Introduction
2.2 The Model of Demand and Supply
2.3 Aggregate Demand and Aggregate Supply
2.4 Demand and Supply Curves
2.5 The General Equilibrium
2.6 The Partial Equilibrium
2.7 Recent Developments in Demand analysis
2.8 Summary
2.9 Further Readings

2.1 INTRODUCTION

The market price of a good is determined by both the supply and demand for it. In 1890, English economist Alfred Marshall published his work, *Principles of Economics*, which was one of the earlier writings on how both supply and demand interacted to determine price. We have already discussed the basic concepts of demand and supply in previous unit. Today, the supply-demand model is one of the fundamental concepts of economics. The price level of a good essentially is determined by the point at which quantity supplied equals quantity demanded. To illustrate, consider the following case in which the supply and demand curves are plotted on the same graph.
On this graph, there is only one price level at which quantity demanded is in balance with the quantity supplied, and that price is the point at which the supply and demand curves cross.

The law of supply and demand predicts that the price level will move toward the point that equalizes quantities supplied and demanded. To understand why this must be the equilibrium point, consider the situation in which the price is higher than the price at which the curves cross. In such a case, the quantity supplied would be greater than the quantity demanded and there would be a surplus of the good on the market. Specifically, from the graph we see that if the unit price is $3 (assuming relative pricing in dollars), the quantities supplied and demanded would be:

\[
\text{Quantity Supplied} = 42 \text{ units} \\
\text{Quantity Demanded} = 26 \text{ units}
\]

Therefore there would be a surplus of 42 - 26 = 16 units. The sellers then would lower their price in order to sell the surplus.

Suppose the sellers lowered their prices below the equilibrium point. In this case, the quantity demanded would increase beyond what was supplied, and there would be a shortage. If the price is held at $2, the quantity supplied then would be:

\[
\text{Quantity Supplied} = 28 \text{ units} \\
\text{Quantity Demanded} = 38 \text{ units}
\]

Therefore, there would be a shortage of 38 - 28 = 10 units. The sellers then would increase their prices to earn more money.
The equilibrium point must be the point at which quantity supplied and quantity demanded are in balance, which is where the supply and demand curves cross. From the graph above, one sees that this is at a price of approximately $2.40 and a quantity of 34 units.

To understand how the law of supply and demand functions when there is a shift in demand, consider the case in which there is a shift in demand:

![Shift in Demand](image)

In this example, the positive shift in demand results in a new supply-demand equilibrium point that is higher in both quantity and price. For each possible shift in the supply or demand curve, a similar graph can be constructed showing the effect on equilibrium price and quantity. The following table summarizes the results that would occur from shifts in supply, demand, and combinations of the two.

### Result of Shifts in Supply and Demand

<table>
<thead>
<tr>
<th>Demand</th>
<th>Supply</th>
<th>Equilibrium Price</th>
<th>Equilibrium Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td>?</td>
<td>+</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>?</td>
<td>-</td>
</tr>
<tr>
<td>+</td>
<td>-</td>
<td>+</td>
<td>?</td>
</tr>
<tr>
<td>-</td>
<td>+</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

Table 1
In the above table, "+" represents an increase, "-" represents a decrease, a blank represents no change, and a question mark indicates that the net change cannot be determined without knowing the magnitude of the shift in supply and demand. If these results are not immediately obvious, drawing a graph for each will facilitate the analysis.

### 2.2 THE MODEL OF SUPPLY AND DEMAND

To this point, we have developed two behavioral statements, or assertions, about how people will act. The first says that the amount buyers are willing and ready to buy depends on price and other factors that are assumed constant. The second says that the amount sellers are willing and ready to sell depends on price and other factors that are assumed constant. In mathematical terms our model is

\[
Q_d = f(\text{price, constants}) \\
Q_s = g(\text{price, constants})
\]

This is not a complete model. Mathematically, the problem is that we have three variables (Qd, Qs, price) and only two equations, and this system will not have a solution. To complete the system, we add a simple equation containing the equilibrium condition:

\[Q_d = Q_s.\]

In words, equilibrium exists if the amount sellers are willing to sell is equal to the amount buyers are willing to buy.

If we combine the supply and demand tables in earlier sections, we get the table below. It should be obvious that the price of $3.00 is the equilibrium price and the quantity of 70 is the equilibrium quantity. At any other price, sellers would want to sell a different amount than buyers want to buy.

<table>
<thead>
<tr>
<th>Price of Widgets</th>
<th>Number of Widgets People Want to Buy</th>
<th>Number of Widgets Sellers Want to Sell</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1.00</td>
<td>100</td>
<td>10</td>
</tr>
<tr>
<td>$2.00</td>
<td>90</td>
<td>40</td>
</tr>
<tr>
<td>$3.00</td>
<td>70</td>
<td>70</td>
</tr>
<tr>
<td>$4.00</td>
<td>40</td>
<td>140</td>
</tr>
</tbody>
</table>

**Table 2**

The same information can be shown with a graph. On the graph, the equilibrium price and quantity are indicated by the intersection of the supply and demand curves.
If one of the many factors that is being held constant changes, then equilibrium price and quantity will change. Further, if we know which factor changes, we can often predict the direction of changes, though rarely the exact magnitude. For example, the market for wheat fits the requirements of the supply and demand model quite well. Suppose there is a drought in the main wheat-producing areas of the United States. How will we show this on a supply and demand graph? Should we move the demand curve, the supply curve, or both? What will happen to equilibrium price and quantity?

A dangerous way to answer these questions is to first try to decide what will happen to price and quantity and then decide what will happen to the supply and demand curves. This is a route to disaster. Rather, one must first decide how the curves will shift, and then from the shifts in the curves decide how price and quantity would change.

What should happen as the result of the drought? One begins by asking whether buyers would change the amount they purchased if price did not change and whether sellers would change the amount sold if price did not change. On reflection, one realizes that this event will change seller behavior at the given price, but is highly unlikely to change buyer behavior (unless one assumes that more than the drought occurs, such as a change in expectations caused by the drought). Further, at any price, the drought will reduce the amount sellers will sell. Thus, the supply curve will shift to the left and the demand curve will not change. There will be a change in supply and a change in quantity demanded. The new equilibrium will have a higher price and a lower quantity. These changes are shown below.

**Assumptions to demand and supply model**

The supply and demand model does not describe all markets--there is too much diversity in the ways buyers and sellers interact for one simple model to explain everything. When we use the supply and demand model to explain a market, we are implicitly making a number of assumptions about that market.
Supply and demand analysis assumes competitive markets. For a supply curve to exist, there must be a large number of sellers in the market; and for a demand curve to exist, there must be many buyers. In both cases there must be enough so that no one believes that what he does will influence price. In terms that were first introduced into economics in the 1950s and that have become quite popular, everyone must be a \textit{price taker} and no one can be a \textit{price searcher}. If there is only one seller, that seller can search along the demand curve to find the most profitable price.\footnote{A price taker cannot influence the price, but must take or leave it. The ordinary consumer knows the role of price taker well. When he goes to the store, he can buy one or twenty gallons of milk with no effect on price. The assumption that both buyers and sellers are price takers is a crucial assumption, and often it is not true with regard to sellers. If it is not true with regard to sellers, a supply curve will not exist because the amount a seller will want to sell will depend not on price but on marginal revenue.} A price taker cannot influence the price, but must take or leave it. The ordinary consumer knows the role of price taker well. When he goes to the store, he can buy one or twenty gallons of milk with no effect on price. The assumption that both buyers and sellers are price takers is a crucial assumption, and often it is not true with regard to sellers. If it is not true with regard to sellers, a supply curve will not exist because the amount a seller will want to sell will depend not on price but on marginal revenue.

The model of supply and demand also requires that buyers and sellers be clearly defined groups. Notice that in the list of factors that affected buyers and sellers, the only common factor was price. Few people who buy hamburger know or care about the price of cattle feed or the details of cattle breeding. Cattle raisers do not care what the income of the buyers is or what the prices of related goods are unless they affect the price of cattle. Thus, when one factor changes, it affects only one curve, not both. When buyers and sellers cannot be clearly distinguished, as on the New York Stock Exchange, where the people who are buyers one minute may be sellers the next, one cannot talk about distinct and separate supply and demand curves.

The model of supply and demand also assumes that both buyers and sellers have good information about the product's qualities and availability. If information is not good, the same product may sell for a variety of prices. Often, however, what seems to be the same product at different prices can be considered a variety of products. A pound of hamburger for which one has to wait 15 minutes in a check-out line can be considered a different product from identical meat that one can buy without waiting.

Finally, for some uses the supply and demand model needs well-defined private-property rights. Elsewhere, we discussed how private-property rights and markets provide one way of coordinating decisions. When property rights are not clearly defined, the seller may be able to ignore some of the costs of production, which will then be imposed on others. Alternatively, buyers may not get all the benefits from purchasing a product; others may get some of the benefits without payment.

Even if the assumptions underlying supply and demand are not met exactly, and they rarely are, the model often provides a fairly good approximation of a situation, good enough so that predictions based on the model are in the right direction. This ability of the model to predict even when some assumptions are not quite satisfied is one reason economists like the model so much.
What should one predict if a new diet calling for the consumption of two loaves of whole wheat bread sweeps through the U.S.? Again one must ask whether the behavior of buyers or sellers will change if price does not change. Reflection should tell you that it will be the behavior of buyers that will change. Buyers would want more wheat at each possible price. The demand curve shifts to the right, which results in higher equilibrium price and quantity. Sellers would also change their behavior, but only because price changed. Sellers would move along the supply curve.

### 2.4 AGGREGATE DEMAND AND AGGREGATE SUPPLY

ISLM aggregates the economy into a market for money balances, a market for goods and services, and a residual market that it ignores by invoking Walras' Law. The ISLM model is a macroeconomic tool that demonstrates the relationship between interest rates and real output in the goods and services market and the money market. The intersection of the IS and LM curves is the "General Equilibrium" where there is simultaneous equilibrium in all the markets of the economy. IS/LM stands for Investment Saving / Liquidity preference Money supply.

Since part of the residual market is the labor market, and because adjustment in this market is slow, ISLM would be a better model if it could capture what is happening in the resource markets. Aggregate supply-aggregate demand analysis makes this incorporation.

The aggregate demand curve is derived from the ISLM model. In the illustration below, equilibrium income is $Y_1$ when the price level is $P_1$. Let the price level rise to a higher level, from $P_1$ to $P_2$. At the higher level, with a constant amount of money, purchasing power is cut. The fixed number of dollars no longer buys as much. The effects on the LM curve are identical to what happens when prices remain fixed and the amount of money falls. The LM curve, in either case, shifts left, interest rates rise, and income falls. The
output levels at both P1 and P2 are shown in the bottom part of the illustration. The aggregate demand curve connects them with points that other price levels generate.

The aggregate supply curve comes from the resource market. Though these markets may adjust slowly, when they finally do fully adjust, price level should have little or no effect on the amount of resources supplied. If a doubling of all prices and wages results in more or less output, someone is suffering from money illusion. The person believes either that he is better off at a higher nominal (but same real) wage, or that he is worse off with higher prices that have been fully compensated with higher wages. If people realize that money is merely an intermediary, and ultimately goods trade for goods, price level should not matter.

The point of the last paragraph is important enough to explain in a more concrete manner. Suppose Edward has a paper route and at the end of each week his income is $25.00. He spends his entire income on 15 hamburgers that cost $1.00 each and 20 soft drinks that cost $.50 each. One day Edward wakes up and finds that his weekly income has doubled to $50, but all prices have also doubled. Is he any better or worse off? Clearly he is not. A week of delivering newspapers still trades for 15 hamburgers and 20 soft drinks. He has no reason to work either more or less.

If behavior does not change when price level does, output will not depend on price level. The result will be the perfectly vertical aggregate supply curve shown in the illustration.
above. In the long run, when prices and wages fully adjust to any change in total spending, resources and output determine output.

In the short-run, however, an adjustment process that is not instantaneous seems more appropriate. Prices can be sticky, especially in resource markets. Expected rates of inflation can affect the way prices are set. Once we allow these possibilities, we have a system in which it may take years to reach long-run equilibrium. It is even possible that the system will never reach equilibrium, but, as the business-cycle writers thought, will fluctuate forever in the adjustment process.

Once we add stickiness to prices and give a role to expected inflation, a change in spending will not simply move the economy up or down a vertical aggregate-supply curve. The upward-sloping curve below shows what is likely in the short run. A change in spending will move the aggregate-demand curve. If the short-run aggregate-supply curve is fairly flat, there will be a large change in output and a small change in price level.

![Graph of Aggregate Supply and Aggregate Demand](image)

**Figure 6**

Aggregate supply and aggregate demand is an attractive framework because it is simple, with the same structure as supply and demand. However, the assumptions behind aggregate supply and aggregate demand are totally different from those behind supply and demand, that is, aggregate supply and aggregate demand curves are not obtained by adding up all the supply and demand curves in an economy. If they were, one would expect that the long-run aggregate-supply curve would be flatter than the short-run aggregate-supply curve, as is the case with a normal supply curve. But the aggregate supply curve grows steeper the longer the time for adjustment.

Aggregate supply and aggregate demand is more general than ISLM, and overcomes some of the limitations of ISLM. It includes price level as a variable, and it shows that
resource markets matter. It also lets one consider cases in which disturbances originate in a resource market, such as a disruption of oil supplies, which ISLM cannot handle.

Aggregate supply and aggregate demand gives insight into the adjustment process. Observation of the real world tells us that when spending suddenly changes, output changes initially more than prices, and only after considerable delay do prices change more than output. Aggregate supply and aggregate demand yields this pattern.

Aggregate demand and aggregate supply show an adjustment process. It does this with a series of short-run equilibria. Alfred Marshall originated this technique with regular supply and demand. He had three periods: the market period or the very short run, in which output was fixed; the short run, in which capital was fixed but utilization of capital was not; and the long run, in which nothing was fixed. So far the expositions of aggregate supply and aggregate demand have been fuzzy about what is fixed in the short run that is not fixed in the long run. This fuzziness remains as a problem of aggregate demand and aggregate supply.

### 2.5 DEMAND AND SUPPLY CURVES

Supply and Demand curves play a fundamental role in Economics. The supply curve indicates how many producers will supply the product (or service) of interest at a particular price. Similarly, the demand curve indicates how many consumers will buy the product at a given price. By drawing the two curves together, it is possible to calculate the market clearing price. This is the intersection of the two curves and is the price at which the amount supplied by the producers will match exactly the quantity that the consumers will buy.

The downward sloping line is the demand curve, while the upward sloping line is the supply curve. The demand curve indicates that if the price were $10, the demand would be zero. However, if the price dropped to $8, the demand would increase to 4 units. Similarly, if the price were to drop to $2, the demand would be for 16 units.

The supply curve indicates how much producers will supply at a given price. If the price were zero, no one would produce anything. As the price increases, more producers would come forward. At a price of $5, there would be 5 units produced by various suppliers. At a price of $10, the suppliers would produce 10 units.

The intersection of the supply curve and the demand curve, shown by \((P^*, Q^*)\), is the market clearing condition. In this example, the market clearing price is \(P^* = 6.67\) and the market clearing quantity is \(Q^* = 6.67\). At the price of $6.67, various producers supply a total of 6.67 units, and various consumers demand the same quantity.
There is no reason why the curves have to be straight lines. They could be different shapes such in the examples below. However, for the sake of simplicity, we will work with straight line demand and supply functions.

Creating the market Demand and Supply curves from the preferences of individual producers and suppliers

In the examples above, the chart contained smooth curves. While such a curve is an excellent approximation when there are many producers (or consumers), each of the curves is actually made up of many small discrete steps. Each of these steps represents
the decision of a single individual (or company). We will see next how these curves are constructed based on the decisions made by individual entities.

We construct the demand and supply curves for a very small market. Suppose there are just 5 consumers and each demands one unit of the product. However, they have distinct prices at which the product is valuable enough for them to buy it. Table 3 shows the price at which each individual will buy the product.

<table>
<thead>
<tr>
<th>Price</th>
<th>Product bought by consumer</th>
<th>Total demand for product</th>
</tr>
</thead>
<tbody>
<tr>
<td>More than 20</td>
<td>None</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>B</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>C</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>D</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>E</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 3

2.6 THE GENERAL EQUILIBRIUM THEORY

**General equilibrium theory** is a branch of theoretical economics. It seeks to explain the behavior of supply, demand and prices in a whole economy with several or many markets. It is often assumed that agents are price takers and in that setting two common notions of equilibrium exist: Walrasian (or competitive) equilibrium, and its generalization; a price equilibrium with transfers.

Broadly speaking, general equilibrium tries to give an understanding of the whole economy using a "bottom-up" approach, starting with individual markets and agents. Macroeconomics, as developed by the Keynesian economists, focused on a "top-down" approach, where the analysis starts with larger aggregates, the "big picture". Therefore general equilibrium theory has traditionally been classed as part of microeconomics.

The difference is not as clear as it used to be, however, since much of modern macroeconomics has emphasized microeconomic foundations, and has constructed general equilibrium models of macroeconomic fluctuations. But general equilibrium macroeconomic models usually have a simplified structure that only incorporates a few markets, like a "goods market" and a "financial market". In contrast, general equilibrium models in the microeconomic tradition typically involve a multitude of different goods markets. They are usually complex and require computers to help with numerical solutions.
In a market system, the prices and production of all goods, including the price of money and interest, are interrelated. A change in the price of one good -- say, bread -- may affect another price, such as bakers' wages. If bakers differ in tastes from others, the demand for bread might be affected by a change in bakers' wages, with a consequent effect on the price of bread. Calculating the equilibrium price of just one good, in theory, requires an analysis that accounts for all of the millions of different goods that are available.

2.6.1 Modern concept of general equilibrium in economics

The modern conception of general equilibrium is provided by a model developed jointly by Kenneth Arrow, Gerard Debreu and Lionel W. McKenzie in the 1950s. Gerard Debreu presents this model in Theory of Value (1959) as an axiomatic model, following the style of mathematics promoted by Bourbaki. In such an approach, the interpretation of the terms in the theory (e.g., goods, prices) is not fixed by the axioms.

Three important interpretations of the terms of the theory have been often cited. First, suppose commodities are distinguished by the location where they are delivered. Then the Arrow-Debreu model is a spatial model of, for example, international trade.

Second, suppose commodities are distinguished by when they are delivered. That is, suppose all markets equilibrate at some initial instant of time. Agents in the model purchase and sell contracts, where a contract specifies, for example, a good to be delivered and the date at which it is to be delivered. The Arrow-Debreu model of intertemporal equilibrium contains forward markets for all goods at all dates. No markets exist at any future dates.

Third, suppose contracts specify states of nature which affect whether a commodity is to be delivered: "A contract for the transfer of a commodity now specifies, in addition to its physical properties, its location and its date, an event on the occurrence of which the transfer is conditional. This new definition of a commodity allows one to obtain a theory of [risk] free from any probability concept..." (Debreu, 1959)

These interpretations can be combined. So the complete Arrow-Debreu model can be said to apply when goods are identified by when they are to be delivered, where they are to be delivered, and under what circumstances they are to be delivered, as well as their intrinsic nature. So there would be a complete set of prices for contracts such as "1 ton of Winter red wheat, delivered on 3rd of January in Minneapolis, if there is a hurricane in Florida during December". A general equilibrium model with complete markets of this sort seems to be a long way from describing the workings of real economies, however its proponents argue that it is still useful as a simplified guide as to how a real economies function.

Some of the recent work in general equilibrium has in fact explored the implications of incomplete markets, which is to say an intertemporal economy with uncertainty, where
there do not exist sufficiently detailed contracts that would allow agents to fully allocate their consumption and resources through time. While it has been shown that such economies will generally still have equilibrium, the outcome may no longer be Pareto optimal. The basic intuition for this result is that if consumers lack adequate means to transfer their wealth from one time period to another and the future is risky, there is nothing to necessarily tie any price ratio down to the relevant marginal rate of substitution, which is the standard requirement for Pareto optimality. However, under some conditions the economy may still be constrained Pareto optimal, meaning that a central authority limited to the same type and number of contracts as the individual agents may not be able to improve upon the outcome - what is needed is the introduction of a full set of possible contracts. Hence, one implication of the theory of incomplete markets is that inefficiency may be a result of underdeveloped financial institutions or credit constraints faced by some members of the public. Research still continues in this area.

2.6.2 Properties and characterization of general equilibrium

See also: Fundamental theorems of welfare economics

Basic questions in general equilibrium analysis are concerned with the conditions under which equilibrium will be efficient, which efficient equilibria can be achieved, when equilibrium is guaranteed to exist and when the equilibrium will be unique and stable.

First Fundamental Theorem of Welfare Economics

The first fundamental welfare theorem asserts that market equilibria are Pareto efficient. In a pure exchange economy, a sufficient condition for the first welfare theorem to hold is that preferences be locally no satiated. The first welfare theorem also holds for economies with production regardless of the properties of the production function. Implicitly, the theorem assumes complete markets and perfect information. In an economy with externalities, for example, it is possible for equilibria to arise that are not efficient.

The first welfare theorem is informative in the sense that it points to the sources of inefficiency in markets. Under the assumptions above, any market equilibrium is tautologically efficient. Therefore, when equilibria arise that are not efficient, the market system itself is not to blame, but rather some sort of market failure.

Second Fundamental Theorem of Welfare Economics

While every equilibrium is efficient, it is clearly not true that every efficient allocation of resources will be equilibrium. However, the Second Theorem states that every efficient allocation can be supported by some set of prices. In other words all that is required to reach a particular outcome is a redistribution of initial endowments of the agents after which the market can be left alone to do its work. This suggests that the issues of
efficiency and equity can be separated and need not involve a trade off. However, the conditions for the Second Theorem are stronger than those for the First, as now we need consumers' preferences to be convex (convexity roughly corresponds to the idea of diminishing rates of marginal substitution, or to preferences where "averages are better than extrema").

**Existence**

Even though every equilibrium is efficient, neither of the above two theorems say anything about the equilibrium existing in the first place. To guarantee that equilibrium exists we once again need consumer preferences to be convex (although with enough consumers this assumption can be relaxed both for existence and the Second Welfare Theorem). Similarly, but less plausibly, feasible production sets must be convex, excluding the possibility of economies of scale.

Proofs of the existence of equilibrium generally rely on fixed point theorems such as Brouwer fixed point theorem or its generalization, the Kakutani fixed point theorem. In fact, one can quickly pass from a general theorem on the existence of equilibrium to Brouwer’s fixed point theorem. For this reason many mathematical economists consider proving existence a deeper result than proving the two Fundamental Theorems.

**Uniqueness**

Although generally (assuming convexity) an equilibrium will exist and will be efficient the conditions under which it will be unique are much stronger. While the issues are fairly technical the basic intuition is that the presence of wealth effects (which is the feature that most clearly delineates general equilibrium analysis from partial equilibrium) generates the possibility of multiple equilibria. When a price of a particular good changes there are two effects. First, the relative attractiveness of various commodities changes, and second, the wealth distribution of individual agents is altered. These two effects can offset or reinforce each other in ways that make it possible for more than one set of prices to constitute an equilibrium.

A result known as the Sommensein-Mantel-Debreu Theorem states that the aggregate (excess) demand function inherits only certain properties of individual's demand functions, and that these (Continuity, Homogeneity of degree zero, Walras' law, and boundary behavior when prices are near zero) are not sufficient to restrict the admissible aggregate excess demand function in a way which would ensure uniqueness of equilibrium.

There has been much research on conditions when the equilibrium will be unique, or which at least will limit the number of equilibria. One result states that under mild assumptions the number of equilibria will be finite (see Regular economy) and odd (see Index Theorem). Furthermore if an economy as a whole, as characterized by an aggregate excess demand function, has the revealed preference property (which is a much stronger condition than revealed preferences for a single individual) or the gross substitute
property then likewise the equilibrium will be unique. All methods of establishing uniqueness can be thought of as establishing that each equilibrium has the same positive local index, in which case by the index theorem there can be but one such equilibrium.

**Determinacy**

Given that equilibria may not be unique, it is of some interest to ask whether any particular equilibrium is at least locally unique. If so, then comparative statics can be applied as long as the shocks to the system are not too large. As stated above, in a Regular economy equilibria will be finite, hence locally unique. One reassuring result, due to Debreu, is that "most" economies are regular. However recent work by Michael Mandler (1999) has challenged this claim. The Arrow-Debreu-McKenzie model is neutral between models of production functions as continuously differentiable and as formed from (linear combinations of) fixed coefficient processes. Mandler accepts that, under either model of production, the initial endowments will not be consistent with a continuum of equilibria, except for a set of Lebesgue measure zero. However, endowments change with time in the model and this evolution of endowments is determined by the decisions of agents (e.g., firms) in the model. Agents in the model have an interest in equilibria being indeterminate:

"Indeterminacy, moreover, is not just a technical nuisance; it undermines the price-taking assumption of competitive models. Since arbitrary small manipulations of factor supplies can dramatically increase a factor's price, factor owners will not take prices to be parametric." (Mandler 1999, p. 17)

When technology is modeled by (linear combinations) of fixed coefficient processes, optimizing agents will drive endowments to be such that a continuum of equilibria exist:

"The endowments where indeterminacy occurs systematically arise through time and therefore cannot be dismissed; the Arrow-Debreu-McKenzie model is thus fully subject to the dilemmas of factor price theory." (Mandler 1999, p. 19)

Critics of the general equilibrium approach have questioned its practical applicability based on the possibility of non-uniqueness of equilibria. Supporters have pointed out that this aspect is in fact a reflection of the complexity of the real world and hence an attractive realistic feature of the model.

**Stability**

In a typical general equilibrium model the prices that prevail "when the dust settles" are simply those that coordinate the demands of various consumers for various goods. But this raises the question of how these prices and allocations have been arrived at and whether any (temporary) shock to the economy will cause it to converge back to the same outcome that prevailed before the shock. This is the question of stability of the equilibrium, and it can be readily seen that it is related to the question of uniqueness. If there are multiple equilibria, then some of them will be unstable. Then, if an equilibrium is unstable and there is a shock, the economy will wind up at a different set of allocations
and prices once the convergence process terminates. However stability depends not only on the number of equilibria but also on the type of the process that guides price changes (for a specific type of price adjustment process see Tatonnement). Consequently some researchers have focused on plausible adjustment processes that guarantee system stability, i.e., that guarantee convergence of prices and allocations to some equilibrium. However, when more than one stable equilibrium exists, where one ends up will depend on where one begins.

### 2.6.3 Computing general equilibrium

Until the 1970s, general equilibrium analysis remained theoretical. However, with advances in computing power, and the development of input-output tables, it became possible to model national economies, or even the world economy, and attempts were made to solve for general equilibrium prices and quantities empirically.

**Applied general equilibrium** (AGE) models were pioneered by Herbert Scarf in 1967, and offered a method for solving the Arrow-Debreu General Equilibrium system in a numerical fashion. This was first implemented by John Shoven and John Whalley (students of Scarf at Yale) in 1972 and 1973, and was a popular method up through the 1970's. In the 1980's however, AGE models faded from popularity due to their inability to provide a precise solution and its high cost of computation. Also, Scarf's method was proven non-computable to a precise solution by Velupillai (2006). (See AGE model article for the full references)

**Computable general equilibrium** (CGE) models surpassed and replaced AGE models in the mid 1980s, as the CGE model was able to provide relatively quick and large computable models for a whole economy, and was the preferred method of governments and the World Bank. CGE models are heavily used today, and while 'AGE' and 'CGE' is used inter-changeably in the literature, Scarf type AGE models have not been constructed since the mid 1980's, and the CGE literature at current is not based on Arrow-Debreu and General Equilibrium Theory as discussed in this article. CGE models, and what is today referred to as AGE models, are based on static, simultaneously solved, macro balancing equations (from the standard Keynesian macro model), giving a precise and explicitly computable result (Mitra-Kahn 2008).

### 2.7 PARTIAL EQUILIBRIUM

A **partial equilibrium** is a type of economic equilibrium, where the clearance on the market of some specific goods is obtained independently from prices and quantities demanded and supplied in other markets. In other words, the prices of all substitutes and complements, as well as income levels of consumers are constant. Here the dynamic process is that prices adjust until supply equals demand. It is a powerfully simple technique that allows one to study equilibrium, efficiency and comparative statics.

Partial equilibrium (PE) analysis of one market always makes the assumption that all relevant variables except the price in question are constant. Thus, we assume that the
prices of all substitutes and complements, as well as income levels of consumers are constant. We can thus focus on the impact a tax in one market, without worrying about second round effects from other markets.

In many cases PE doesn’t make sense, but in others it is a very useful tool to anticipate the main effects of a policy. Generally, the more narrowly we define a market, the more appropriate PE analysis is. So, PE of a tax on pencils is normally viewed as more reasonable than PE analysis of a broadbased tax like the GST. In the former case, it’s probably unlikely that a tax on pencils will cause significant changes in the price of substitutes (pens and printers) or inputs (graphite or wood), so the maintained assumption that no other prices change in response to the policy is likely a good one. Note that in the case of a broad-based tax like the GST, such an assumption is not reasonable.

Assumptions

In what follows, we will make a lot of strong assumptions to make the analysis clearer (and normally simpler too). These assumptions include:

• The market under consideration is for a private good (like clothing or food or paper) and there is no trade (imports or exports).

• All product and factor markets are perfectly competitive.

• Production exhibits non-increasing returns to scale.

• There are no externalities associated with the production or consumption of the goods.

• There is no government intervention of any kind.

These assumptions mean that:

1. The supply curve can be interpreted as the marginal social cost curve.

2. The demand curve can be interpreted as the marginal social benefit curve.

2.8 RECENT DEVELOPMENTS IN APPLIED DEMAND ANALYSIS

A technical evaluation of goods or services on the basis of factors affecting the supply of and demand for a particular product or service in general. Supply-demand analysis is supposed to determine if an imbalance exists or will exist between supply and demand for those goods or services. For example, if the supply of a security is expected to exceed demand, the security should be sold or not purchased because its price can be expected to decline. Supply-demand analysis mainly incorporates information on new stock
offerings, government borrowing, and contributions to pension funds, mutual fund cash balances, and a number of other similar factors.

Classical approaches to analyze the demand are no doubt have immortal impact in microeconomic concepts. But every theory and approach has certain obvious drawbacks and limitations.

Since the economic scenario of world is changing at the fast speed, the need is felt to make some developments in existing patterns or demand analysis. Some of the recent developments in demand analysis are being discussed as follows:

- New simple data-analytic techniques for analyzing consumption data;
- New tests based on computer simulations for hypothesis testing;
- New methodological results on how to estimate demand equations;
- Innovative applications to alcohol demand;
- Extension of the system-wide framework to analyze the effects of advertising on consumption;
- And novel approaches to forecasting consumption patterns; and new theoretical results on aggregation of demand systems over consumers.

**Activity 2**

1. Explain how supply and demand interact to determine equilibrium price and output?
2. Discuss the assumptions behind demand and supply Model.
3. What do you understand by aggregate demand? How it is related to aggregate supply
4. Distinguish between general and partial equilibrium approaches.

**2.8 SUMMARY**

Supply-demand analysis is supposed to determine if an imbalance exists or will exist between supply and demand for a good or service. The model of demand and supply has discussed using suitable examples. Aggregate supply and aggregate demand gives insight into the adjustment process. The supply curve indicates how many producers will supply the product (or service) of interest at a particular price. Similarly, the demand curve indicates how many consumers will buy the product at a given price. General equilibrium theory seeks to explain the behavior of supply, demand and prices in a whole economy with several or many markets. A partial equilibrium is a type of economic equilibrium, where the clearance on the market of some specific goods is obtained independently from prices and quantities demanded and supplied in other markets. After discussing the given
topics the brief discussion on recent developments in demand analysis had given in the unit.

2.9 FURTHER READINGS

UNIT 3

THEORIES OF DEMAND

Objectives

After studying this unit, you should know the:

- Basic approaches to the concept of demand
- The theory of utility
- The relevance of Income and Substitution Effect
- The concept of indifference curve and revealed Preference
- The slutsky theorem and its relevance in real world situations
- The Hicks theory and its assumptions
- Price Formation and Discovery

Structure

3.1 Introduction
3.2 The Utility Theory
3.3 Income and substitution Effect
3.4 Indifference Curve
3.5 Revealed Preference
3.6 The Slutsky Theorem
3.7 The Hicks Theory
3.8 Price Formation and Discovery
3.9 Summary
3.10 Further Readings

3.1 INTRODUCTION

Demand is perhaps one of the most fundamental concepts of economics and it is the backbone of a market economy. Demand refers to how much (quantity) of a product or service is desired by buyers. The quantity demanded is the amount of a product people are willing to buy at a certain price; the relationship between price and quantity demanded is known as the demand relationship.

In this unit we are going to discuss certain main approaches to Demand and relative theories to it. Further some important aspects that directly or indirectly affect the demand will also be discussed.
3.2 THE UTILITY THEORY

In economics, utility is a measure of the relative satisfaction from, or desirability of, consumption of various goods and services. Given this measure, one may speak meaningfully of increasing or decreasing utility, and thereby explain economic behavior in terms of attempts to increase one's utility. For illustrative purposes, changes in utility are sometimes expressed in units called utils.

The doctrine of utilitarianism saw the maximization of utility as a moral criterion for the organization of society. According to utilitarians, such as Jeremy Bentham (1748-1832) and John Stuart Mill (1806-1876), society should aim to maximize the total utility of individuals, aiming for "the greatest happiness for the greatest number of people". Another theory forwarded by John Rawls (1921-2002) would have society maximize the utility of the individual receiving the minimum amount of utility.

In neoclassical economics, rationality is precisely defined in terms of imputed utility-maximizing behavior under economic constraints. As a hypothetical behavioral measure, utility does not require attribution of mental states suggested by "happiness", "satisfaction", etc.

Utility can be applied by economists in such constructs as the indifference curve, which plots the combination of commodities that an individual or a society would accept to maintain a given level of satisfaction. Individual utility and social utility can be construed as the dependent variable of a utility function (such as an indifference curve map) and a social welfare function respectively. When coupled with production or commodity constraints, these functions can represent Pareto efficiency, such as illustrated by Edgeworth boxes in contract curves. Such efficiency is a central concept of welfare economics.

3.2.1 Cardinal and ordinal utility

Economists distinguish between cardinal utility and ordinal utility. When cardinal utility is used, the magnitude of utility differences is treated as an ethically or behaviorally significant quantity. On the other hand, ordinal utility captures only ranking and not strength of preferences. An important example of a cardinal utility is the probability of achieving some target.

Utility functions of both sorts assign real numbers ("utils") to members of a choice set. For example, suppose a cup of orange juice has utility of 120 utils, a cup of tea has a utility of 80 utils, and a cup of water has a utility of 40 utils. When speaking of cardinal utility, it could be concluded that the cup of orange juice is better than the cup of tea by exactly the same amount by which the cup of tea is better than the cup of water. One is not entitled to conclude, however, that the cup of tea is two thirds as good as the cup of juice, because this conclusion would depend not only on magnitudes of utility differences, but also on the "zero" of utility.
It is tempting when dealing with cardinal utility to aggregate utilities across persons. The argument against this is that interpersonal comparisons of utility are suspect because there is no good way to interpret how different people value consumption bundles.

When ordinal utilities are used, differences in utils are treated as ethically or behaviorally meaningless: the utility values assigned encode a full behavioral ordering between members of a choice set, but nothing about strength of preferences. In the above example, it would only be possible to say that juice is preferred to tea to water, but no more.

Neoclassical economics has largely retreated from using cardinal utility functions as the basic objects of economic analysis, in favor of considering agent preferences over choice sets. As will be seen in subsequent sections, however, preference relations can often be rationalized as utility functions satisfying a variety of useful properties.

Ordinal utility functions are equivalent up to monotone transformations, while cardinal utilities are equivalent up to positive linear transformations.

### 3.2.2 Utility functions

While preferences are the conventional foundation of microeconomics, it is often convenient to represent preferences with a utility function and reason indirectly about preferences with utility functions. Let X be the consumption set, the set of all mutually-exclusive packages the consumer could conceivably consume (such as an indifference curve map without the indifference curves). The consumer's utility function \( u : X \rightarrow \mathbb{R} \) ranks each package in the consumption set. If \( u(x) \geq u(y) \), then the consumer strictly prefers x to y or is indifferent between them.

For example, suppose a consumer's consumption set is \( X = \{ \text{nothing}, 1 \text{ apple}, 1 \text{ orange}, 1 \text{ apple and 1 orange}, 2 \text{ apples}, 2 \text{ oranges} \} \), and its utility function is \( u(\text{nothing}) = 0 \), \( u(1 \text{ apple}) = 1 \), \( u(1 \text{ orange}) = 2 \), \( u(1 \text{ apple and 1 orange}) = 4 \), \( u(2 \text{ apples}) = 2 \) and \( u(2 \text{ oranges}) = 3 \). Then this consumer prefers 1 orange to 1 apple, but prefers one of each to 2 oranges.

In microeconomic models, there are usually a finite set of L commodities, and a consumer may consume an arbitrary amount of each commodity. This gives a consumption set of \( \mathbb{R}_+^L \), and each package \( x \in \mathbb{R}_+^L \) is a vector containing the amounts of each commodity. In the previous example, we might say there are two commodities: apples and oranges. If we say apples is the first commodity, and oranges the second, then the consumption set \( X = \mathbb{R}_+^2 \) and u \((0, 0) = 0\), \( u(0, 1) = 1\), \( u(1, 0) = 2\), \( u(1, 1) = 4\), \( u(2, 0) = 2\), \( u(0, 2) = 3 \) as before. Note that for u to be a utility function on X, it must be defined for every package in X.
A utility function \( u : X \to \mathbb{R} \) rationalizes a preference relation \( \preceq \) on \( X \) if for every \( x, y \in X \), \( u(x) \leq u(y) \) if and only if \( x \preceq y \). If \( u \) rationalizes \( \preceq \), then this implies \( \preceq \) is complete and transitive, and hence rational.

In order to simplify calculations, various assumptions have been made of utility functions.

- **CES** (constant elasticity of substitution, or isoelastic) utility
- **Exponential utility**
- **Quasilinear utility**
- **Homothetic utility**

Most utility functions used in modeling or theory are well-behaved. They are usually monotonic, quasi-concave, continuous and globally non-satiated. There are some important exceptions, however. For example lexicographic preferences are not continuous and hence cannot be represented by a continuous utility function.

### 3.2.3 Expected utility

Main article: Expected utility hypothesis

The expected utility theory deals with the analysis of choices among risky projects with (possibly multidimensional) outcomes.

The expected utility model was first proposed by Nicholas Bernoulli in 1713 and solved by Daniel Bernoulli in 1738 as the St. Petersburg paradox. Bernoulli argued that the paradox could be resolved if decisionmakers displayed risk aversion and argued for a logarithmic cardinal utility function.

The first important use of the expected utility theory was that of John von Neumann and Oskar Morgenstern who used the assumption of expected utility maximization in their formulation of game theory.

### 3.2.4 Additive von Neumann-Morgenstern Utility

In older definitions of utility, it makes sense to rank utilities, but not to add them together. A person can say that a new shirt is preferable to a baloney sandwich, but not that it is twenty times preferable to the sandwich.

The reason is that the utility of twenty sandwiches is not twenty times the utility of one sandwich, by the law of diminishing returns. So it is hard to compare the utility of the shirt with 'twenty times the utility of the sandwich'. But Von Neumann and Morgenstern suggested an unambiguous way of making a comparison like this.
Their method of comparison involves considering probabilities. If a person can choose between various randomized events (lotteries), then it is possible to additively compare the shirt and the sandwich. It is possible to compare a sandwich with probability 1, to a shirt with probability p or nothing with probability 1-p. By adjusting p, the point at which the sandwich becomes preferable defines the ratio of the utilities of the two options.

A notation for a lottery is as follows: if options A and B have probability p and 1-p in the lottery, write it as a linear combination:

\[ L = pA + (1-p)B \]

More generally, for a lottery with many possible options:

\[ L = \sum p_i A_i, \]

with the sum of the \( p_i \)s equalling 1.

By making some reasonable assumptions about the way choices behave, von Neumann and Morgenstern showed that if an agent can choose between the lotteries, then this agent has a utility function which can be added and multiplied by real numbers, which means the utility of an arbitrary lottery can be calculated as a linear combination of the utility of its parts.

This is called the expected utility theorem. The required assumptions are four axioms about the properties of the agent's preference relation over 'simple lotteries', which are lotteries with just two options. Writing \( B \preceq A \) to mean 'A is preferred to B', the axioms are:

1. completeness: For any two simple lotteries \( L \) and \( M \), either \( L \preceq M \), \( L = M \), or \( M \preceq L \).
2. transitivity: if \( L \preceq M \) and \( M \preceq N \), then \( L \preceq N \).
3. convexity/continuity (Archimedean property): If \( L \preceq M \preceq N \), then there is a \( p \) between 0 and 1 such that the lottery \( pL + (1-p)N \) is equally preferable to \( M \).
4. independence: if \( L = M \), then \( pL + (1-p)N = pM + (1-p)N \).

In more formal language: A von Neumann-Morgenstern utility function is a function from choices to the real numbers:

\[ u : X \to \mathbb{R} \]

which assigns a real number to every outcome in a way that captures the agent's preferences over both simple and compound lotteries. The agent will prefer a lottery \( L_2 \)
to a lottery $L_1$ if and only if the expected utility of $L_2$ is greater than the expected utility of $L_1$:

$$L_1 \leq L_2 \iff u(L_1) \leq u(L_2).$$

Repeating in category language: $u$ is a morphism between the category of preferences with uncertainty and the category of reals as an additive group.

Of all the axioms, independence is the most often discarded. A variety of generalized expected utility theories have arisen, most of which drop or relax the independence axiom.

- CES (constant elasticity of substitution, or isoelastic) utility is one with constant relative risk aversion
- Exponential utility exhibits constant absolute risk aversion

### 3.2.5 Utility of money

One of the most common uses of a utility function, especially in economics, is the utility of money. The utility function for money is a nonlinear function that is bounded and asymmetric about the origin. These properties can be derived from reasonable assumptions that are generally accepted by economists and decision theorists, especially proponents of rational choice theory. The utility function is concave in the positive region, reflecting the phenomenon of diminishing marginal utility. The boundedness reflects the fact that beyond a certain point money ceases being useful at all, as the size of any economy at any point in time is itself bounded. The asymmetry about the origin reflects the fact that gaining and losing money can have radically different implications both for individuals and businesses. The nonlinearity of the utility function for money has profound implications in decision making processes: in situations where outcomes of choices influence utility through gains or losses of money, which are the norm in most business settings, the optimal choice for a given decision depends on the possible outcomes of all other decisions in the same time-period.

### 3.2.6 Discussion and criticism

Different value systems have different perspectives on the use of utility in making moral judgments. For example, Marxists, Kantians, and certain libertarians (such as Nozick) all believe utility to be irrelevant as a moral standard or at least not as important as other factors such as natural rights, law, conscience and/or religious doctrine. It is debatable whether any of these can be adequately represented in a system that uses a utility model.

Another criticism comes from the assertion that neither cardinal nor ordinary utility are empirically observable in the real world. In case of cardinal utility it is impossible to measure the level of satisfaction "quantitatively" when someone consume/purchase an apple. In case of ordinal utility, it is impossible to determine what choice were made
when someone purchase, for example, an orange. Any act would involve preference over infinite possibility of set choices such as (apple, orange juice, other vegetable, vitamin C tablets, exercise, not purchasing, etc).

3.3 THE INCOME AND SUBSTITUTION EFFECTS

When the price of q1, p1, changes there are two effects on the consumer. First, the price of q1 relative to the other products (q2, q3, . . . qn) has changed. Second, due to the change in p1, the consumer's real income changes. When we compute the change in the optimal consumption as a result of the price change, we do not usually separate these two effects. Sometimes we might want to separate the effects.

3.3.1 The Substitution Effect

The Substitution Effect is the effect due only to the relative price change, controlling for the change in real income. In order to compute it we ask what is the bundle that would make the consumer just as happy as before the price change, but if they had to make their choice faced with the new prices. To find this point we consider a budget line characterized by the new prices but with a level of income such that it is tangent to the initial indifference curve. In the diagram on the next page, the initial consumer equilibrium is at point A where the initial budget line is tangent to the higher indifference curve. Consumption at this point is 11 units of good 1 and 8 units of good 2. After an increase in the price of good 1, the consumer moves to point E, where the new budget line is tangent to the lower indifference curve. Consumption of good 1 has fallen to 4 units while consumption of good 2 has increased to 10 units. The substitution effect is the movement from point A to point G. This point is characterized by two things. (1) It is on the same indifference curve as the original consumption bundle; AND (2) it is the point where a budget line that is parallel to the new budget line is just tangent to initial indifference curve. This "intermediate" budget line is attempting to hold real income fixed so we can isolate the substitution effect. The point G reflects the consumer's choice if faced with the new prices (the budget line has the slope reflecting the new prices) and the compensated income (i.e., an income level that holds real income fixed). The substitution effect is the difference between the original consumption and the new "intermediate" consumption. In this case consumption of good 1 falls from 11 to 6.84 while consumption of good 2 increases to 14.27.

When p1 goes up the Substitution Effect will always be non-positive (i.e., negative or zero).

3.3.2 The Income Effect

The Income Effect is the effect due to the change in real income. For example, when the price goes up the consumer is not able to buy as many bundles that she could purchase before. This means that in real terms she has become worse off. The effect is measured as
the difference between the “intermediate" consumption” at G and the final consumption of q1 and q2 at E.

Unlike the Substitution Effect, the Income Effect can be both positive and negative depending on whether the product is a normal or inferior good. By the way we constructed them, the Substitution Effect plus the Income Effect equals the total effect of the price change.

**Alternative Way of Analyzing a Price Change**

One can also analyze the income and substitution effects by first considering the income change necessary to move the consumer to the new utility level at the initial prices. This constitutes the income effect. The movement along the new indifference curve from the intermediate point to the new equilibrium as the slope of the price line changes is then the substitution effect. See if you can identify the “intermediate” point on the lower indifference curve by shifting the budget line (Hint: q1 and q2 both fall.).

**Figure 1**
3.4 THE INDIFFERENCE CURVE

In microeconomic theory, an indifference curve is a graph showing different bundles of goods, each measured as to quantity, between which a consumer is indifferent. That is, at each point on the curve, the consumer has no preference for one bundle over another. In other words, they are all equally preferred. One can equivalently refer to each point on the indifference curve as rendering the same level of utility (satisfaction) for the consumer. Utility is then a device to represent preferences rather than something from which preferences come. The main use of indifference curves is in the representation of potentially observable demand patterns for individual consumers over commodity bundles.

Map and properties of indifference curves

![Indifference Curves](image)

**Figure 2: An example of an indifference map with three indifference curves represented**

A graph of indifference curves for an individual consumer associated with different utility levels is called an indifference map. Points yielding different utility levels are each associated with distinct indifference curves. An indifference curve describes a set of personal preferences and so can vary from person to person. An indifference curve is like a contour line on a topographical map. Each point on the map represents the same elevation. If you move "off" an indifference curve traveling in a northeast direction you are essentially climbing a mound of utility. The higher you go the greater the level of utility. The non-satiation requirement means that you will never reach the "top".
Indifference curves are typically represented to be:

- 1. defined only in the positive (+, +) quadrant of commodity-bundle quantities.
- 2. negatively sloped. That is, as quantity consumed of one good (X) increases, total satisfaction would increase if not offset by a decrease in the quantity consumed of the other good (Y). Equivalently, satiation, such that more of either good (or both) is equally preferred to no increase, is excluded. (If utility $U = f(x, y)$, $U$, in the third dimension, does not have a local maximum for any $x$ and $y$ values.) The negative slope of the indifference curve reflects the law of diminishing marginal utility. That is as more of a good is consumed total utility increases at a decreasing rate - additions to utility per unit consumption are successively smaller. Thus as you move down the indifference curve you are trading consumption of units of Y for additional units of X. The price of a unit of X in terms of Y increases.
- 3. complete, such that all points on an indifference curve are ranked equally preferred and ranked either more or less preferred than every other point not on the curve. So, with (2), no two curves can intersect (otherwise non-satiation would be violated).
- 4. transitive with respect to points on distinct indifference curves. That is, if each point on $I_2$ is (strictly) preferred to each point on $I_1$, and each point on $I_3$ is preferred to each point on $I_2$, each point on $I_3$ is preferred to each point on $I_i$. A negative slope and transitivity exclude indifference curves crossing, since straight lines from the origin on both sides of where they crossed would give opposite and intransitive preference rankings.
- 5. (strictly) convex (sagging from below). With (2), convex preferences implies a bulge toward the origin of the indifference curve. As a consumer decreases consumption of one good in successive units, successively larger doses of the other good are required to keep satisfaction unchanged.

**Assumptions**

Let $a$, $b$, and $c$ be bundles (vectors) of goods, such as $(x, y)$ combinations above, with possibly different quantities of each respective good in the different bundles. The first assumption is necessary for a well-defined representation of stable preferences for the consumer as agent; the second assumption is convenient.

**Rationality** (called an ordering relationship in a more general mathematical context): Completeness + transitivity. For given preference rankings, the consumer can choose the best bundle(s) consistently among $a$, $b$, and $c$ from lowest on up.

**Continuity**: This means that you can choose to consume any amount of the good. For example, I could drink 11 mL of soda, or 12 mL, or 132 mL. I am not confined to drinking 2 liters or nothing. See also continuous function in mathematics.

Of the remaining properties above, suppose, property (5) (convexity) is violated by a bulge of the indifference curves out from the origin for a particular consumer with a
given budget constraint. Consumer theory then implies zero consumption for one of the two goods, say good Y, in equilibrium on the consumer's budget constraint. This would exemplify a corner solution. Further, decreases in the price of good Y over a certain range might leave quantity demanded unchanged at zero beyond which further price decreases switched all consumption and income away from X and to Y. The eccentricity of such an implication suggests why convexity is typically assumed.

Examples of Indifference Curves

In Figure 3, the consumer would rather be on $I_3$ than $I_2$, and would rather be on $I_2$ than $I_1$, but does not care where he/she is on a given indifference curve. The slope of an indifference curve (in absolute value), known by economists as the marginal rate of substitution, shows the rate at which consumers are willing to give up one good in exchange for more of the other good. For most goods the marginal rate of substitution is not constant so their indifference curves are curved. The curves are convex to the origin, describing the negative substitution effect. As price rises for a fixed money income, the consumer seeks less the expensive substitute at a lower indifference curve. The substitution effect is reinforced through the income effect of lower real income (Beattie-LaFrance).

If two goods are perfect substitutes then the indifference curves will have a constant slope since the consumer would be willing to trade at a fixed ratio. The marginal rate of substitution between perfect substitutes is likewise constant. An example of a utility function that is associated with indifference curves like these would be $U(x,y) = \alpha x + \beta y$.

If two goods are perfect complements then the indifference curves will be L-shaped. An example would be something like if you had a cookie recipe that called for 3 cups flour to 1 cup sugar. No matter how much extra flour you had, you still could not make more cookie dough without more sugar. Another example of perfect complements is a left shoe and a right shoe. The consumer is no better off having several right shoes if she has only one left shoe. Additional right shoes have zero marginal utility without more left shoes. The marginal rate of substitution is either zero or infinite. An example of the type of utility function that has an indifference map like that above is $U(x,y) = \min\{\alpha x, \beta y\}$.

The different shapes of the curves imply different responses to a change in price as shown from demand analysis in consumer theory. The results will only be stated here. A price-budget-line change that kept a consumer in equilibrium on the same indifference curve:

- in Fig. 3 would reduce quantity demanded of a good smoothly as price rose relatively for that good.
- in Fig. 4 would have either no effect on quantity demanded of either good (at one end of the budget constraint) or would change quantity demanded from one end of the budget constraint to the other.
- in Fig. 5 would have no effect on equilibrium quantities demanded, since the budget line would rotate around the corner of the indifference curve.
Figure 3 encore: An example of an indifference map with three indifference curves represented.

Figure 4: Three indifference curves where Goods X and Y are perfect substitutes. The gray line perpendicular to all curves indicates the curves are mutually parallel.

Figure 5: Indifference curves for perfect complements X and Y. The elbows of the curves are collinear.

3.5 REVEALED PREFERENCE
**Revealed preference theory**, pioneered by American economist Paul Samuelson, is a method by which it is possible to discern the best possible option on the basis of consumer behavior. Essentially, this means that the preferences of consumers can be revealed by their purchasing habits. Revealed preference theory came about because the theories of consumer demand were based on a diminishing marginal rate of substitution (MRS). This diminishing MRS is based on the assumption that consumers make consumption decisions based on their intent to maximize their utility. While utility maximization was not a controversial assumption, the underlying utility functions could not be measured with great certainty. Revealed preference theory was a means to reconcile demand theory by creating a means to define utility functions by observing behavior.

**Theory**

If a person chooses a certain bundle of goods (ex. 2 apples, 3 bananas) while another bundle of goods is affordable (ex. 3 apples, 2 bananas), then we say that the first bundle is revealed preferred to the second. It is then assumed that the first bundle of goods is always preferred to the second. This means that if the consumer ever purchases the second bundle of goods then it is assumed that the first bundle is unaffordable. This implies that preferences are transitive. In other words if we have bundles A, B, C, ..., Z, and A is revealed preferred to B which is revealed preferred to C and so on then it is concluded that A is revealed preferred to C through Z. With this theory economists can chart indifference curves which adhere to already developed models of consumer theory.

**The Weak Axiom of Revealed Preference**

The Weak Axiom of Revealed Preference (WARP) is a characteristic on the choice behavior of an economic agent. For example, if an individual chooses A and never B when faced with a choice of both alternatives, they should never choose B when faced with a choice of A, B and some additional options. More formally, if A is ever chosen when B is available, revealed preference theory, pioneered by American economist Paul Samuelson, is a method by which it is possible to discern the best possible option on the basis of consumer behavior. Essentially, this means that the preferences of consumers can be revealed by their purchasing habits. Revealed preference theory came about because the theories of consumer demand were based on a diminishing marginal rate of substitution (MRS). This diminishing MRS is based on the assumption that consumers make consumption decisions based on their intent to maximize their utility. While utility maximization was not a controversial assumption, the underlying utility functions could not be measured with great certainty. Revealed preference theory was a means to reconcile demand theory by creating a means to define utility functions by observing behavior.

If a person chooses a certain bundle of goods (ex. 2 apples, 3 bananas) while another bundle of goods is affordable (ex. 3 apples, 2 bananas), then we say that the first bundle is revealed preferred to the second. It is then assumed that the first bundle of goods is always preferred to the second. This means that if the consumer ever purchases the
second bundle of goods then it is assumed that the first bundle is unaffordable. This implies that preferences are transitive. In other words if we have bundles A, B, C, ..., Z, and A is revealed preferred to B which is revealed preferred to C and so on then it is concluded that A is revealed preferred to C through Z. With this theory economists can chart indifference curves which adhere to already developed models of consumer theory.

then there can be no optimal set containing both alternatives for which B is chosen and A is not.

This characteristic can be stated as a characteristic of Walrasian demand functions as seen in the following example. Let \( p_a \) be the price of apples and \( p_b \) be the price of bananas, and let the amount of money available be \( m=5 \). If \( p_a =1 \) and \( p_b=1 \), and if the bundle (2,3) is chosen, it is said that that the bundle (2,3) is revealed preferred to (3,2), as the latter bundle could have been chosen as well at the given prices. More formally, assume a consumer has a demand function \( x \) such that they choose bundles \( x(p,w) \) and \( x(p',w') \) when faced with price-wealth situations \( (p,w) \) and \( (p',w') \) respectively. If \( p \cdot x(p',w') \leq w \) then the consumer chooses \( x(p,w) \) even when \( x(p',w') \) was available under prices \( p \) at wealth \( w \), so \( x(p,w) \) must be preferred to \( x(p',w') \).

### 3.6 SLUTSKY THEOREM

The Slutsky's Equation breaks down a change in demand due to price change into the substitution effect and the income effect. The equation takes the form:

\[
\frac{dx}{dp} = \frac{dh}{dp} - x \frac{dx}{dm}
\]

The term on the left is the change in demand when price changes, where \( x \) is the (Marshallian) demand for a good and \( p \) is the price. The term \( h \) is the Hicksian or the compensated demand. The term \( \frac{dh}{dp} \) measures the substitution effect. The term \( m \) is the income, and \( x \cdot \frac{dx}{dm} \) measures the income effect. See below for more explanations and the derivation of the equation.

We can make sense of the substitution and the income effects by this intuitive story. Suppose a consumer is consuming the optimal amount of two goods \( x \) and \( y \), given his income and suddenly the price of \( x \) drops. The consumer will respond to this price change in two ways. First, as \( x \) becomes relatively cheaper the consumer will shift some of his consumption of \( y \) to \( x \) (assume \( x \) and \( y \) are not perfect complement). Second, as the price of \( x \) drops, even if the consumer does not make any consumption shift from \( y \) to \( x \), he has more purchasing power because of the savings that results from the price drop in \( x \). This savings allows the consumer to buy more goods (\( x \) or \( y \)). The shift in consumption from \( y \) to \( x \) is the substitution effect, and the increase in purchasing power due to the savings is the income effect.

When we read the Slutsky's equation, the term \( \frac{dh}{dp} \) is the substitution effect. This is because the compensated depend \( h(p_1, p_2, u) \) fixes the consumer's utility level, and when
the consumer's purchasing power remains constant, the term \( \frac{dh}{dp} \) only measures the shift in consumption when the price changes. On the other hand, the income effect depends on the amount of good the consumer is consuming \((x)\), and the consumer's reaction to an income change \((\frac{dx}{dm})\) that comes from the "savings". Thus the term \( x \times \frac{dx}{dm} \) measures the income effect.

**Derivation**

Notice that in equilibrium, the (Marshallian) demand and the compensated demand are the same. That is, \( x(p_1, p_2, m) = h(p_1, p_2, v(p_1, p_2, m)) \), where \( v \) is the value function of the utility maximization problem. To simplify notation, we write \( u = v(p_1, p_2, m) \), a fixed level of utility, and we write the budget constraint as \( p_1x_1 + p_2x_2 - m = 0 \).

Now equate the two demands as above,

\[ x(p_1, p_2, m) = h(p_1, p_2, u) \]

Without loss of generality, differentiate with respect to \( p_1 \),

\[ \frac{dx}{dp_1} + \frac{dx}{dm} \frac{dm}{dp_1} = \frac{dh}{dp_1} \]

Note that the budget constraint in the Marshallian demand depends on \( p \), so we have to use total derivative when differentiating the left side of the equation. The second term is merely an application of the chain rule. The term \( \frac{dm}{dp_1} \) is the derivative of the budget constraint \( p_1x_1 + p_2x_2 - m \) with respect to \( p_1 \), ie, \( \frac{dm}{dp_1} = x_1 \). Substitute this in and the equation becomes the Slutsky's equation

\[ \frac{dx}{dp_1} + \frac{dx}{dm} x_1 = \frac{dh}{dp_1} \]

**Endowment income effect**

When the consumer is endowed with the goods instead of a fixed income, the budget constraint is \( p_1x_1 + p_2x_2 - p_1w_1 - p_2w_2 = 0 \). Writing the budget constraint this way, and by differentiating the budget constraint with respect to \( p_1 \), it is easy to see that the Slutsky's equation becomes

\[ \frac{dx}{dp_1} = \frac{dh}{dp_1} - \frac{dx}{dm} (x_1 - w_1) \]

In other words, the substitution effect remains the same, but the income effect applies to the excess demand rather than the demand itself.

**Other Slutsky equations**

Given the two examples and the derivation above, we can see that the Slutsky's equation always has the same format, and each format is different only because the budget constraint is different. Students can try deriving Slutsky's equations for other situations, such as one with labour supply, or intertemporal choices.

### 3.7 THE HICK’S THEORY
Hicks’s work in this area was first published in 1934 in an article he wrote jointly with R. G. D. Allen (1934a), but it is more comprehensively set out in Chapters 1-3 of Value and Capital (1939a). Their work in this area was a development of earlier work by Edgeworth, Pareto and Slutsky, and they established that ordinal utility (elaborated in terms of indifference curves and budget lines) could derive the same propositions as cardinal utility (elaborated in terms of measurable marginal utilities) but that the former achieved the same results more clearly and more precisely. However, they accepted that the two theories were saying the same thing, e.g. ‘Tangency between the price line and an indifference curve is the expression … of the proportionality between marginal utilities and prices’ (Hicks, 1939a, p. 17). This can be demonstrated conveniently as in Figure 2. The slope of indifference curve, $I$, can be written as $dY/dX$ and measures the marginal rate of substitution (MRS) of X for Y. However, along an indifference curve the total utility of the consumer is constant. Therefore it is possible to demonstrate that the slope of an indifference curve can also be defined as the ratio of the marginal utilities (MU) of X for Y.2 Thus the slope of indifference curve, $I = dY/dX = MU_x/MU_y = MRS$ of X for Y. However, the slope of the budget line, $AB = Px/Py$. Therefore at the point of tangency $MU_x/MU_y = Px/Py = MRS$ of X for Y, i.e. the ordinal and cardinal conditions for a maximum are equivalent. It is also of interest to note the similarities between this analysis and our previous discussion of isoquants. The elasticity of substitution is a property of an isoquant but the same general principle holds for an indifference curve. Specifically the marginal rate of substitution diminishes along a convex indifference curve and the marginal rate of technical substitution (MRTS) diminishes along a convex isoquant. Figure 2 can then be relabelled to demonstrate the cost minimising output level, i.e. measuring capital along the Y axis and labour along the X axis, cost minimisation occurs where $PL/Pk = MRTS$ of L for K. In fact, it was the realisation of this symmetry that led Hicks in the direction of the 1934 Hicks-Allen article.

Figure 6
Apart from considerations of presentation and clarity the major advance which indifference curve analysis offered was its ability to distinguish between the income and substitution effects of a price change. Moreover, Hicks and Allen pointed out that the
sign of this income effect could not be predicted from the simple assumption that the consumer seeks to maximise utility. However, they demonstrated that in the absence of income effects the substitution effects were remarkably regular. For example, they were able to show that not only is the direct substitution effect of a change in the price of X on the quantity demanded of X always negative, but that a number of secondary substitution theorems could also be deduced – the most important being the proposition that the substitution effect of a change in the price of X on the quantity demanded of Y must be exactly equal to the effect of a change in the price of Y on the quantity demanded of X. The latter substitution theorem refers to the cross elasticity of demand and its importance derives from the fact that it successfully cleared up the elasticity problem that had initially prompted the Hicks-Allen article. Specifically, Henry Schultz had estimated some cross elasticities of demand (the demand for X against the price of Y and vice versa) and found them to be non-symmetric whereas Marshallian demand theory suggested they should be symmetric. Hicks and Allen were then able to point out that ‘Schultz had left out the income effects which for direct elasticities may indeed be negligible, as Marshall (in effect) supposed them to be; but for cross elasticities there is no reason why they should be negligible’ (Hicks, 1974b).

This quotation is important in another respect. Basically, Marshall had ignored the income effect by assuming a constant marginal utility of money. The quotation suggests that from the point of view of ordinary demand analysis Hicks accepted Marshall’s neglect of the income effect. Indeed, such a neglect is justifiable as long as the good in question represents a small part of the consumer’s budget and Marshall was always careful to make this assumption. Thus in Hicks’s view the ordinal approach ‘was not so clear an advance (on the older marginal utility approach) as is usually supposed’ (1976, p. 137). By the same token Hicks has never accepted the extension of his analysis via Samuelson’s revealed preference approach as either necessary or desirable. ‘Marshall’s consumer who decides on his purchase by comparing the marginal utility of what is to be bought with the marginal utility of the money he will have to pay for it is more like an actual consumer’ (Hicks, 1976, p. 138). In fact in his Revision of Demand Theory (1956a) Hicks rejects the revealed preference approach even for econometric purposes, stating that for the data to which econometrics is usually applied on ordinal scale of preferences seems the most sensible hypothesis with which to begin the analysis. However, despite his own reservations, Hicks’s ordinal approach in the end met with very little criticism and quickly became a standard part of the economist’s tool kit. Moreover, many observers would claim that Hicks has underestimated the step forward that his use of indifference curves entailed, e.g. Blaug explains that European marginal utility doctrine around the First World War proliferated ‘in subtle distinctions and metaphysical classifications. [Thus], one is made to realize how much has been swept away by the Hicksian Revolution – all to the good we would say’ (Blaug, 1976, p. 388).

3.8 PRICE FORMATION AND DISCOVERY

This factor focuses on the process by which the price for an asset is determined. For example, in some markets prices are formed through an auction process (e.g. eBay), in
other markets prices are negotiated (e.g., new cars) or simply posted (e.g. local supermarket) and buyers can choose to buy or not.

Assume a representative firm in an imperfectly competitive market, producing a single good for which imperfect substitutes are produced abroad. The planned price of the good is determined as a mark-up over marginal costs

\[ pr = \mu + mc, \quad (2.2) \]

where \( \mu \) is the mark-up and \( mc \) are marginal costs. The mark-up is not necessarily constant and may be a function of relative prices. This allows for a pricing-to-market effect, with the mark-up inversely related to the elasticity of demand (see, for example, Krugman, 1987)

\[ \mu = \kappa + \lambda(q - pr) + \xi_p, \quad \kappa, \lambda \geq 0, \quad (2.3) \]

where \( q \) is the domestic currency price of imperfect substitute tradeable goods produced abroad and \( \lambda \) reflects the exposure of domestic firms to foreign demand. Thus, the greater the pricing-to-market effect (smaller \( \lambda \)) the less is the passthrough from foreign price or exchange rate shocks to domestic prices. \( \xi_p \) includes other terms that affect the mark-up, such as indirect taxes, the marginal cost of capital and changes in the market power of firms that are not reflected in the pricing-to-market effect. Production technology is given by a simple Cobb-Douglas production function, with constant returns to scale

\[ y = \log \zeta + \gamma n + (1 - \gamma)k, \quad 0 < \gamma < 1, \quad (2.4) \]

where \( y \) is value added output, \( \zeta \) is total factor productivity, \( n \) is total hours of work and \( k \) is capital input. Assuming that firms maximise profits with respect to labour inputs gives marginal costs as

\[ mc = w - z - \log \gamma, \quad (2.5) \]

where \( z = (y - n) \). Substituting (2.3) and (2.5) into (2.2) gives the planned producer price level as a mark-up over unit labour cost and import prices, with prices homogenous in both these arguments

\[ p_r = \left( \frac{\kappa - \log \gamma + \xi_p}{1 + \lambda} \right) + \left( \frac{1}{1 + \lambda} \right) (w - z) + \left( \frac{\lambda}{1 + \lambda} \right) q. \quad (2.6) \]

Activity 3

1. discuss briefly the theories of demand

2. What do you mean by revealed preference? discuss its relevance in real world situations
3. What the Indifference curve contribute to the concept of demand? Discuss giving suitable examples.


5. Assume a person has a utility function $U = XY$, and money income of $10,000, facing an initial price of $X$ of $10$ and price of $Y$ of $15$. If the price of $X$ increases to $15$, answer the following questions:

   a. What was the initial utility maximizing quantity of $X$ and $Y$?

   b. What is the new utility maximizing quantity of $X$ and $Y$ following the increase in the price of $X$?

   c. What is the Hicks compensating variation in income that would leave this person equally well off following the price increase? What is the Slutsky compensating variation in income?

D. Calculate the pure substitution effect and the real income effect on $X$ of this increase in the price of $X$. Distinguish between the calculation of these effects using the Hicksian analysis vs. the Slutsky analysis.

### 3.9 SUMMARY

In market economy, demand theories allocate resources in the most efficient way possible. Where utility theory underlies the assumption of desire and consumption of various goods and services, indifference curve shows different bundles of goods, each measured as to quantity, between which a consumer is indifferent. The Substitution Effect is the effect due only to the relative price change, controlling for the change in real income. The Income Effect is the effect due to the change in real income. Revealed preference means that the preferences of consumers can be revealed by their purchasing habits. Similarly theories evolved by Slutsky and Hicks have also discussed in depth in this unit.

### 3.10 FURTHER READINGS

• Nash Jr., John F. The Bargaining Problem. Econometrica 1950
M.A. PREVIOUS ECONOMICS
PAPER I
MICRO ECONOMIC ANALYSIS
BLOCK 2
THEORIES OF PRODUCTION AND COSTS
PAPER I
MICRO ECONOMIC ANALYSIS

BLOCK 2
THEORIES OF PRODUCTION AND COSTS

CONTENTS

<table>
<thead>
<tr>
<th>Unit 1</th>
<th>Production Function</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit 2</td>
<td>Economies of Scale and Cost Analysis</td>
<td>27</td>
</tr>
<tr>
<td>Unit 3</td>
<td>Elasticity of Substitution and related aspects of Production function</td>
<td>49</td>
</tr>
</tbody>
</table>
The purpose of this block is to introduce you to the various theories of production and costs. The block has three units.

The first unit analyses the production function. It begins with explaining the basic concepts of production function. Law of variable proportions and returns to scale along with criticism to Production Function are explained in the subsequent sections.

The second unit deals with economies of scale and cost analysis. Overview to Economies of Scale and marginal rate of technical substitution are discussed in detail. Isoquant Analysis and returns to factors are also dealt in detail.

The third and the last unit is concerned with Elasticity of Substitution and related aspects of production function. Measuring the substitutability; Cobb-Douglas production function; Constant elasticity of substitution (CES); Elasticity of substitution in multi input cases; Euler’s theorem; Variable elasticity of substitution (VES) and Technical progress and production function are the main areas that are being discussed in this unit.
UNIT 1

PRODUCTION FUNCTION

Objectives

After studying this unit you should be able to:

- Define the production function.
- Understand the basic concepts production function.
- Analyze the approach to law of variable proportions.
- Have the knowledge of returns to scale and its various aspects.

Structure

1.1 Introduction
1.2 Basic concepts of production function
1.3 Law of variable proportions
1.4 Returns to scale
1.5 Criticism to Production Function
1.6 Summary
1.7 Further readings

1.1 INTRODUCTION

The production function relates the output of a firm to the amount of inputs, typically capital and labor. It is important to keep in mind that the production function describes technology, not economic behavior. A firm may maximize its profits given its production function, but generally takes the production function as a given element of that problem. (In specialized long-run models, the firm may choose its capital investments to choose among production technologies.)

In economics, a production function is a function that specifies the output of a firm, an industry, or an entire economy for all combinations of inputs. A meta-production function (sometimes metaproduction function) compares the practice of the existing entities converting inputs \( X \) into output \( y \) to determine the most efficient practice production function of the existing entities, whether the most efficient feasible practice production or the most efficient actual practice production. In either case, the maximum output of a technologically-determined production process is a mathematical function of input factors of production. Put another way, given the set of all technically feasible combinations of output and inputs, only the combinations encompassing a maximum output for a specified set of inputs would constitute the production function. Alternatively, a production function can be defined as the specification of the minimum input requirements needed to produce designated quantities of output, given available
technology. It is usually presumed that unique production functions can be constructed for every production technology.

By assuming that the maximum output technologically possible from a given set of inputs is achieved, economists using a production function in analysis are abstracting away from the engineering and managerial problems inherently associated with a particular production process. The engineering and managerial problems of technical efficiency are assumed to be solved, so that analysis can focus on the problems of allocative efficiency. The firm is assumed to be making allocative choices concerning how much of each input factor to use, given the price of the factor and the technological determinants represented by the production function. A decision frame, in which one or more inputs are held constant, may be used; for example, capital may be assumed to be fixed or constant in the short run, and only labour variable, while in the long run, both capital and labour factors are variable, but the production function itself remains fixed, while in the very long run, the firm may face even a choice of technologies, represented by various, possible production functions.

The relationship of output to inputs is non-monetary, that is, a production function relates physical inputs to physical outputs, and prices and costs are not considered. But, the production function is not a full model of the production process: it deliberately abstracts away from essential and inherent aspects of physical production processes, including error, entropy or waste. Moreover, production functions do not ordinarily model the business processes, either, ignoring the role of management, of sunk cost investments and the relation of fixed overhead to variable costs. (For a primer on the fundamental elements of microeconomic production theory, see production theory basics).

The primary purpose of the production function is to address allocative efficiency in the use of factor inputs in production and the resulting distribution of income to those factors. Under certain assumptions, the production function can be used to derive a marginal product for each factor, which implies an ideal division of the income generated from output into an income due to each input factor of production.

**Production function as an equation**

There are several ways of specifying the production function.

In a general mathematical form, a production function can be expressed as:

\[ Q = f(X_1, X_2, X_3, \ldots, X_n) \]

where:

- \( Q \) = quantity of output
- \( X_1, X_2, X_3, \ldots, X_n \) = factor inputs (such as capital, labour, land or raw materials).

This general form does not encompass joint production, that is a production process, which has multiple co-products or outputs.
One way of specifying a production function is simply as a table of discrete outputs and input combinations, and not as a formula or equation at all. Using an equation usually implies continual variation of output with minute variation in inputs, which is simply not realistic. Fixed ratios of factors, as in the case of laborers and their tools, might imply that only discrete input combinations, and therefore, discrete maximum outputs, are of practical interest.

One formulation is as a linear function:

\[ Q = a + bX_1 + cX_2 + dX_3, \ldots \]

where \( a, b, c, \) and \( d \) are parameters that are determined empirically.

Other forms include the constant elasticity of substitution production function (CES) which is a generalized form of the Cobb-Douglas function, and the quadratic production function which is a specific type of additive function. The best form of the equation to use and the values of the parameters \((a,b,c,\) and \(d)\) vary from company to company and industry to industry. In a short run production function at least one of the \(X\)'s (inputs) is fixed. In the long run all factor inputs are variable at the discretion of management.

**Production function as a graph**

Any of these equations can be plotted on a graph. A typical (quadratic) production function is shown in the following diagram. All points above the production function are unobtainable with current technology, all points below are technically feasible, and all points on the function show the maximum quantity of output obtainable at the specified levels of inputs. From the origin, through points A, B, and C, the production function is rising, indicating that as additional units of inputs are used, the quantity of outputs also increases. Beyond point C, the employment of additional units of inputs produces no additional outputs, in fact, total output starts to decline. The variable inputs are being used too intensively (or to put it another way, the fixed inputs are under utilized). With too much variable input use relative to the available fixed inputs, the company is experiencing negative returns to variable inputs, and diminishing total returns. In the diagram this is illustrated by the negative marginal physical product curve (MPP) beyond point Z, and the declining production function beyond point C.
From the origin to point A, the firm is experiencing increasing returns to variable inputs. As additional inputs are employed, output increases at an increasing rate. Both marginal physical product (MPP) and average physical product (APP) is rising. The inflection point A, defines the point of diminishing marginal returns, as can be seen from the declining MPP curve beyond point X. From point A to point C, the firm is experiencing positive but decreasing returns to variable inputs. As additional inputs are employed, output increases but at a decreasing rate. Point B is the point of diminishing average returns, as shown by the declining slope of the average physical product curve (APP) beyond point Y. Point B is just tangent to the steepest ray from the origin hence the average physical product is at a maximum. Beyond point B, mathematical necessity requires that the marginal curve must be below the average curve (See production theory basics for an explanation.).
Stages of production

To simplify the interpretation of a production function, it is common to divide its range into 3 stages. In Stage 1 (from the origin to point B) the variable input is being used with increasing efficiency, reaching a maximum at point B (since the average physical product is at its maximum at that point). The average physical product of fixed inputs will also be rising in this stage (not shown in the diagram). Because the efficiency of both fixed and variable inputs is improving throughout stage 1, a firm will always try to operate beyond this stage. In stage 1, fixed inputs are underutilized.

In Stage 2, output increases at a decreasing rate, and the average and marginal physical product is declining. However the average product of fixed inputs (not shown) is still rising. In this stage, the employment of additional variable inputs increase the efficiency of fixed inputs but decrease the efficiency of variable inputs. The optimum input/output combination will be in stage 2. Maximum production efficiency must fall somewhere in this stage. Note that this does not define the profit maximizing point. It takes no account of prices or demand. If demand for a product is low, the profit maximizing output could be in stage 1 even though the point of optimum efficiency is in stage 2.

In Stage 3, too much variable input is being used relative to the available fixed inputs: variable inputs are over utilized. Both the efficiency of variable inputs and the efficiency of fixed inputs decline throughout this stage. At the boundary between stage 2 and stage 3, fixed input is being utilized most efficiently and short-run output is maximum.

Shifting a production function

As noted above, it is possible for the profit maximizing output level to occur in any of the three stages. If profit maximization occurs in either stage 1 or stage 3, the firm will be operating at a technically inefficient point on its production function. In the short run it can try to alter demand by changing the price of the output or adjusting the level of promotional expenditure. In the long run the firm has more options available to it, most notably, adjusting its production processes so they better match the characteristics of demand. This usually involves changing the scale of operations by adjusting the level of fixed inputs. If fixed inputs are lumpy, adjustments to the scale of operations may be more significant than what is required to merely balance production capacity with demand. For example, you may only need to increase production by a million units per year to keep up with demand, but the production equipment upgrades that are available may involve increasing production by 2 million units per year.
If a firm is operating (inefficiently) at a profit maximizing level in stage one, it might, in the long run, choose to reduce its scale of operations (by selling capital equipment). By reducing the amount of fixed capital inputs, the production function will shift down and to the left. The beginning of stage 2 shifts from B1 to B2. The (unchanged) profit maximizing output level will now be in stage 2 and the firm will be operating more efficiently.

If a firm is operating (inefficiently) at a profit maximizing level in stage three, it might, in the long run, choose to increase its scale of operations (by investing in new capital equipment). By increasing the amount of fixed capital inputs, the production function will shift up and to the right.

**Homogeneous and homothetic production functions**

There are two special classes of production functions that are frequently mentioned in textbooks but are seldom seen in reality. The production function \( Q = f(X_1,X_2) \) is said to be **homogeneous of degree \( n \)**, if given any positive constant \( k \), \( f(kX_1,kX_2) = k^nf(X_1,X_2) \). When \( n > 1 \), the function exhibits increasing returns, and decreasing returns when \( n < 1 \). When it is homogeneous of degree 1, it exhibits constant returns. Homothetic functions are functions whose marginal technical rate of substitution (slope of the isoquant) is homogeneous of degree zero. Due to this, along rays coming from the origin, the slope of the isoquants will be the same.
Aggregate production functions

In macroeconomics, production functions for whole nations are sometimes constructed. In theory they are the summation of all the production functions of individual producers, however this is an impractical way of constructing them. There are also methodological problems associated with aggregate production functions.

1.2 BASIC CONCEPTS OF PRODUCTION FUNCTION

Some basic concepts of production function are discussed as following

1.2.1 Production

Production refers to the output of goods and services produced by businesses within a market. This production creates the supply that allows our needs and wants to be satisfied. To simplify the idea of the production function, economists create a number of time periods for analysis.

1. Short run production

The short run is a period of time when there is at least one fixed factor input. This is usually the capital input such as plant and machinery and the stock of buildings and technology. In the short run, the output of a business expands when more variable factors of production (e.g. labour, raw materials and components) are employed.

1. Long run production

In the long run, all of the factors of production can change giving a business the opportunity to increase the scale of its operations. For example a business may grow by adding extra labour and capital to the production process and introducing new technology into their operations.

The length of time between the short and the long run will vary from industry to industry. For example, how long would it take a newly created business delivering sandwiches around a local town to move from the short to the long run? Let us assume that the business starts off with leased premises to make the sandwiches; two leased vehicles for deliveries and five full-time and part-time staff. In the short run, they can increase production by using more raw materials and by bringing in extra staff as required. But if demand grows, it wont take the business long to perhaps lease another larger building, buy in some more capital equipment and also lease some extra delivery vans – by the time it has done this, it has already moved into the long run.

The point is that for some businesses the long run can be a matter of weeks! Whereas for industries that requires very expensive capital equipment which may take several months or perhaps years to become available, then the long run can be a sizeable period
of time.

1.2.2 The meaning of productivity

When economists and government ministers talk about productivity they are referring to how productive labour is. But productivity is also about other inputs. So, for example, a company could increase productivity by investing in new machinery which embodies the latest technological progress, and which reduces the number of workers required to produce the same amount of output. The government’s objective is to improve labour and capital productivity in the British economy in order to improve the supply-side potential of the country.

Productivity of the variable factor labour and the law of diminishing returns

In the example of productivity given below, the labour input is assumed to be the only variable factor by a firm. Other factor inputs such as capital are assumed to be fixed in supply. The “returns” to adding more labour to the production process are measured in two ways:

**Marginal product (MP) =** Change in total output from adding one extra unit of labour

**Average product (AP) =** Total Output divided by the total units of labour employed

In the example below, a business hires extra units of labour to produce a higher quantity of wheat. The table below tracks the output that results from each level of employment.

<table>
<thead>
<tr>
<th>Units of Labour Employed</th>
<th>Total Physical Product of wheat (tonnes of wheat)</th>
<th>Marginal Product (tonnes of wheat)</th>
<th>Average Product (tonnes of wheat)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>24</td>
<td>14</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>36</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>40</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>42</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>42</td>
<td>0</td>
<td>6</td>
</tr>
</tbody>
</table>

| **Table 1** |

**Diminishing returns** is said to occur when the **marginal product of labour starts to fall**. In the example above, extra labour is added to a fixed supply of land when a farming business is harvesting wheat. The marginal product of extra workers is
maximized when the 4th worker is employed. Thereafter the output from new workers is falling although total output continues to rise until the seventh worker is employed.

Notice that once marginal product falls below average product we have reached the point where average product is maximized – i.e. we have reached the point of productive efficiency.

1.2.3 Explaining the law of diminishing returns

The law of diminishing returns occurs because factors of production such as labour and capital inputs are not perfect substitutes for each other. This means that resources used in producing one type of product are not necessarily as efficient (or productive) when switched to the production of another good or service. For example, workers employed in producing glass for use in the construction industry may not be as efficient if they have to be re-employed in producing cement or kitchen units. Likewise many items of capital equipment are specific to one type of production. They would be much less efficient in generating output if they were to be switched to other uses. We say that factors of production such as labour and capital can be “occupationally immobile”; they can be switched from one use to another, but with a consequent loss of productivity.

There is normally an inverse relationship between the productivity of the factors of production and the unit costs of production for a business. When productivity is low, the unit costs of supplying a good or service will be higher. It follows that if a business can achieve higher levels of efficiency among its workforce, there may well be a benefit from lower costs and higher profits.

1.2.4 Costs of production

Costs are defined as those expenses faced by a business when producing a good or service for a market. Every business faces costs and these must be recouped from selling goods and services at different prices if a business is to make a profit from its activities. In the short run a firm will have fixed and variable costs of production. Total cost is made up of fixed costs and variable costs

Fixed Costs

These costs relate do not vary directly with the level of output. Examples of fixed costs include:

1. Rent paid on buildings and business rates charged by local authorities.
2. The depreciation in the value of capital equipment due to age.
3. Insurance charges.
4. The costs of staff salaries e.g. for people employed on permanent contracts.
5. Interest charges on borrowed money.
6. The costs of purchasing new capital equipment.
7. Marketing and advertising costs.

**Variable Costs**

Variable costs vary directly with output. I.e. as production rises, a firm will face higher total variable costs because it needs to purchase extra resources to achieve an expansion of supply. Examples of variable costs for a business include the costs of raw materials, labour costs and other consumables and components used directly in the production process.

We can illustrate the concept of fixed cost curves using the table below. The greater the total volume of units produced, the lower will be the fixed cost per unit as the fixed costs are spread over a higher number of units. This is one reason why mass-production can bring down significantly the unit costs for consumers – because the fixed costs are being reduced continuously as output expands.

In our example below, a business is assumed to have fixed costs of £30,000 per month regardless of the level of output produced. The table shows total fixed costs and average fixed costs (calculated by dividing total fixed costs by output).

<table>
<thead>
<tr>
<th>Output (000s)</th>
<th>Total Fixed Costs (£000s)</th>
<th>Average Fixed Cost (AFC)</th>
<th>Fixed Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>30</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>30</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>30</td>
<td>7.5</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>30</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>30</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>30</td>
<td>4.3</td>
<td></td>
</tr>
</tbody>
</table>

*Table 2*

When we add variable costs into the equation we can see the total costs of a business.

The table below gives an example of the short run costs of a firm.
Table 3

Average Total Cost (ATC) is the cost per unit of output produced. ATC = TC divided by output

Marginal cost (MC) is defined as the change in total costs resulting from the production of one extra unit of output. In other words, it is the cost of expanding production by a very small amount.

Long run costs of production

The long run is a period of time in which all factor inputs can be changed. The firm can therefore alter the scale of production. If as a result of such an expansion, the firm experiences a fall in long run average total cost, it is experiencing economies of scale. Conversely, if average total cost rises as the firm expands, diseconomies of scale are happening.

The table below shows a simple example of the long run average cost of a business in the long run when average costs are falling, then economies of scale are being exploited by the business.
<table>
<thead>
<tr>
<th>Long Run Output (units per month)</th>
<th>Total Costs (£s)</th>
<th>Long Run Average Cost (£s per unit)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,000</td>
<td>8,500</td>
<td>8.5</td>
</tr>
<tr>
<td>2,000</td>
<td>15,000</td>
<td>7.5</td>
</tr>
<tr>
<td>5,000</td>
<td>36,000</td>
<td>7.2</td>
</tr>
<tr>
<td>10,000</td>
<td>65,000</td>
<td>6.5</td>
</tr>
<tr>
<td>20,000</td>
<td>120,000</td>
<td>6.0</td>
</tr>
<tr>
<td>50,000</td>
<td>280,000</td>
<td>5.6</td>
</tr>
<tr>
<td>100,000</td>
<td>490,000</td>
<td>4.9</td>
</tr>
<tr>
<td>500,000</td>
<td>2,300,000</td>
<td>4.6</td>
</tr>
</tbody>
</table>

**Table 4**

### 1.3 LAW OF VARIABLE PROPORTIONS

Marginal productivity is *not* obvious in the production function $Y = f(L, K)$. We must first fix one of the factors and let the other factor vary. This is shown in Figure 3, by the "reduced" production function $Y = f(L, K_0)$, where only labor ($L$) varies while capital is held fixed at $K_0$. Reduced production functions where all factors but one are held constant are often referred to as the "total product" curve.

![Figure 3 - Total Product Curve](image)

The total product curve in Figure 3 can be read in conjunction with the average and marginal product curves in Figure 4. The total product curve is originally due to Frank H. Knight (1921: p.100), and much of the subsequent analysis is due to him and John M. Cassels (1936). Although both these sets of curves have long been implicit in much earlier discussions (e.g. Edgeworth, 1911), average and marginal products were confused by early Neoclassicals with surprising frequency. The particular shape of the total
product curve shown in Figure 3 exhibits what has been baptized by John M. Cassels (1936) as *the Law of Variable Proportions* -- effectively what Ragnar Frisch (1965: p.120) quirkily renamed the *ultra-passum law* of production.

The *marginal product* of the factor $L$ is given by the *slope* of the total product curve, thus $\text{MP}_L = \frac{\partial Y}{\partial L} = \frac{df}{dL}$. As we see, at low levels of $L$ up to $L_2$ in Figure 3, we have *rising* marginal productivity of the factor. At levels of $L$ above $L_2$ we have *diminishing* marginal productivity of that factor. Thus, marginal productivity of $L$ reaches its maximum at $L_2$. We can thus trace out a marginal product of $L$ curve, $\text{MP}_L$, in Figure 4. The labels there correspond to those of Figure 3. Thus the $\text{MP}_L$ curve in Figure 4 rises until the inflection point $L_2$, and falls after it. It becomes negative after $L_5$ - which would be equivalent to the "top" of the reduced production function, what Frisch (1965: p.89) calls a "strangulation point". A negative marginal product is akin to a situation when one adds the fiftieth worker to a field whose only accomplishment is to get in everyone else's way - and thus does not increase output at all but actually reduces it.

The slope of the different *rays* through the origin ($O_1$, $O_2$, $O_3$, etc.) in Figure 3 reflect *average products* of the factor $L$, i.e. $\text{AP}_L = \frac{Y}{L}$. The steeper the ray, the higher the average product. Thus, at low levels of output such as $Y_1$, the average product represented by the slope of $O_1$ is rather low, while at some levels of output such as $Y_3$, the average product (here the slope of $O_3$) is much higher. Indeed, as we can see, average product is at its highest at $Y_3$, what is sometimes called the *extensive margin* of production. Notice that at $Y_2$ and $Y_4$ we have the same average product (i.e. the ray $O_2$ passes through both points). The average product curve $\text{AP}_L$ corresponding to Figure 3 is also drawn in Figure 4.

As we can see in Figure 3, the slope of the total product curve is equal to the slope of the ray from the origin at $L_3$, thus average product and marginal product are equal at this
point (as shown in Figure 4). We also know that as the ray from the origin associated with \( L_3 \) is the highest, thus average product curve intersects the marginal product curve, \( MP_L = AP_L \), exactly where the average product curve is at its maximum. Notice that at values below \( L_3 \), \( MP_L > AP_L \), marginal product is greater than average product whereas above \( L_3 \), we have the reverse, \( MP_L < AP_L \). We shall make use of these results later on.

As we can see in Figure 4, it seems that we can have increasing as well as diminishing marginal productivity of labor, as suggested by Walker (1891) and finally acknowledged by Clark (1899: p.164). However, we have already gone a long way in arguing for diminishing marginal productivity that it seems that we must be excluding points where there is rising marginal productivity, i.e. those points to the left of \( L_2 \).

How might such a restriction be justified? In effect, the argument is that in situations of increasing marginal productivity, one can always discard factors and increase output (cf. F.H. Knight, 1921: p.100-104; G. Cassel, 1918: p.279; J.M. Cassels, 1936; A.P. Lerner, 1944: p.153-5). Consider the following example. Assume we have an acre of ripened land to which we are going to apply various quantities of workers. The one lonely worker produces 10 bushels of wheat; two workers will produce 22 bushels of wheat; three workers will produce 36 bushels. Thus, we see:

<table>
<thead>
<tr>
<th>Qty. of Labor</th>
<th>Total Product</th>
<th>Average Product</th>
<th>Marginal Product</th>
</tr>
</thead>
<tbody>
<tr>
<td>One Laborers</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Two Laborers</td>
<td>22</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>Three Laborers</td>
<td>36</td>
<td>12</td>
<td>14</td>
</tr>
</tbody>
</table>

Table 5

Thus, there is increasing average product and increasing marginal product of labor in this example. Why? One can think of it as follows. When we apply one lonely worker to an entire acre of ripened land, his running around the entire acre trying to harvest it will produce a lower average product than if we had three workers, each working a third of the field by himself.

This should already reveal why we would never see a situation of increasing average product. Basically, when we are faced with a situation of a single worker on an acre of land, why should we force him to work on the entire acre and only produce 10 units of output? Average product (and total product) would be higher if instead of forcing that single worker to try to harvest the entire acre, we let him confine himself to a third of that acre, and let the other two-thirds of the plot lie untouched. In this case, the average product of the single worker is as it would have been had there been three harvesters, i.e. 12 units of output. In other words, in situations of rising average and marginal product, total output is increased by discarding two-thirds of the land! Thus, situations of increasing marginal productivity will simply never be seen.
Of course, this logic is not unassailable. While the idea may apply naturally to some cases, it can be questioned in cases where division of labor is crucial as, say, we might have in an automobile factory. Suppose that the average productivity of a worker is highest when there are twenty men working on a factory floor, each worker specializing in fitting a special part of the automobile. We cannot subsequently do the same operation we did before. In other words, we cannot remove nineteen men and let 19/20ths of the car remain unbuilt. The only remaining man, whose productivity was highest when he only fitted wheels on axles, will not yield any output if he is permitted to perform only his specialized task bereft of the other nineteen men. Instead of having cars as output in that case, we would have axles-with-wheels.

Consequently, we see that in order to produce any cars whatsoever, the lone man must be forced to perform all the tasks, not only the fitting of wheels on axles. If this is true, his productivity by himself, where product is measured in number of cars produced rather than axles-with-wheels, will be lower than if he worked together with his nineteen colleagues.

The automobile case shows an example of indivisibility in production, a traditional explanation of increasing marginal productivity (e.g. Edgeworth, 1911; Lerner, 1944). Production is divisible if it "permits any particular method of production, involving certain proportions between factors and products, to be repeated in exactly the same way on larger or on a smaller scale." (Lerner, 1944: p.143). In other words, in a perfectly divisible world, there cannot be changes in method when increasing or decreasing the scale of production. In our automobile example we have indivisibility: when we remove the nineteen men, the remaining man who previously only placed axles on wheels must change his method and do all the tasks in the construction of the automobile. In contrast, our agricultural example was divisible: a laborer working exclusively on his portion of the field will not change his method of harvesting that third of the field when the other laborers on the other the remaining two-thirds of the field are removed.

In sum, increasing marginal productivity, especially in cases where specialization is vital, can be ostensibly encountered in the real world where there are indivisibilities in production. Nevertheless, much of the Neoclassical work on the production function omits this. This is, as noted earlier, is often taken axiomatically, but the question of whether one finds it an acceptable assumption is largely an empirical one.

1.4 RETURNS TO SCALE

The concept of returns to scale can easily be understood under three heads – Classification, Justification and Characterization. These three aspects are as follows:

1.4.1 Classification

Returns to scale are technical properties of the production function, y = f (x₁, x₂, ..., xₙ). If we increase the quantity of all factors employed by the same (proportional) amount, output will increase. The question of interest is whether the resulting output will increase
by the same proportion, more than proportionally, or less than proportionally. In other words, when we double all inputs, does output double, more than double or less than double? These three basic outcomes can be identified respectively as *increasing returns to scale* (doubling inputs more than doubles output), *constant returns to scale* (doubling inputs doubles output) and *decreasing returns to scale* (doubling inputs less than doubles output).

The concept of returns to scale are as old as economics itself, although they remained anecdotal and were not carefully defined until perhaps Alfred Marshall (1890: Bk. IV, Chs. 9-13). Marshall used the concept of returns to scale to capture the idea that firms may alternatively face "economies of scale" (i.e. advantages to size) or "diseconomies of scale" (i.e. disadvantages to size). Marshall's presented an assortment of rationales for why firms may face changing returns to scale and the rationales he offered up were sometimes technical (and thus applicable in general), sometimes due to changing prices (thus only applicable to situations of imperfect competition). As we are focusing on technical aspects of production, we shall postpone the latter for our discussion of the Marshallian firm.

The definition of the concept of returns in to scale in a technological sense was further discussed and clarified by Knut Wicksell (1900, 1901, 1902), P.H. Wicksteed (1910), Piero Sraffa (1926), Austin Robinson (1932) and John Hicks (1932, 1936). Although any particular production function can exhibit increasing, constant or diminishing returns throughout, it used to be a common proposition that a single production function would have different returns to scale at different levels of output (a proposition that can be traced back at least to Knut Wicksell (1901, 1902)). Specifically, it was natural to assume that when a firm is producing at a very small scale, it often faces increasing returns because by increasing its size, it can make more efficient use of resources by division of labor and specialization of skills. However, if a firm is already producing at a very large scale, it will face decreasing returns because it is already quite unwieldy for the entrepreneur to manage properly, thus any increase in size will probably make his job even more complicated. The movement from increasing returns to scale to decreasing returns to scale as output increases is referred to by Frisch (1965: p.120) as the *ultrapassum law* of production.

We can conceive of different returns to scale diagramatically in the simplest case of a one-input/one-output production function $y = f(x)$ as in Figure 5 (note: this is not a total product curve!). As all our inputs (in this case, the only input, $x$) increase, output ($y$) increases, but at different rates. At low levels of output (around $y_1$), the production function $y = f(x)$ is convex, thus it exhibits increasing returns to scale (doubling inputs more than doubles output). At high levels of output (around $y_3$), the production function $y = f(x)$ is concave, thus it exhibits decreasing returns to scale (doubling inputs less than doubles output).

Note: the relationship between convexity and concave production functions and returns to scale can be violated unless the $f(0) = 0$ assumption is imposed. Heuristically, a function exhibits decreasing returns if every ray from the origin cuts the graph of the production
function from below. A production function which is strictly concave but intersects the horizontal axis at a positive level (thus $f(0) < 0$) will not exhibit decreasing returns to scale. Similarly, a non-concave production function which intersects the vertical axis at a positive amount (thus $f(0) > 0$) will exhibit decreasing returns to scale.

![Figure 5 - Returns to Scale for One-Output/One-Input Production Function](image)

### 1.4.2 Justification

The economic justification for these different returns to scale turns out to be far from simple. At the most naive level, we justify increasing returns to scale by appealing to some "division of labor" argument. A single man and a single machine may be able to produce a handful of cars a year, but we will have to have a very amply skilled worker and very flexible machine, able to singlehandedly build every component of a car. Now, as Adam Smith (1776) famously documented, if we add more labor and more machines, each laborer and machine can specialize in a particular sub-task in the car production process, doing so with greater precision in less time so that more cars get built per year than before. The ability to divide tasks, of course, is not available to the single man and single machine.

Specialization reflects, then, the advantage of large scale production over small scale. In Figure 5, assume we increase all factor inputs from $x_1$ to $x_2$, reflecting, say, the movement from a single man-and-machine to fifteen men-with-machines. The total output increases, of course, but so does the productivity of each man-and-machine since fifteen men-and-machines can divide tasks and specialize. So increasing factors fifteen-fold, increases output more than fifteenfold. In effect, we have increasing returns to scale.

We should note that by justifying increasing returns by specialization implies that increasing returns is necessarily associated with a change of method. But this implies there are indivisibilities in production. In other words, the specialized tasks available at large scale are not available at the smaller scale; consequently, as the scale of production increases, these indivisibilities are overcome and thus methods not previously available

Nonetheless, we should note that there are direct examples of pure increasing returns to scale. For instance, consider a cylinder such as an oil pipeline and the mathematical relationship between the steel it contains (= $2\pi rl$ where $r$ is the radius and $l$ the length of the pipe) and the volume of oil it can carry ($= \pi r^2l$). If one adds sufficient steel to the cylinder to double its circumference, one will be more than doubling its volume. Thus, doubling inputs (steel in pipeline) more than doubles output (flow of oil). In this example, increasing returns does not involve changes in technique.

However, these pure examples are rare and the rationale for increasing returns is usually given by specialization. Thus, we can say equivalently that increasing scale captures the idea that there is technical progress with increasing scale. This is how we find it explicitly expressed in the work of Allyn A. Young (1928) and Nicholas Kaldor (1966) and, indeed, modern Neoclassical endogenous growth theory. As such, as it is generally discussed, increasing returns is more than a "pure scale" matter; it is about emerging techniques and changes in technique, of which we shall have more to say later.

Decreasing returns to scale are more difficult to justify. We see that, in Figure 5, moving from $x_2$ to $x_3$, the production function is concave, so that by doubling inputs we less than double output. The naive justification is that the size of production has overstretched itself. The advantages of specialization are being outweighed by the disadvantages of, say, managerial coordination of an enterprise of such great scale.

However often employed (e.g. Marshall, 1890: Ch. 12; Hicks, 1939: p.83; Kaldor, 1934), this "managerial breakdown" explanation is not really legitimate. This is because "returns to scale" requires that we double all inputs, yet we have not increased one of the factors: namely, the managers themselves. In the managerial breakdown argument, the manager implicitly remains as a "fixed factor", thus we are no longer talking of "decreasing returns to scale" in its pure technical sense but rather of diminishing marginal productivity, which is quite a different concept.

In principle, then, decreasing returns to scale is hard to justify technically because every element in production can always be identically replicated (i.e. all inputs increase). To take another common but misleading example, suppose we increase the number of fishing boats in the North Sea. In this case, we would expect each boat to catch relatively less fish. Similarly, taking Pareto's (1896, II: 714) example, doubling the number of train lines from Paris will lead us to expect that each train will carry less passengers. But these examples are not examples of decreasing returns to scale because we have not, appropriately speaking, doubled all inputs: we have kept the North Sea and Paris constant. In other words, we have changed factor proportions: we have more fishing boats per square mile of North Sea and more trains per Parisian passenger.

The accurate exercise for decreasing returns to scale in the first case is to double the number of boats and double the size of the North Sea (and thus double the number of
fish). Similarly, we would need to double the number of trains and double the number of Parisians. In other words, one needs to replicate the North Sea and Paris completely, is entirely possible. The only way one might obtain decreasing returns to scale in these circumstances is if there were externalities of some sort, e.g. the existence of second replica implicates the operation of the first ("there is only one Paris...", etc.). Yet, barring this, two identical North Seas or two identical cities of Paris should not interfere with each other. Thus, decreasing returns simply do not make technical sense since replication does not complicate things.

Another reason for doubting the existence of decreasing returns to scale is more empirical. Specifically, it would not be "rational" for an enterprise to ever produce in such a situation. To see this, suppose there is an entrepreneur who has a given set of laborers and machines willing to work for him. He can either put all these factors into a single factory, or just construct a series of smaller, but identical factories. Obviously, if he is facing decreasing returns to scale, then organizing them into several, decentralized, separate factories is better than throwing them all together into a single, centralized factory. Consequently, one of the justifications sometimes found for arguing for decreasing returns to scale is that production faces indivisibilities so that "dividing" a factory into several factories is simply not possible.

Technically speaking, then, only constant and increasing returns can make sense; decreasing returns are harder to accept. The asymmetric nature of different returns to scale was explicitly admitted by Alfred Marshall in a footnote, "the forces which make for Increasing Return are not of the same order as those that make for Diminishing Return: and there are undoubtedly cases in which it is better to emphasize this difference by describing causes rather than results." (Marshall, 1890: p.266, fn.1). However, the problematic nature of this asymmetry of causes for a competitive economy, as we shall discuss later, were fully uncovered by Piero Sraffa (1925, 1926). It will be noticed that although most textbooks since have continued to refer to the possibility of decreasing returns to scale, they also often add parenthetically that they are assuming a fixed factor, or indivisibilities or some other imperfection that violates somewhat its pure definition. Be that as it may, it is important to note that decreasing returns to scale, in its proper symmetric definition, is rarely held among modern economists. In contrast, increasing returns, as noted, have become irrevocably associated with technical progress.

1.4.3 Characterization

We can characterize the "returns to scale" properties of a production function via the homogeneity properties of the production function. In principle, consider a general function:

\[ y = f(x_1, x_2, ..., x_m) \]

Now, a function of this type is called homogeneous of degree \( r \) if by multiplying all arguments by a constant scalar \( \lambda \), we increase the value of the function by \( \lambda^r \), i.e.
\[ \lambda^r y = f(\lambda x_1, \lambda x_2, \ldots, \lambda x_m) \]

If \( r = 1 \), we call this a **linearly homogenous** function. Now, if we interpret this function to be a production function, then the implications are obvious. If \( r = 1 \), then \( \lambda^r = \lambda \), so increasing inputs by factor \( \lambda \) will increase output by the same factor \( \lambda \). This, of course, is the very definition of constant returns to scale.

If \( r > 1 \), then \( \lambda^r > \lambda \), which implies that when we increase inputs by scalar \( \lambda \), output will increase by more than proportionally. This is the definition of increasing returns to scale. Finally, if \( 0 < r < 1 \), then \( \lambda^r < \lambda \), which implies that increasing inputs by a scalar \( \lambda \) will lead output to increase by less than proportionally. This is the definition of decreasing returns to scale.

The relationship between the elasticity of scale and the output elasticities tell us that, indeed, there is a relationship between returns to scale and marginal productivities. However, it is important not to get confused between the two and assume that, say, diminishing marginal productivity is somehow related to decreasing returns to scale. This is not true. Constant returns or increasing returns to scale are compatible with diminishing marginal productivity.

For instance, examine. Let \( r = 1 \) everywhere, so that we have constant returns, thus \( \lambda^r Y^* = \lambda Y \). Now, if we increase capital only by the scalar \( \lambda \) and leave labor unchanged, we obtain a new configuration \((\lambda K^*, L^*)\) at point f. Notice that we now achieve a level of output \( Y = \mu Y^* = f(\lambda K^*, L^*) \) which is lower than \( Y = \lambda Y^* = f(\lambda K^*, \lambda L^*) \), thus implying that \( \mu < \lambda \). Thus, increasing both capital and labor by the amount \( \lambda \) leads to an increase in output by \( \lambda \). But increasing capital employment only by the amount \( \lambda \) leads to an increase in output of merely \( \mu \).

We can see the (non-)relationship between returns to scale and marginal productivity more clearly if we take a specific functional form for the production function. Consider the famous Cobb-Douglas production function (Wicksell, 1901: p.128; Cobb and Douglas, 1928). This is the following:

\[ Y = f(K, L) = AK^\alpha L^\beta \]

where \( A, \alpha \) and \( \beta \) are positive constants, and \( K \) and \( L \) are capital and labor respectively. Increasing both capital and labor by the scalar \( \lambda \), then we obtain:

\[ A(\lambda K)^\alpha (\lambda L)^\beta = \lambda^{\alpha+\beta} AK^\alpha L^\beta = \lambda^{\alpha+\beta} Y \]

so output increases by the factor \( \lambda^{\alpha+\beta} \). If \( \alpha + \beta = 1 \), then we have constant returns to scale. Decreasing returns to scale implies that \( \alpha + \beta < 1 \) and increasing returns to scale implies that \( \alpha + \beta > 1 \). Now, the marginal products of capital and labor are:

\[ f_K = \frac{\partial Y}{\partial K} = \alpha AK^{\alpha-1} L^\beta = \alpha Y/K \]
\[ f_L = \frac{\partial Y}{\partial L} = \beta AK^\alpha L^{\beta - 1} = \beta Y/L \]

To see diminishing marginal productivity, we must show that marginal products decline as the relevant factors rise. Pursuing this, we see that:

\[ f_{KK} = \frac{\partial f_K}{\partial K} = \alpha (\alpha - 1) AK^{\alpha - 2} L^{\beta} = \alpha (\alpha - 1)Y/K^2 \]

\[ f_{LL} = \frac{\partial f_L}{\partial L} = \beta (\beta - 1) AK^{\alpha} L^{\beta - 2} = \beta (\beta - 1)Y/L^2 \]

Notice that for diminishing marginal productivity, \( f_{KK} < 0 \) and \( f_{LL} < 0 \), and this will be true if \( 0 < \alpha < 1 \) and \( 0 < \beta < 1 \). However, this does not necessarily imply what kind of returns to scale we will obtain. Obviously, decreasing returns to scale fulfills this automatically (\( \alpha + \beta < 1 \)) and so does constant returns (\( \alpha + \beta = 1 \)). However, notice that some cases of increasing returns (\( \alpha + \beta > 1 \)) also satisfy this, e.g. \( \alpha = 0.6 \) and \( \beta = 0.7 \). Thus, while diminishing marginal productivity may be implied by decreasing or constant

### 1.5 CRITICISMS OF PRODUCTION FUNCTIONS

During the 1950s, 60s, and 70s there was a lively debate about the theoretical soundness of production functions. (See the Capital controversy.) Although most of the criticism was directed primarily at aggregate production functions, microeconomic production functions were also put under scrutiny. The debate began in 1953 when Joan Robinson criticized the way the factor input, capital, was measured and how the notion of factor proportions had distracted economists.

According to the argument, it is impossible to conceive of an abstract quantity of capital which is independent of the rates of interest and wages. The problem is that this independence is a precondition of constructing an iso-product curve. Further, the slope of the iso-product curve helps determine relative factor prices, but the curve cannot be constructed (and its slope measured) unless the prices are known beforehand.

Often natural resources are omitted from production functions. When Solow and Stiglitz sought to make the production function more realistic by adding in natural resources, they did it in a manner that economist Georgescu-Roegen criticized as a "conjuring trick" that failed to address the laws of thermodynamics. Neither Solow nor Stiglitz addressed his criticism, despite an invitation to do so in the September 1997 issue of the journal Ecological Economics.

### Activity 1

1. What do you understand by Production Function? Discuss the basic concepts of Production Function.
2. Give a brief note on various types of costs associated with Production Function.
3. Discuss the concept of returns to scale.
4. Justify the relevance of law of variable proportions in theory of production function.

1.6 SUMMARY

When most people think of fundamental tasks of a firm, they think first of production. Economists describe this task with the production function, an abstract way of discussing how the firm gets output from its inputs. It describes, in mathematical terms, the technology available to the firm. In production, returns to scale refers to changes in output subsequent to a proportional change in all inputs (where all inputs increase by a constant factor). The Law of Variable Proportions Is also called the Law of Decreasing marginal returns. It states that "An increase in some inputs relative to other fixed inputs will, in a given state of technology, cause the output to increase, however after a certain point extra output resulting from the same additions of extra inputs will become less and less. After above discussions a brief discussion on criticism of production function is given.

1.7 FURTHER READINGS

UNIT 2

ECONOMIES OF SCALE AND COSTS ANALYSIS

Objectives

After studying this unit you should be able to:

- Understand the basic concepts of Economies of Scale.
- Know the approaches to Marginal Rate of Technical Substitution
- Appreciate the concept and strategies pertaining to Isoquant Analysis
- Be aware about the approach to Returns to Factors
- Have the understanding about the Multi-Product Firms

Structure

2.1 Introduction
2.2 Overview to Economies of Scale
2.3 Marginal Rate of Technical Substitution
2.4 Isoquant Analysis
2.5 Returns to Factors
2.6 The Multi-Product Firm
2.7 Summary
2.8 Further readings

2.1 INTRODUCTION

Economies of scale, in microeconomics, are the cost advantages that a business obtains due to expansion. They are factors that cause a producer’s average cost per unit to fall as scale is increased. An economy of scale is a long run concept and refers to reductions in unit cost as the size of a facility, or scale, increases. Diseconomies of scale are the opposite. Economies of scale may be utilized by any size firm expanding its scale of operation. The common ones are purchasing (bulk buying of materials through long-term contracts), managerial (increasing the specialization of managers), financial (obtaining lower-interest charges when borrowing from banks and having access to a greater range of financial instruments), and marketing (spreading the cost of advertising over a greater range of output in media markets). Each of these factors reduces the long run average costs (LRAC) of production by shifting the short-run average total cost (SRATC) curve down and to the right.
2.2 OVERVIEW TO ECONOMIES OF SCALE

Economies of scale is a practical concept that is important for explaining real world phenomena such as patterns of international trade, the number of firms in a market, and how firms get "too big to fail". Economies of scale are related to and can easily be confused with the theoretical economic notion of returns to scale. Where economies of scale refer to a firm's costs, returns to scale describe the relationship between inputs and outputs in a long-run (all inputs variable) production function. A production function has constant returns to scale if increasing all inputs by some proportion results in output increasing by that same proportion. Returns are decreasing if, say, doubling inputs results in less than double the output, and increasing if more than double the output. If a mathematical function is used to represent the production function, returns to scale are represented by the degree of homogeneity of the function. Production functions with constant returns to scale are first degree homogeneous; increasing returns to scale are represented by degrees of homogeneity greater than one, and decreasing returns to scale by degrees of homogeneity less than one.

The confusion between the practical concept of economies of scale and the theoretical notion of returns to scale arises from the fact that large fixed costs, such as occur from investment in a factory or from research and development, are an important source of real world economies of scale. In conventional microeconomic theory there can be no increasing returns to scale when there are fixed costs, since this implies at least one input that cannot be increased.
A natural monopoly is often defined as a firm which enjoys economies of scale for all reasonable firm sizes; because it is always more efficient for one firm to expand than for new firms to be established, the natural monopoly has no competition. Because it has no competition, it is likely the monopoly has significant market power. Hence, some industries that have been claimed to be characterized by natural monopoly have been regulated or publicly-owned.

In the short run at least one factor of production is fixed. Therefore the SRAC curve will fall and then rise as diminishing returns sets in. In the long run however all factors of production vary and therefore the LRAC curve will fall and then rise according to economies and diseconomies of scale.

There are two typical ways to achieve economies of scale:

1. High fixed cost and constant marginal cost
2. Low or no fixed cost and declining marginal cost

Economies of scale refers to the decreased per unit cost as output increases. More clearly, the initial investment of capital is diffused (spread) over an increasing number of units of output, and therefore, the marginal cost of producing a good or service is less than the average total cost per unit (note that this is only in an industry that is experiencing economies of scale).

An example will clarify. AFC is average fixed cost.

If a company is currently in a situation with economies of scale, for instance, electricity, then as their initial investment of $1000 is spread over 100 customers, their AFC is $1000/100 = $10.

If that same utility now has 200 customers, their AFC becomes $1000/200 = $5. their fixed cost is now spread over 200 units of output. In economies of scale this results in a lower average total cost.

The advantage is that "buying bulk is cheaper on a per-unit basis." Hence, there is economy (in the sense of "efficiency") to be gained on a larger scale.

Economies of scale tend to occur in industries with high capital costs in which those costs can be distributed across a large number of units of production (both in absolute terms and, especially, relative to the size of the market). A common example is a factory. An investment in machinery is made, and one worker, or unit of production, begins to work on the machine and produces a certain number of goods. If another worker is added to the machine he or she is able to produce an additional amount of goods without adding significantly to the factory's cost of operation. The amount of goods produced grows significantly faster than the plant's cost of operation. Hence, the cost of producing an
additional good is less than the good before it, and an economy of scale emerges. Economies of scale are also derived partially from learning by doing.

The exploitation of economies of scale helps explain why companies grow large in some industries. It is also a justification for free trade policies, since some economies of scale may require a larger market than is possible within a particular country — for example, it would not be efficient for Liechtenstein to have its own car maker, if they would only sell to their local market. A lone car maker may be profitable, however, if they export cars to global markets in addition to selling to the local market. Economies of scale also play a role in a "natural monopoly."

Typically, because there are fixed costs of production, economies of scale are initially increasing, and as volume of production increases, eventually diminishing, which produces the standard U-shaped cost curve of economic theory. In some economic theory (e.g., "perfect competition") there is an assumption of constant returns to scale.

**Increasing Returns to Scale**

Economists usually explain "increasing returns to scale" by indivisibility. That is, some methods of production can only work on a large scale -- either because they require large-scale machinery, or because (getting back to Adam Smith, here) they require a great deal of division of labor. Since these large-scale methods cannot be divided up to produce small amounts of output, it is necessary to use less productive methods to produce the smaller amounts. Thus, costs increase less than in proportion to output -- and average costs decline as output increases.

Increasing Returns to Scale is also known as "economies of scale" and as "decreasing costs." All three phrases mean exactly the same.

**Constant Returns to Scale**

We would expect to observe constant returns where the typical firm (or industry) consists of a large number of units doing pretty much the same thing, so that output can be expanded or contracted by increasing or decreasing the number of units. In the days before computer controls, machinery was a good example. Essentially, one machinist used one machine tool to do a series of operations to produce one item of a specific kind -- and to double the output you had to double the number of machinists and machine tools.

Constant Returns to Scale is also known as "constant costs." Both phrases mean exactly the same.

**Decreasing Returns to Scale**

Decreasing returns to scale are associated with problems of management of large, multi-unit firms. Again with think of a firm in which production takes place by a large number of units doing pretty much the same thing -- but the different units need to be coordinated
by a central management. The management faces a trade-off. If they don't spend much on management, the coordination will be poor, leading to waste of resources, and higher cost. If they do spend a lot on management, that will raise costs in itself. The idea is that the bigger the output is, the more units there are, and the worse this trade-off becomes -- so the costs rise either way.

Decreasing Returns to Scale is also known as "diseconomies of scale" and as "increasing costs." All three phrases mean exactly the same.

### 2.3 MARGINAL RATE OF TECHNICAL SUBSTITUTION

In economics, the marginal rate of technical substitution (MRTS) or the Technical Rate of Substitution (TRS) is the amount by which the quantity of one input has to be reduced \((- \Delta x_2)\) when one extra unit of another input is used \((\Delta x_1 = 1)\), so that output remains constant.

where \(MP_1\) and \(MP_2\) are the marginal products of input 1 and input 2, respectively.

Along an isoquant, the MRTS shows the rate at which one input (e.g. capital or labor) may be substituted for another, while maintaining the same level of output. The MRTS can also be seen as the slope of an Isoquant at the point in question. Since the Isoquant is generally downward sloping and marginal products are generally positive, the MRTS is generally negative.

At a given level of output, the slope of the curve relating the two variables gives the rate of change of one variable with respect to another variable. Thus, the rate of change of input \(Y\) with respect to \(X\) --- that is, the rate at which \(Y\) may be substituted for \(X\) in the production process --- is given by the slope of the curve relating \(Y\) to \(X\). This is the slope of the isoquant.

Since the slope is negative and the objective here is to express the substitution rate as a positive quantity, a negative sign is attached to the slope (a convenience factor).

\[
MRTS = \frac{(Y_1 - Y_2)}{(X_1 - X_2)} = \frac{\Delta Y}{\Delta X}
\]

For example, in the Ore Mining problem, given a target output of 29 tons of ore, moving from 3 to 4 workers yields an MRTS of 250 (horsepower).

\[
MRTS = \frac{- (750 - 500)}{(3 - 4)} = -250
\]

Stated differently, for every unit of labor added 250 horsepower may be discharged without changing total output. Or, to discharge one laborer, management must increase the power of the ore mining machine by 250 horsepower.
Note that we can show the MRTS to equal the ratio of the marginal products of X and Y; remember:

$$\Delta Y = \Delta Q / MP_Y$$

and,

$$\Delta X = \Delta Q / MP_X$$

substituting these in above yields, \[ MRTS = MP_X / MP_Y \]

We can also interpret the marginal rate of technical substitution (MRTS) graphically. The MRTS at point b on the q2 isoquant in the figure equals the absolute value of the slope of the straight line that is tangent to the isoquant at that point. The MRTS is approximately equal to the absolute value of the slope of the line from point b to point c, which equals $\Delta K/\Delta L$ (the rise divided by the run). As $\Delta L$ and $\Delta K$ become small, this approximation becomes exact.

Figure: Marginal Rates of Technical Substitution: The MRTS at point b on the q* isoquant equals the absolute value of the slope of the straight line that is tangent to the isoquant at that point. The MRTS is approximately equal to the negative of the slope (-$\Delta K/\Delta L$) of the line from point b to point c. As $\Delta L$ and $\Delta K$ become small, this approximation becomes exact.

Figure 2
2.4 ISOQUANT ANALYSIS

In economics, an isoquant (derived from quantity and the Greek word iso, meaning equal) is a contour line drawn through the set of points at which the same quantity of output is produced while changing the quantities of two or more inputs. While an indifference curve helps to answer the utility-maximizing problem of consumers, the isoquant deals with the cost-minimization problem of producers. Isoquants are typically drawn on capital-labor graphs, showing the tradeoff between capital and labor in the production function, and the decreasing marginal returns of both inputs. Adding one input while holding the other constant eventually leads to decreasing marginal output, and this is reflected in the shape of the isoquant. A family of isoquants can be represented by an isoquant map, a graph combining a number of isoquants, each representing a different quantity of output.

An isoquant shows that the firm in question has the ability to substitute between the two different inputs at will in order to produce the same level of output. An isoquant map can also indicate decreasing or increasing returns to scale based on increasing or decreasing distances between the isoquants on the map as you increase output. If the distance between isoquants increases as output increases, the firm's production function is exhibiting decreasing returns to scale; doubling both inputs will result in placement on an isoquant with less than double the output of the previous isoquant. Conversely, if the distance is decreasing as output increases, the firm is experiencing increasing returns to scale; doubling both inputs results in placement on an isoquant with more than twice the output of the original isoquant.

As with indifference curves, two isoquants can never cross. Also, every possible combination of inputs is on an isoquant. Finally, any combination of inputs above or to the right of an isoquant results in more output than any point on the isoquant. Although the marginal product of an input decreases as you increase the quantity of the input while holding all other inputs constant, the marginal product is never negative since a rational firm would never increase an input to decrease output.

**Shapes of Isoquant Curve:** If the two inputs are perfect substitutes, the resulting isoquant map generated is represented in fig. A; with a given level of production Q3, input X is effortlessly replaced by input Y in the production function. The perfect substitute inputs do not experience decreasing marginal rates of return when they are substituted for each other in the production function.

If the two inputs are perfect complements, the isoquant map takes the form of fig. B; with a level of production Q3, input X and input Y can only be combined efficiently in a certain ratio represented by the kink in the isoquant. The firm will combine the two inputs in the required ratio to maximize output and minimize cost. If the firm is not producing at this ratio, there is no rate of return for increasing the input that is already in excess. Isoquants are typically combined with isocost lines in order to provide a cost-minimization production optimization problem.
An isoquant map where $Q_3 > Q_2 > Q_1$. A typical choice of inputs would be labor for input $X$ and capital for input $Y$. More of input $X$, input $Y$, or both is required to move from isoquant $Q_1$ to $Q_2$, or from $Q_2$ to $Q_3$.

Figure 4

A) Example of an isoquant map with two inputs that are perfect substitutes.

B) Example of an isoquant map with two inputs that are perfect complements.
Notice from the tabular presentation of the production function in Table 1 that different combinations of resources may yield the same level of output. For example, several combinations of labour and capital yield 290 units of output. Some of the information provided in Table 1 can be presented more clearly in graphical form. In Figure 6, the quantity of labour employed is measured along the horizontal axis and the quantity of capital is measured along the vertical axis. The combinations that yield 290 units of output are presented in the figure as points a, b, c and d. These points can be connected to form an isoquant, Q1, which shows the possible combinations of the two resources that produce 290 units of output. Likewise, Q2 shows combinations of inputs that yield 415 units of output, and Q3 shows combinations that yield 475 units of output. (The colours of the isoquants match those of the corresponding entries in the production function table in Table 1.)

An isoquant, such as Q1 in Figure 6, is a curve that shows all the technologically efficient combinations of two resources, such as labour and capital, that produce a certain amount of output. Iso is from the Greek word meaning ‘equal’, and quant is short for ‘quantity’; so isoquant means ‘equal quantity’. Along a particular isoquant, such as Q1, the amount of output produced remains constant, in this case 290 units, but the combination of resources varies. To produce a particular level of output, the firm can employ resource combinations ranging from capital-intensive combinations (much capital and little labour) to labour-intensive combinations (much labour and little capital).

For example, a paving contractor can put in a new driveway with ten workers using shovels and hand-rollers; the same job can also be done with only two workers, a road grader and a paving machine. A Saturday-afternoon charity car wash to raise money to send the school band on a Gold Coast holiday at Nara Sea World is labour-intensive, involving perhaps a dozen workers per car. In contrast, a professional car wash is fully automated, requiring only

<table>
<thead>
<tr>
<th>Units of Capital Employed per Period</th>
<th>Units of Labour Employed per Period</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>40</td>
</tr>
<tr>
<td>2</td>
<td>90</td>
</tr>
<tr>
<td>3</td>
<td>150</td>
</tr>
<tr>
<td>4</td>
<td>200</td>
</tr>
<tr>
<td>5</td>
<td>240</td>
</tr>
<tr>
<td>6</td>
<td>270</td>
</tr>
<tr>
<td>7</td>
<td>290</td>
</tr>
</tbody>
</table>

Table 1
ISOCOST LINES

Isoquants graphically illustrate a firm’s production function for all quantities of output the firm could possibly produce. Given these isoquants, how much should the firm produce? More specifically, what is the firm’s profit-maximising level of output? The answer depends on the cost of resources and on the amount of money the firm plans to spend. Assume a unit of labour costs the firm $15 000 per year, and the cost for each unit of capital is $25 000 per year. The total cost (TC) of production is:

\[ TC = (w \times L) + (r \times K) \]
\[ = $15 000 L + $25 000 K \]

where \( w \) is the annual wage rate, \( L \) is the quantity of labour employed, \( r \) is the annual cost of capital, and \( K \) is the quantity of capital employed. An isocost line identifies all combinations of capital and labour the firm can hire for a given total cost. Again, iso is from the Greek word meaning ‘equal’, so an isocost line is a line representing equal total cost to the firm. In Figure 5, for example, the line \( TC = $150 000 \) identifies all combinations of labour and capital that cost the firm a total of $150 000. The entire $150
000 could pay for 6 units of capital per year; if the entire budget is spent only on labour, 10 workers per year could be hired; or the firm can employ any combination of resources along the isocost line.

Recall that the slope of any line is the vertical change between two points on the line divided by the corresponding horizontal change (the rise over the run). At the point where the isocost line meets the vertical axis, the quantity of capital that can be purchased equals the total cost divided by the annual cost of capital, or TC/r. At the point where the isocost line meets the horizontal axis, the quantity of labour that can be hired equals the firm’s total cost divided by the annual wage, or TC/w. The slope of any isocost line in Figure 5 can be calculated by considering a movement from the vertical intercept to the horizontal intercept. That is, we divide the vertical change (–TC/r) by the horizontal change (TC/w), as follows:

\[
\text{Slope of isocost line} = \frac{\text{–} \frac{TC}{r}}{\frac{TC}{w}} = \frac{-w}{r}
\]

The slope of the isocost line equals minus the price of labour divided by the price of capital, or –w/r, which indicates the relative prices of the inputs. In our example, the absolute value of the slope of the isocost line equals w/r, or:

\[
\text{Slope of isocost line} = \frac{w}{r} = \frac{15 000}{25 000} = 0.6
\]

The wage rate of labour is 0.6 of the annual cost of capital, so hiring one more unit of labour, without incurring any additional cost, implies that the firm must employ 0.6 units less capital. A firm is not confined to a particular isocost line. Thus, a firm’s total cost depends on how much the firm plans to spend. This is why in Figure 8A.2 we include three isocost lines, not just one, each corresponding to a different total budget. In fact, there is a different isocost line for every possible budget. These isocost lines are parallel because each reflects the same relative resource price. Resource prices are assumed to be constant regardless of the amount employed.
The optimal choice of input combinations

We bring the isoquants and the isocost lines together in Figure 6. Suppose the firm has decided to produce 415 units of output and wants to minimise its total cost. The firm could select point $f$, where 6 units of capital are combined with 4 units of labour. This combination, however, would cost $210,000 at prevailing prices. Since the profit-maximising firm wants to produce its chosen output at the minimum cost, it tries to find the isocost line closest to the origin that still touches the isoquant. Only at a point of tangency does a movement in either direction along an isoquant shift the firm to a higher cost level. Hence, it follows that: The point of tangency between the isocost line and the isoquant shows the minimum cost required to produce a given output.

Consider what is going on at the point of tangency. At point $e$ in Figure 6, the isoquant and the isocost line have the same slope. As mentioned already, the absolute value of the slope of an isoquant equals the marginal rate of technical substitution between labour and capital, and the absolute value of the slope of the isocost line equals the ratio of the input prices. So, when a firm produces output in the least costly way, the marginal rate of technical substitution must equal the ratio of the resource prices, or:

\[
\text{Slope} = -\frac{L}{K} = -\frac{15,000}{25,000} = -0.6
\]
MRTSLK = MPL/MPK = \frac{w}{r} = \frac{15 000}{25 000} = 0.6

This equality shows that the firm adjusts resource use so that the rate at which one input can be substituted for another in production — that is, the marginal rate of technical substitution — equals the rate at which one resource can be traded for another in resource markets, that is the resource price ratio \( w/r \). If this equality does not hold, it means that the firm could adjust its input mix to produce the same output for a lower cost. Finally, to demonstrate the consistency between the golden rule for consumer equilibrium and the producer’s equivalent least-cost input combination rule, consider again the firm’s input equilibrium condition above:

\[ MRTS_{LK} = \frac{MP_L}{MP_K} = \frac{w}{r} \]

Now, simply cross-multiply the wage rate \( w \) to the denominator and the MPK to the numerator of their respective opposite sides to yield:

\[ \frac{MP_L}{w} = \frac{MP_K}{r} \]

This is the least-cost input rule for firms operating in competitive resource markets — that is, employs a combination of resources such that the marginal product per dollar spent is equated across all resources used. Here only two resources, capital and labour, are employed. If this least-cost input condition is not met, then, assuming the eventual onset of diminishing returns to variable resources, it is possible to reallocate the amount of resource use between capital and labour until this equilibrium condition does hold.

![Figure 8](image)
Isoquants for a fixed proportions production function

Consider the fixed proportions production function \( F(z_1, z_2) = \min\{z_1, z_2\} \). The 1-isoquant is the set of all pairs \((z_1, z_2)\) for which \( F(z_1, z_2) = 1\), or \( \min\{z_1, z_2\} = 1 \). That is, the 1-isoquant is the set of all pairs of numbers whose smallest member is 1: the set of all pairs \((1, z_2)\) for \( z_2 \geq 1 \) and all pairs \((z_1, 1)\) for \( z_1 \geq 1 \). This set is shown in the following figure, together with the isoquant for the output 2.

Now consider the fixed proportions production function \( F(z_1, z_2) = \min\{z_1/2, z_2\} \), which models a technology in which 2 units of input 1 and 1 units of input 2 are required to produce every unit of output. The 1-isoquant for this technology is the set of all pairs \((z_1, z_2)\) for which \( \min\{z_1/2, z_2\} = 1 \). This isoquant, together with the 2-isoquant is shown in the following figure.
For a general fixed proportions production function $F(z_1, z_2) = \min\{az_1, bz_2\}$, the isoquants take the form shown in the following figure.

**Figure 10**

 Isoquants for a technology in which there are two possible techniques

Consider a technology in which there are two possible techniques. In each technique there is no possibility of substituting one input for another, but various mixes of the two techniques may be used by the firm. For example, perhaps machines can be operated at two possible speeds, fast and slow. If they run fast, then a relatively small amount of labor is used together with a relatively large amount of raw material (since some is
wasted). If they run slowly, then a relatively large amount of labor is used together with a relatively small amount of raw material. The firm can run some of its machines fast, and some slowly. An isoquant for such a technology has the form shown in the following figure. (I am considering only raw material and labor as inputs, ignoring the machine.)

The two corners of the isoquant correspond to the case in which all the machines in the factory run slowly, and the case in which they all run fast. The points in between, on the downward sloping section, correspond to cases in which the firm runs some of its machines fast and some slowly.

Isoquants for a production function in which the inputs are perfect substitutes
If the production function models a technology in which the inputs are perfect substitutes, then it takes the form

\[ F(z_1, z_2) = az_1 + bz_2. \]

In this case the y-isoquant is the set of all pairs \((z_1, z_2)\) for which

\[ az_1 + bz_2 = y, \]

a straight line with slope \(-a/b\). Thus the isoquants are parallel straight lines:
2.5 RETURNS TO FACTORS

The return attributable to a particular common factor is called as return to factor. Recall that in the section Increasing, Decreasing, and Constant Returns to Scale that we can put several factors in varied quantity to get expected level of output as return to these factors. Increasing returns to scale would be when we double all factors, and production more that doubles.

In our example we have two factors K and L, so we'll double K and L and see what happens:

\[ Q = K^a L^b \]

Now lets double all our factors, and call this new production function \( Q' \)

\[ Q' = (2K)^a (2L)^b \]

Rearranging leads to:

\[ Q' = 2^{a+b} K^a L^b \]

Now we can substitute back in our original production function, \( Q \):

\[ Q' = 2^{a+b} Q \]
To get $Q' > 2Q$, we need $2^{a+b} > 2$. This occurs when $a + b > 1$.

As long as $a+b>1$, we will have increasing returns to scale. We also need decreasing returns to scale in each factor. Decreasing returns for each factor occurs when we double only one factor, and the output less than doubles. Try it first for $K$: $Q = K^aL^b$

Now lets double $K$, and call this new production function $Q'$

$$Q' = (2K)^aL^b$$

Rearranging leads to:

$$Q' = 2^aK^aL^b$$

Now we can substitute back in our original production function, $Q$:

$$Q' = 2^aQ$$

To get $2Q > Q'$ (since we want decreasing returns for this factor), we need $2 > 2^a$. This occurs when $1 > a$.

Similarly for $L$: $Q = K^aL^b$

Now lets double $L$, and call this new production function $Q'$

$$Q' = K^a(2L)^b$$

Rearranging leads to:

$$Q' = 2^bK^aL^b$$

Now we can substitute back in our original production function, $Q$:

$$Q' = 2^bQ$$

To get $2Q > Q'$ (since we want decreasing returns for this factor), we need $2 > 2^a$. This occurs when $1 > b$.

So there are your conditions. You need $a+b > 1$, $1 > a$, $1 > b$. By doubling factors, we can easily create conditions where we have increasing returns to scale overall, but decreasing returns to scale in each factor.
2.6 THE MULTI PRODUCT FIRM

Firms producing multiple products are known as multi product firms. Multi product firms’ contribution in making an economy strong is unavoidable. Three sources of advantages to multi-product firm are considered: economies of scope, risk reduction and demand complementarity. Each source if of sufficient magnitude leads to a market equilibrium dominated by multi-product firms. Multi-product firms do not only dominate manufacturing output, they also differ in observable characteristics from single-product firms. A common feature of multi-product firm models is that the presence of headquarter fixed costs implies that the more “able” firms will self-select into becoming multi-product firms. It is evident that firms that eventually expanded their product scope were stronger performers even before the product expansion took place, thus further supporting the selection argument. While existing models yield similar predictions in terms of the ex-ante “quality” of firms that become multi-product firms, their predictions regarding the ex-post performance of such firms differ.

Multi-product firms end up having higher overall productivity than single-product firms. This result is driven by the assumption that each product’s productivity is the sum of a firm level ability component, and a product-specific expertise. Though product-specific expertise is assumed to be uncorrelated across products, the presence of the first component induces positive correlation in the productivities of the products offered by each firm, so that “more able” firms will be more productive in all products. In contrast, in Nocke and Yeaple (2006), it is assumed that marginal costs for each product are increasing in the number of products produced by the firm. This implies that at the equilibrium, multi-product firms will have lower productivity for their inframarginal products, and lower overall productivity than single-product firms, even though such firms were ex-ante better. The point estimate on the multi-product firm dummy indicates that multi-product firms are on average 1 percent more productive, though the estimate is not statistically significant).

Overall, the evidence suggests that multi-product firms are stronger performers, not only in terms of total sales and exports, but also in terms of total factor and labor productivity. It is assumed that firms do not differ in their product-specific expertise; this assumption leads to the prediction that output should be evenly distributed across products within each firm. In contrast, BRS assume that firms possess “core competencies”, so that output should be highly skewed towards products for which firms have particular expertise.

There is a large literature in economics focusing on the size distribution of firms. The natural question arises what share of the differences in the distribution of output across firms can be attributed to the extensive versus the intensive margin. Are bigger firms bigger because they produce more output per product or because they produce more products? One important prediction of the theoretical model developed by BRS (2006b) is that a firm’s extensive (number of products) and intensive margins (output per product) are positively correlated.
Changes in Product Mix over Time

We now examine in detail the nature of product mix changes that led to the observed expansion of the extensive margin. We classify firm activity into one of four mutually exclusive groups: no activity, add products only, drop products only, and both add and drop products. A product is added in period t if it is produced in period t but not in period t-1. A product is dropped in period t, if it was produced in period t-1 but it is not produced in period t.

Changes in product mix provide a non-negligible contribution to changes in output of continuing firms. We decompose the aggregate change in output of continuing firms into changes in output due to changes in product mix (i.e., the extensive margin) and changes in output due to existing products (i.e., the intensive margin) Let \( Y_{jt} \) denote the output of product i produced by firm j at time t, C the set of products that a firm produces in both periods t and t-1 (i.e., the intensive margin), and E the set of products that the firms produces only in t or t-1 (i.e., the extensive margin). Then changes in a firm’s total output between periods t and t-1 can be decomposed as follows:

\[
\Delta Y_{jt} = \sum_{i \in E} \Delta Y_{ijt} + \sum_{i \in C} \Delta Y_{ijt},
\]

We decompose output changes due to the extensive margin further into changes in output due to product additions (A) and product droppings (D):

\[
\sum_{i \in E} \Delta Y_{ijt} = \sum_{i \in A} \Delta Y_{ijt} + \sum_{i \in D} \Delta Y_{ijt}.
\]

Continuing products can be further decomposed into the contributions from growing (G) and shrinking products (S):

\[
\sum_{j \in C} \Delta Y_{ijt} = \sum_{j \in G} \Delta Y_{ijt} + \sum_{j \in S} \Delta Y_{ijt}.
\]

The first two terms capture the growth due to changes in the firms’ extensive product margin and the final two terms capture changes in the intensive margin. The changes in output stemming from the extensive margin are almost entirely driven by output growth due to product additions. Consequently, gross changes in output stemming from the extensive margin are of similar order of magnitude as net changes.
Multi-product oligopoly.

This proposition can be stated in terms of complete products. For instance, the first claim says that total rivals' output of the low-quality good increases while their output of the high-quality good decreases and the second claim says that the opposite happens for firm. Committing to selling only the high-quality good has conflicting effects on r. On the one hand, r is "tougher" in upgrade market 2, leading others to curb production there to the benefit of r. On the other hand, r is "softer" in upgrade market 1, causing others to increase production to the detriment of r.

Champsaur and Rochet (1989) found that price-setting multiproduct duopolists avoid competing head-to-head in order to reduce the intensity of price competition; equilibrium involves each firm committing to an interval of qualities that does not intersect that offered by the other. For similar reasons, firms hold back on the breadth of their product portfolios in the analyses of Anderson and de Palma (1992, 2006). The underlying theoretical reason for why we reach different conclusions is simply that prices tend to be strategic complements, while quantities tend to be strategic substitutes.

One might wonder whether the possibility we identify of firms choosing to compete head-to-head is merely a theoretical curiosity. It is not; there are many examples of firms so competing with vertically differentiated product lines. For example, in the market for plasma televisions, manufacturers such as Mitsubishi, Samsung, and Sony each offer product lines spanning the set of high-resolution possibilities. Similarly, the microprocessors of AMD and Intel compete at many points in quality space, as do the cars of Audi, BMW, Mercedes-Benz, and Jaguar. Note further that in these industries, capacity choices are important, and developing new products takes time. Thus, the spirit of our formulation in this section seems appropriate.

If the softening of price competition is a main objective in designing a line of quality-differentiated products, the outcomes in these industries are less intuitive. However, if we entertain the possibility that competition in these industries is reasonably approximated by quantity setting, then these observations are not surprising.

We conclude that product line commitments introduce interesting strategic effects. Nonetheless, in some cases firms will not gain from strategically omitting products, providing further justification for the simultaneous-move game specified in our base analysis of earlier sections.

Activity 2

1. Discuss in depth the concept of Economies of Scale. What do you understand by increasing returns to scale and how it is different from constant and decreasing return to scale?
2. What do you understand by return to factor? Discuss any 10 factors you think are necessary to be considered in manufacturing process.
3. What are Multi-Product Firms? Discuss its characteristics.
4. Draw an isoquant for the production function:

\[ F(z_1, z_2) = z_1^{1/2} + z_2^{1/2}. \]

5. Find the MRTS for the production function

\[ F(z_1, z_2) = z_1^{1/2} + z_2^{1/2}. \]

2.7 SUMMARY

Economies of scale and diseconomies of scale refer to an economic property of production that affects cost if quantity of all input factors is increased by some amount. In production, returns to scale refer to changes in output subsequent to a proportional change in all inputs (where all inputs increase by a constant factor). Definition Rate at which a producer is technically able to substitute (without affecting the quality of the output) a small amount of one input (such as capital) for a small amount of another input (such as labor). Further the concept of Isoquant and its analysis has been discussed as In economics, an isoquant is a contour line drawn through the set of points at which the same quantity of output is produced while changing the quantities of two or more inputs. Finally a brief discussion on Multi-Product Firm was given to provide readers the great understanding about the concept.

2.8 FURTHER READINGS

- Mas-Colell, Andreu; Whinston, Michael D.; and Jerry R. Green. *Microeconomic Theory*. Oxford University Press
UNIT 3

ELASTICITY OF SUBSTITUTION AND RELATED ASPECTS OF PRODUCTION FUNCTION

Objectives

After studying this unit you should be able to:

- Define the elasticity of substitution and its measurement.
- Understand the approach of Cobb Douglas production function.
- Analyze the Euler’s theorem and its generalization.
- Have the knowledge of CES and VES production functions.
- Know the concept of Technical Progress of production function

Structure

3.1 Introduction
3.2 Measuring the substitutability
3.3 Cobb-Douglas production function
3.4 Constant elasticity of substitution (CES)
3.5 Elasticity of substitution in multi input cases
3.6 Euler’s theorem
3.7 Variable elasticity of substitution (VES)
3.8 Technical progress and production function
3.9 Summary
3.10 further readings

3.1 INTRODUCTION

Elasticity of substitution is the elasticity of the ratio of two inputs to a production (or utility) function with respect to the ratio of their marginal products (or utilities).

The elasticity of substitution was designed as "a measure of the ease with which the varying factor can be substituted for others" (Hicks, 1932: p.117). [on the relationship between the Hicks and Robinson definitions, see R.F. Kahn (1933) and F. Machlup (1935).]

As Abba Lerner (1933) was quick to point out, the elasticity of substitution $\sigma$ is effectively a measure of the curvature of an isoquant. Heuristically, this can be understood by referring to Figure 5.1. Suppose we move from point e to point e’ on the isoquant. At point e, the MRTS is $f_K/f_L$, as represented by the slope of the line tangent to point e, while the labor-capital ratio is $L/K$, as represented by the slope of the chord connecting e to the origin. When we move to e’, the MRTS increases to $f_{K’}/f_{L’}$ while
the labor-capital ratio increases to \( L'/K' \). The elasticity of substitution, thus, compares the movement in the chord from \( L/K \) to \( L'/K' \).

### 3.2 MEASURING SUBSTITUTABILITY

Let us now turn to the issue of measuring the degree of substitutability between any pair of factors. One of the most famous ones is the elasticity of substitution, introduced independently by John Hicks (1932) and Joan Robinson (1933). Formally, the elasticity of substitution measures the percentage change in factor proportions due to a change in marginal rate of technical substitution. In other words, for our canonical production function, \( Y = f (K, L) \), the elasticity of substitution between capital and labor is given by:

\[
\sigma = \frac{d \ln (L/K)}{d \ln (f_K/f_L)} = \frac{[d(L/K)/d(f_K/f_L)] \cdot [(f_K/f_L)/(L/K)]}{[d(f_K/f_L)/(f_K/f_L)]}
\]

The elasticity of substitution was designed as "a measure of the ease with which the varying factor can be substituted for others" (Hicks, 1932: p.117). [on the relationship between the Hicks and Robinson definitions, see R.F. Kahn (1933) and F. Machlup (1935).]

As Abba Lerner (1933) was quick to point out, the elasticity of substitution \( \sigma \) is effectively a measure of the curvature of an isoquant. Heuristically, this can be understood by referring to Figure 5.1. Suppose we move from point e to point e' on the isoquant. At point e, the MRTS is \( f_K/f_L \), as represented by the slope of the line tangent to point e, while the labor-capital ratio is \( L/K \), as represented by the slope of the chord connecting e to the origin. When we move to e', the MRTS increases to \( f_{K'}/f_{L'} \) while the labor-capital ratio increases to \( L'/K' \). The elasticity of substitution, thus, compares the movement in the chord from \( L/K \) to \( L'/K' \) (denoted heuristically by \( \Delta^R \) in Figure 5.1) to the movement in the MRTS from \( f_K/f_L \) to \( f_{K'}/f_{L'} \) (represented by \( \Delta^M \)). The elasticity of substitution is thus, intuitively speaking, merely \( \sigma = \Delta^R/\Delta^M \).
It is immediately deducible that, intuitively, the *more* curved or convex the isoquant is, the *less* the resulting change in the factor proportions will be ($\Delta^R$ is lower for the same $\Delta^M$), thus the elasticity of substitution $\sigma$ is *lower* for very curved isoquants. In the extreme case of Leontief (no-substitution) technology, where the L-shaped isoquants are as "curved" as can be (as shown in our earlier Figure 4.1), a change in MRTS will *not* lead to *any* change in the factor proportions, i.e. $\Delta^R = 0$ for any $\Delta^M$. Thus, $\sigma = 0$ for Leontief isoquants.

The other extreme case of *perfect substitution* or *linear* production technology is shown in Figure 5.2. This represents the case when machines are perfectly substitutable for laborers. In other words, adding a laborer and taking out a machine will not lead to any change in the marginal products of either of them as one is perfectly substitutable for another. A production function which exhibits this can be written as a linear function:

$$Y = f(K, L) = \alpha K + \beta L$$

where $\alpha$, $\beta$ are constants. Notice that $dY/dK = \alpha$ and $dY/dL = \beta$, thus the marginal products of capital and labor are constant and MRTS $= \alpha / \beta$, which is also constant. Thus, as shown in Figure 5.2, the isoquants are straight lines, indicating a constant marginal rate of technical substitution.
Notice that as the MRTS does not change at all along the isoquant, then $\Delta M = 0$. Consequently, the elasticity of substitution of perfect substitute production functions is infinite, i.e. $\sigma = \infty$.

In sum, then, we see that in general, for any production technology, as $\sigma \to \infty$, we approach perfect substitutability between factors, while as $\sigma \to 0$, we approach no substitution between factors. Intuitively, it is clear why. If $\sigma$ is very high, then a small percentage change in the MRTS will engender a very large percentage change in the labor-capital ratio. In order for the input mix to react so violently, they must be very good substitutes. Conversely, if $\sigma$ is very low, a large percentage shift in MRTS barely budges the factor input mix. Thus, if factor proportions are held on to so tightly, they must be needed in relatively fixed proportions.

As such, we can see that the assumption of diminishing marginal productivity, which the early economists struggled with, gains a more interesting and straightforward meaning when viewed in terms of the elasticity of substitution. As we see, diminishing marginal productivity necessarily implies that $\sigma < \infty$. Thus, as Joan Robinson points out, what the assumption of diminishing marginal productivity "really states is that there is a limit to the extent to which one factor of production can be substituted for another, or, in other words, the elasticity of substitution between factors is not infinite" (J. Robinson, 1933: p.330).

The elasticity of substitution can be expressed in various forms. Let $Y = f(K, L)$ be our production function. Now, we know:

$$\sigma = [d(L/K)/d(f_k/f_L) \cdot (f_k/f_L)/(L/K)]$$

Now, totally differentiating the expression $f_k/f_L$ with respect to $K$ and $L$, we obtain:
\[d\left(\frac{f_K}{f_L}\right) = \left[\partial \left(\frac{f_K}{f_L}\right)/\partial K\right] \cdot dK + \left[\partial \left(\frac{f_K}{f_L}\right)/\partial L\right] \cdot dL\]

and, by the definition of the isoquant, \(f_K/f_L = -dL/dK\), or \(dK = -(f_L/f_K)dL\), so:

\[d\left(\frac{f_K}{f_L}\right) = \left[\partial \left(\frac{f_K}{f_L}\right)/\partial K\right] \cdot (-f_L/f_K)dL + \left[\partial \left(\frac{f_K}{f_L}\right)/\partial L\right]dL\]

or simply:

\[d\left(\frac{f_K}{f_L}\right) = \left[\partial \left(\frac{f_K}{f_L}\right)/\partial K\right] \cdot \left(-\frac{f_L}{f_K}\right)dL + \left[\partial \left(\frac{f_K}{f_L}\right)/\partial L\right]dL\]

Now, totally differentiating the expression \(L/K\), we obtain:

\[d\left(\frac{L}{K}\right) = \frac{KdL - LdK}{K^2}\]

or, again as \(dK = -(f_L/f_K)dL\) by the isoquant, this becomes:

\[d\left(\frac{L}{K}\right) = \left[K + L\left(\frac{f_L}{f_K}\right)\right]dL/K\]

Thus dividing this through by \(d\left(\frac{f_K}{f_L}\right)\):

\[d\left(\frac{L}{K}\right)/d\left(\frac{f_K}{f_L}\right) = \left[f_K + f_L\left(\frac{f_L}{f_K}\right)\right]/\left[K^2\left(f_K\left[\partial \left(\frac{f_K}{f_L}\right)/\partial L\right] - f_L\left[\partial \left(\frac{f_K}{f_L}\right)/\partial K\right]\right)\right]\]

Now, dividing through by \(L/K\) and multiplying by \(f_K/f_L\), we obtain the expression for the elasticity of substitution

\[\sigma = \left[d\left(\frac{L}{K}\right)/d\left(\frac{f_K}{f_L}\right)\right] \cdot \left[(f_K/f_L)/(L/K)\right] = \left\{f_K\left[f_K + f_L\left(\frac{f_L}{f_K}\right)\right]/\left[f_LKL(f_K\left[\partial \left(\frac{f_K}{f_L}\right)/\partial L\right] - f_L\left[\partial \left(\frac{f_K}{f_L}\right)/\partial K\right]\right)\right]\right\}\]

All that remains is to evaluate the terms \(\partial \left(\frac{f_K}{f_L}\right)/\partial L\) and \(\partial \left(\frac{f_K}{f_L}\right)/\partial K\). Now,

\[\partial \left(\frac{f_K}{f_L}\right)/\partial K = [f_{K,K}f_L - f_{K,K}f_K]/f_L^2\]

\[\partial \left(\frac{f_K}{f_L}\right)/\partial L = [f_{K,L}f_L - f_{L,L}f_K]/f_L^2\]

Thus, combining, we see that:

\[f_{K}L\left[\partial \left(\frac{f_K}{f_L}\right)/\partial L\right] - f_L\left[\partial \left(\frac{f_K}{f_L}\right)/\partial K\right] = f_{K}\left[f_{K,1}f_{1,L} - f_{K,1}f_{K,1}\right]/f_L^2 - f_{L}\left[f_{K,K}f_L - f_{L,K}f_K\right]/f_L^2\]

\[= (2\left[f_{K,L}f_{1,K} - f_{L,L}f_{K}^2 - f_{K,K}f_L\right]/f_L^2)\]
as, by Young's Theorem, $f_{KL} = f_{LK}$. Thus, we see that plugging back into our expression, we now have:

$$\sigma = \left( f_{LK}f_{K} + f_{LL} \right) / \left( KL \left( 2f_{KL}f_{LL}f_{L}f_{K} - f_{LL}f_{K}^2 - f_{KL}f_{L}^2 \right) \right)$$

which is our alternative expression for $\sigma$. This expression is notable for the fact that the term within the brackets in the denominator is merely the determinant of the bordered Hessian formed by the production function. Recall that for our particular case this is:

$$|B| = f_{L}f_{L}f_{KL}f_{K} - f_{LL}f_{K}^2 - f_{KL}f_{L}^2$$

or:

$$|B| = 2f_{KL}f_{LL}f_{K} - f_{LL}f_{K}^2 - f_{KL}f_{L}^2$$

Also note that the term $f_{L/K}$ is actually the cofactor of the $LK$th term in the Hessian matrix, i.e. $f_{L/K} = |B_{LK}|$. Thus, the elasticity of substitution can be written as:

$$\sigma = \left( (f_{K}K + f_{L}L) / KL \right) \cdot (|B_{LK}| / |B|)$$

Now, recall that quasi-concavity implies that $|B| > 0$, thus automatically we obtain the result that $\sigma > 0$ for quasi-concave production functions with two factors. This, of course, is as is should be expected. Namely, recall that quasi-concavity of the production function implies convexity of the isoquants and that, in turn, implies a diminishing MRTS. Now, a diminishing MRTS, as is obvious from the earlier diagrammatic exposition, implies that $K/L$ and $f_{K}/f_{L}$ move in opposite directions as we go along an isoquant, or, equivalently, that $L/K$ and $f_{K}/f_{L}$ move in the same direction. But this last is precisely what $\sigma$ measures, thus its positivity.

Finally, notice that as, by Young's Theorem, $f_{LK} = f_{KL}$, we have the immediate implication that $|B_{LK}| = |B_{KL}|$ and thus that:

$$\sigma = d \ln \left( L/K \right) / d \ln \left( f_{K}/f_{L} \right) = d \ln \left( K/L \right) / d \ln \left( f_{L}/f_{K} \right)$$

so that the elasticity of substitution is symmetric.

### 3.3 COBB-DOUGLAS PRODUCTION FUNCTIONS

In economics, the **Cobb-Douglas** functional form of production functions is widely used to represent the relationship of an output to inputs. It was proposed by Knut Wicksell.
(1851-1926), and tested against statistical evidence by Charles Cobb and Paul Douglas in 1900-1928. For production, the function is

\[ Y = AL^\alpha K^\beta, \]

where:

- \( Y \) = total production (the monetary value of all goods produced in a year)
- \( L \) = labor input
- \( K \) = capital input
- \( A \) = total factor productivity
- \( \alpha \) and \( \beta \) are the output elasticities of labor and capital, respectively. These values are constants determined by available technology.

Output elasticity measures the responsiveness of output to a change in levels of either labor or capital used in production, *ceteris paribus*. For example if \( \alpha = 0.15 \), a 1% increase in labor would lead to approximately a 0.15% increase in output.

Further, if:

\[ \alpha + \beta = 1, \]

the production function has constant returns to scale. That is, if \( L \) and \( K \) are each increased by 20%, \( Y \) increases by 20%. If
\[ \alpha + \beta < 1, \]

returns to scale are decreasing, and if

\[ \alpha + \beta > 1 \]

returns to scale are increasing. Assuming perfect competition, \( \alpha \) and \( \beta \) can be shown to be labor and capital's share of output.

Cobb and Douglas were influenced by statistical evidence that appeared to show that labor and capital shares of total output were constant over time in developed countries; they explained this by statistical fitting least-squares regression of their production function. There is now doubt over whether constancy over time exists.

If we have \( \sigma = 1 \), then a 10\% change in MRTS will yield a 10\% change in the input mix. This unit-elasticity curve will give our isoquants their traditional, very nice, gently convex shape. A famous case is the well-known Cobb-Douglas production function introduced by Charles W. Cobb and Paul H. Douglas (1928), although anticipated by Knut Wicksell (1901: p.128, 1923) and, some have argued, J.H. von Thünen (1863). [for a review of theoretical and empirical literature on the Cobb-Douglas production function, see Douglas (1934, 1967), Nerlove (1965) and Samuelson (1979)]

The Cobb-Douglas production function normally has the form akin to the following for our canonical case:

\[ Y = AK^\alpha L^\beta \]

where \( A, \alpha \) and \( \beta \) are constants. Let us derive the elasticity of substitution from this. As we know from before, in Cobb-Douglas production functions, \( f_K = \alpha Y/K \) and \( f_L = \beta Y/L \), thus:

\[ f_K/f_L = (\alpha/\beta) \cdot (L/K) \]

Consequently, it follows that:

\[ \sigma = (\beta/\alpha) \cdot [(\alpha /\beta) \cdot (L/K)]/(L/K) = 1 \]

as we announced.

We can now turn to an interesting exercise: namely, that if we have a constant returns to scale production function and the elasticity of substitution is 1, then the form of the production function is necessarily Cobb-Douglas. To see this, recall that when we have constant returns to scale and \( \sigma = 1 \), then we can write it as:

\[ \sigma = d \ln y/d \ln f_L = 1 \]
Integrating:

\[ \ln y = \ln f_L + a \]

where \( a \) is a constant of integration. Consequently, taking the antilog:

\[ y = f_L e^a \]

as \( f_L = y - \phi_k k \) by constant returns, then:

\[ y = (y - \phi_k k)b \]

where \( b = e^a \). Then:

\[ (b-1)y = b\phi_k k \]

Now, as \( \phi_k = dy/dk \), then this can be rewritten as \( (b-1)y = bk(dy/dk) \), or:

\[ (1/y) \cdot dy = ((b-1)/bk) \cdot dk \]

integrating:

\[ \int 1/y \ dy = \int (b-1)/bk \ dk \]

which yields:

\[ \ln y = [(b-1)/b] \cdot \ln k + c \]

\[ = \ln [k^{(b-1)/b}] + c \]

where \( c \) is a constant of integration. Taking the anti-log:

\[ y = e^c k^{(b-1)/b} \]

Letting \( e^c = A \) and \( (b-1)/b = \alpha \), then this becomes:

\[ y = Ak^\alpha \]

Consequently, as \( k^\alpha = (K/L)^\alpha = K^\alpha L^{-\alpha} \), then multiplying the expression through by \( L \), we obtain:

\[ Y = AK^\alpha L^{1-\alpha} \]

which is the Cobb-Douglas form. Thus, Cobb-Douglas is the only form which a constant returns to scale production function with \( \sigma = 1 \) can take.
Slopes

For the utility function, the slope of this curve in the X,U plane is just the marginal utility of X, holding Y constant. For the production function, the slope is the marginal product of one of the two factors, holding the other constant. Using calculus, the slope is simply the partial derivative of the Cobb-Douglas function with respect to X holding Y constant, or vice-versa.

\[
\frac{\partial U}{\partial X \mid Y = \bar{Y}} = \text{marginal utility of } X
\]

One of the reasons the Cobb-Douglas is so popular is that its derivatives are so simple. The function itself is not, but taking the partial of U with respect to X gives:

\[
\frac{\partial U}{\partial X} = \beta X^{\beta - 1} Y^{1-\beta} = \frac{\beta U}{X}
\]

With respect to Y

\[
\frac{\partial U}{\partial Y} = (1 - \beta) X^{\beta} Y^{1-\beta} - 1 = \frac{(1 - \beta) U}{Y}
\]

The slope of the horizontal projection

The horizontal projection of the utility function is an indifference curve. For a production function, it is an isoquant. The slope of either of these two curves in just the rise over the run as it would be in any 2-D space. For the square root version of the Cobb-Douglas, \( \beta = 0.5 \), the slope is very easy to calculate. Consider an indifference curve. First pick a point on the curve, call it \( X_1, Y_1 \). On the same indifference curve a second point might be \( X_2, Y_2 \) and think of it has lower and to the right of the first point. The run is just \( X_2 - X_1 \) while the rise is \( Y_2 - Y_1 \). But since \( Y_2 \) is less than \( Y_1 \) the rise is negative. Since both points are on the indifference curve, the utility must be the same; we have:

\[
U(X_1, Y_1) = U(X_2, Y_2)
\]

Subtracting, we can write:

\[
U_1^2 - U_2^2 = 0
\]

Where:

\[
U_i = \sqrt{X_i, Y_i}.
\]
To get the slope of the indifference curve, we can just add and subtract $X_1Y_2$ to:

$$X_1Y_1 - X_2Y_2 + X_1Y_2 - X_1Y_2 = 0$$

This is a key substitution, one that only works for the square root version of the Cobb-Douglas. Reorganizing this last expression

$$X_1(Y_1 - Y_2) + (X_1 - X_2)Y_2 = 0$$

The elements a slope, the rise $(Y_1 - Y_2)$ over run $(X_1 - X_2)$ are starting to take shape. Divide both sides by $(X_1 - X_2)$

$$X_1 \frac{(Y_1 - Y_2)}{(X_1 - X_2)} + Y_2 = 0$$

and call the slope, $\sigma$

$$\sigma X_1 + Y_2 = 0$$

and solve for $\sigma = -\frac{Y_2}{X_1}$

This is the slope of the straight line in the graph above. To get the instantaneous slope at the point $X_1Y_1$, just move the two points closer and closer together, that is move $X_2Y_2$ down toward $X_1Y_1$. In the limit, that is as the distance between the points goes to zero, we have:
The slope is known as the marginal rate of substitution. In the case of the isoquant, the argument is identical; instead of X and Y we have K and L. But this gives:

\[ \sigma = -\frac{K}{L} \]

so long as L is on the horizontal axis, taking the place of X. For the isoquant, the slope is known as the marginal rate of technical substitution. Another way to get slopes of horizontal projections is to use multivariate calculus. It gets the same result, but requires that you understand a total differential of a function of two variables. In the case of utility, the total differential of U is:

\[ dU = \frac{\partial U}{\partial X} dX + \frac{\partial U}{\partial Y} dY \]

where \( \frac{\partial U}{\partial X} \) is a partial derivative, that is the derivative of U with respect to X holding Y constant, said “the partial of U with respect to Y.” 1 As we saw above, along an indifference curve or an isoquant, the change in U, \( dU = 0 \). We then have:

\[ 0 = \frac{\partial U}{\partial X} dX + \frac{\partial U}{\partial Y} dY \]

from which the slope, \( dY/dX \) of the indifference curve can be calculated.

\[ \frac{dY}{dX} = -\frac{\partial U/\partial X}{\partial U/\partial Y} \]

Using the Cobb-Douglas for Utility or Profit Maximization Now whether we are talking about maximizing utility or minimizing cost, setting the slope of the 2-D projection is set equal to the slope of the constraint is gives one equation for the solution to the maximization problem. For example, in the maximization of utility problem the slope of the budget constraint is just the (negative of the ) opportunity cost of X in terms of the good Y, in other words, \( p_X/p_Y \), the price of X divided by the price of Y. Maximizing utility is requires that these two slopes are the same:

\[ \frac{Y}{X} = -\frac{p_X}{p_Y} \]

This is the tangency condition which must be solved simultaneously with the budget constraint in order to find a maximum:

\[ B = p_X X + p_Y Y \]
where \( B \) is the budget. Substituting the tangency condition into the budget constraint for \( Y \), we have:

\[
B = p_X X + p_Y \left( \frac{p_X}{p_Y} X \right)
\]

Simplifying:

\[
B = 2p_X X \\
X = \frac{B}{2p_X}
\]

and finally, substituting \( X \) into the budget constraint

\[
B = p_X \left( \frac{B}{2p_X} \right) + p_Y Y
\]

**Example 1**

Let \( p_X = 1 \) and \( p_Y = 2 \) and
\( B = 10 \). Solve the consumer’s maximization problem.

**Solution:**

\[
X = \frac{B}{2p_X} = \frac{10}{2(1)} = 5; \text{ and } Y = \frac{B}{2p_Y} = \frac{10}{2(2)} = 2.5
\]

Check to see that the budget is exhausted. Total utility is \( U = \sqrt{12.5} \).

### 3.4 CONSTANT ELASTICITY OF SUBSTITUTION (CES) PRODUCTION FUNCTIONS

Now, recall that the bordered Hessian, \( |B| \), is evaluated at a particular point on the production function. Different points on the production function might yield different \( |B| \). Consequently, as \( |B| \) enters directly into \( \sigma \), it is not surprising that \( \sigma \) could be different at different places on the production function. Thus, in general, \( \sigma \) is not constant.

A special class of production functions, known as Constant Elasticity of Substitution (CES) production functions, were introduced by Arrow, Chenery, Minhas and Solow (1961) (thus it is also known as the ACMS function). It was generalized to the n-factor case by Hirofumi Uzawa (1963) and Daniel McFadden (1963). A CES function, as its name indicates, possesses a constant \( \sigma \) throughout. The CES production function takes the following famous form in the two-input case:

\[
Y = \tau \left[ \alpha K^{-\rho} + (1-\alpha) L^{-\rho} \right]^{1/\rho}
\]

where \( \tau \) denotes the degree of homogeneity of the function; \( \tau > 0 \) is the efficiency parameter which represents the "size" of the production function; \( \alpha \) is the distribution parameter which will help us explain relative factor shares (so \( 0 \leq \alpha \leq 1 \)); and \( \rho \) is the substitution parameter, which will help us derive the elasticity of substitution. Notice that marginal products are:
Thus, immediately we see that MRTS is:

\[
\frac{f_K}{f_L} = \left(\frac{\alpha}{1-\alpha}\right)(L/K)^{\rho+1}
\]

Thus, in order for there to be decreasing MRTS (i.e. convex isoquants), we must assume that the substitution parameter takes on the value \( \rho \geq -1 \). It can be shown that for the constant returns to scale case (\( r = 1 \)), the elasticity of substitution of a CES production function will be \( \sigma = 1/(1+\rho) \), thus we can see immediately that it does not depend on where on the production function we are as \( \rho \) is given exogenously.

Notice also that if we have a Cobb-Douglas production function with constant returns to scale, then \( r = 1 \) and \( \rho = 0 \) so that \( \sigma = 1 \). It is not difficult to show that in this case, the CES production function takes the familiar Cobb-Douglas constant returns to scale form (apply l'Hôpital's rule to obtain this). Other substitution parameter values are also rather straightforward: \( \rho \to \infty \) implies \( \sigma \to 0 \), i.e. Leontief (no-substitution); \( \rho \to -1 \) implies \( \sigma \to \infty \) (i.e. perfect substitutes).

### 3.5 Elasticities of Substitution in Multi-Input Cases

It should be noted that the positivity of \( \sigma \) relies to a good extent on the fact that we are, so far, assuming that \( L \) and \( K \) are substitutes. Specifically, as noted, \( \sigma \) measures the degree of substitutability between two goods and thus the only allowance for complementarity we make is the Leontief case, when \( \sigma = 0 \). However, we are, so far, restricting ourselves to a two-input world, where the degree of complementarity is necessarily restricted. In a more general case, when there are many inputs available, the degree of complementarity may be such that the elasticity of substitution is negative, i.e. \( \sigma < 0 \).

Extending the concept of the elasticity of substitution from a two-input production function into one with three or more inputs invites complications. When measuring the elasticity of substitution between two factors when there are other factors in the production function, one must take care of controlling for possible cross effects. There are different schools of thought on the appropriate measure for the elasticity of substitution between inputs \( i \) and \( j \) in the context of a wider, multiple-input production function \( y = f(x_1, x_2, \ldots, x_m) \).

Three famous measures will be briefly mentioned. The simplest and most obvious measure is the direct elasticity of substitution between two factors \( x_i \) and \( x_j \) and is denoted:

\[
\sigma_{ij}^D = \left(\frac{f_i x_i + f_j x_j}{x_i x_j}\right) \cdot \left(\frac{|B_{ij}|}{|B|}\right)
\]
Specifically, \( x_i \) and \( x_j \) are the quantities of the inputs, \( f_i \) and \( f_j \) are their marginal products, \( |B| \) is the determinant of the bordered Hessian and \( |B_{ij}| \) is the cofactor of \( f_{ij} \) (in our earlier case, this was \( |B_{KL}| = f_{KL} \)). Thus, the direct elasticity is identical to our earlier two-input case, thus, effectively, it is assuming that the other factors in the production function are fixed and thus can be ignored.

Roy G.D. Allen (1938: p.503-5) proposed a different measure, the *Allen elasticity of substitution* (also known as the *partial elasticity of substitution*) and is defined as:

\[
\sigma_{ij}^A = \frac{(\sum_i f_i x_i / x_i) \cdot (|B_{ij}|/|B|)}{|B|} 
\]

where, notice, the numerator holds a larger sum. Notice that if the total number of factors is two, this reduces to the direct elasticity of substitution, i.e. \( \sigma_{ij}^D = \sigma_{ij}^A \). This is perhaps the most popular measure of the elasticity of substitution in general applications, although, intuitively, it seems somewhat amorphous.

We can obtain an interesting alternative expression for the Allen elasticity of substitution. As we shall see later, it turns out that from cost-minimization decision of the firm, we will obtain:

\[
\sigma_{ij}^A = \varepsilon_{ij}/s_j
\]

where \( \varepsilon_{ij} = \partial \ln x_i / \partial \ln w_j \), i.e. the elasticity of the demand for the \( i \)th factor \( x_i \) with respect to the price of the \( j \)th factor \( w_j \). The term \( s_j = w_j x_j / \sum_i w_i x_i \), where the numerator \( w_j x_j \) is the expenditure by the producer on the \( j \)th factor and the denominator \( \sum_i w_i x_i \) is total expenditures. Thus, \( s_j \) is the the \( j \)th factor's share of total expenditures by the producer. This will be useful later in determining the properties of the derived demand for factors.

An alternative measure of elasticity of substitution in the multi-factor case was proposed by Michio Morishima (1967) known as the *Morishima elasticity of substitution* and defined as:

\[
\sigma_{ij}^M = (f_j / f_i) \cdot (|B_{ij}|/|B|) - (f_j / f_i) \cdot (|B_{ij}|/|B|)
\]

which has the seemingly unusual property of being asymmetric, i.e. \( \sigma_{ij}^M \neq \sigma_{ji}^M \). This, as Blackorby and Russell (1981, 1989) argue, *should* be natural for a multi-factor case. It is an algebraic matter to note that we re-express the Morishima measure in terms of the Allen measure as follows:

\[
\sigma_{ij}^M = (f_j x_j / f_i x_i) (\sigma_{ij}^A - \sigma_{ji}^A)
\]

where \( \sigma_{ij}^A \) and \( \sigma_{ji}^A \) are Allen elasticities of substitution. One of the implications we should observe is that the Morishima measure also classifies factors somewhat differently from Allen's measure. More specifically, for any two inputs, \( x_i \) and \( x_j \), it may be that \( \sigma_{ij}^M \)
> 0 but that $\sigma_{ij}^{A} < 0$, so that by the Morishima measure, the inputs are substitutes, but by the Allen measure, the inputs are complements. In general, factors that are substitutes by the Allen measure, will be substitutes by the Morishima measure; but factors that are complements by the Allen measure may still be substitutes by the Morishima measure. Thus, the Morishima measure has a bias towards treating inputs as substitutes (or, alternatively, the Allen measure has a bias towards treating them as complements). This apparently paradoxical result in the Allen and Morishima measures is actually not too disturbing: it reflects the fluidity of the concept of elasticity of substitution in a multiple factor world. For a comparison between them (and defense of the Morishima elasticity), see Blackorby and Russell (1981, 1989).

### 3.6 EULER'S THEOREM

In number theory, Euler's theorem (also known as the Fermat-Euler theorem or Euler's totient theorem) states that if $n$ is a positive integer and $a$ is a positive integer coprime to $n$, then

$$a^{\varphi(n)} \equiv 1 \pmod{n}$$

where $\varphi(n)$ is Euler's totient function and "... $\equiv ... \pmod{n}$" denotes ... congruence ... modulo $n$.

The theorem is a generalization of Fermat's little theorem, and is further generalized by Carmichael's theorem.

The theorem may be used to easily reduce large powers modulo $n$. For example, consider finding the last decimal digit of $7^{222}$, i.e. $7^{222} \pmod{10}$. Note that 7 and 10 are coprime, and $\varphi(10) = 4$. So Euler's theorem yields $7^4 \equiv 1 \pmod{10}$, and we get $7^{222} \equiv 7^{4 \times 55 + 2} \equiv (7^4)^{55} \times 7^2 \equiv 1^{55} \times 7^2 \equiv 49 \equiv 9 \pmod{10}$.

In general, when reducing a power of $a$ modulo $n$ (where $a$ and $n$ are coprime), one needs to work modulo $\varphi(n)$ in the exponent of $a$:

if $x \equiv y \pmod{\varphi(n)}$, then $a^x \equiv a^y \pmod{n}$.

Euler's theorem also forms the basis of the RSA encryption system: encryption and decryption in this system together amount to exponentiating the original text by $\varphi(n)$, so Euler's theorem shows that the decrypted result is the same as the original.

### Proofs

1. Leonhard Euler published a proof in 1736. Using modern terminology, one may prove the theorem as follows: the numbers $a$ which are relatively prime to $n$ form a group under multiplication mod $n$, the group $G$ of (multiplicative) units of the ring $\mathbb{Z}/n\mathbb{Z}$. This group
has \( \varphi(n) \) elements. The element \( a := a \pmod{n} \) is a member of the group \( G \), and the order \( o(a) \) of \( a \) (the least \( k > 0 \) such that \( a^k = 1 \)) must have a multiple equal to the size of \( G \). (The order of \( a \) is the size of the subgroup of \( G \) generated by \( a \), and Lagrange's theorem states that the size of any subgroup of \( G \) divides the size of \( G \).)

Thus for some integer \( M > 0 \), \( M \cdot o(a) = \varphi(n) \). Therefore \( a^{\varphi(n)} = a^{o(a) \cdot M} = (a^{o(a)})^M = 1^M = 1 \). This means that \( a^{\varphi(n)} \equiv 1 \pmod{n} \).

2. Another direct proof: if \( a \) is coprime to \( n \), then multiplication by \( a \) permutes the residue classes mod \( n \) that are coprime to \( n \); in other words (writing \( R \) for the set consisting of the \( \varphi(n) \) different such classes) the sets \( \{ x : x \in R \} \) and \( \{ ax : x \in R \} \) are equal; therefore, their products are equal. Hence, \( P \equiv a^{\varphi(n)}P \pmod{n} \) where \( P \) is the first of those products. Since \( P \) is coprime to \( n \), it follows that \( a^{\varphi(n)} \equiv 1 \pmod{n} \).

**Generalization of Euler’s Theorem**

A function \( F(L,K) \) is homogeneous of degree \( n \) if for any values of the parameter \( \lambda \)

\[
F(\lambda L, \lambda K) = \lambda^n F(L,K)
\]

The analysis is given only for a two-variable function because the extension to more variables is an easy and uninteresting generalization.

Euler's Theorem: For a function \( F(L,K) \) which is homogeneous of degree \( n \)

\[
(\partial F/\partial L)L + (\partial F/\partial K)K = nF(L,K).
\]

Proof: Differentiate the condition

\[
F(\lambda L, \lambda K) = \lambda^n F(L,K)
\]

with respect to \( \lambda \) to obtain

\[
(\partial F/\partial \lambda L)L + (\partial F/\partial \lambda K)K = n\lambda^{n-1}F(L,K)
\]

and let \( \lambda \) equal unity. The result is

\[
(\partial F/\partial L)L + (\partial F/\partial K)K = nF(L,K).
\]

When \( F(L,K) \) is a production function then Euler's Theorem says that if factors of production are paid according to their marginal productivities the total factor payment is equal to the degree of homogeneity of the production function times output. The case of \( n=1 \) is an important special case. For that case if factors of production are paid according to their marginal productivities then output will exactly cover the factor payments.
A corollary to Euler’s Theorem for production functions is that the sum of the elasticities of output with respect to factor inputs is equal to the degree of homogeneity of the production function; i.e.,

\[ L(\partial F/\partial L)/F + K(\partial F/\partial K)/F = n. \]

This result is obtained simply dividing through the equation for Euler’s Theorem by the level of output.

Generalizations

The equation that was obtained by differentiating the defining condition for homogeneity of degree \( n \) with respect to the parameter \( \lambda \) can be differentiated a second time with respect to \( \lambda \) and the value of \( \lambda \) set equal to unity. The result is:

\[
(\partial^2 F/\partial L^2)L^2 + (\partial^2 F/\partial K \partial L)KL + (\partial^2 F/\partial L \partial K)LK + (\partial^2 F/\partial L^2)L^2 = n(n-1)F(L,K)
\]

Since the cross derivatives are equal the above can be expressed as:

\[
(\partial^2 F/\partial L^2)L^2 + 2(\partial^2 F/\partial K \partial L)KL + (\partial^2 F/\partial L^2)L^2 = n(n-1)F(L,K)
\]

For the special case of \( n=1 \) the above equation reduces to:

\[
(\partial^2 F/\partial L^2)L^2 + 2(\partial^2 F/\partial K \partial L)KL + (\partial^2 F/\partial L^2)L^2 = 0
\]

The process of differentiating with respect to \( \lambda \) and setting \( \lambda \) equal to unity can be continued. The result is:

**A Generalization of Euler’s Theorem**

\[
\sum_{i=0}^{m} C(m \, i)(\partial^m F/\partial K^i \partial L^{m-i})L^iK^{m-i} = n(n-1)(n-2)...(n-m+1)F(L,K)
\]

for \( m \) any positive integer less than or equal to \( n+1 \) and where \( C(m \, i) \) are the binomial coefficients \( m!/(m-i)!i! \). In the above formula a partial derivative of the form \( \partial^m F/\partial K^0 \partial L^m \) is just \( \partial^m F/\partial L^m \).

---

**Example 2**

Let \( F(L,K)=L^2K^3 \). This is a homogeneous function of degree 5. Then

\[ \partial F/\partial L = 2LK^3 \text{ and } \partial F/\partial K = 3L^2K^3 \]

so

\[ (\partial F/\partial L)L + (\partial F/\partial K)K = 2L^2K^3 + 3L^2K^3 = 5L^2K^3 \]
The second derivatives are

\[
(\partial F^2/\partial L^2) = 2K^3 \\
(\partial F^2/\partial K\partial L) = (\partial F^2/\partial L\partial K) = 6LK^2 \\
(\partial F^2/\partial K^2) = 6L^2K 
\]

Thus

\[
(\partial F^2/\partial L^2)L^2 + 2(\partial F^2/\partial K\partial L)LK + (\partial F^2/\partial L^2)K^2 \\
[1(2) + 2(6) + 1(6)]L^2K^3 \\
= 20L^2K^3 = 5(4)L^2K^3
\]

The third order derivatives are, without distinguishing between the equal cross derivatives,

\[
(\partial F^3/\partial L^3) = 0 \\
(\partial F^3/\partial K\partial L^2) = 6K^2 \\
(\partial F^3/\partial K^2\partial L) = 12LK \\
(\partial F^3/\partial K^3) = 6L^2 
\]

Thus

\[
(\partial F^3/\partial L^3)L^3 + 3(\partial F^3/\partial K\partial L^2)L^2K + 3(\partial F^3/\partial K^2\partial L)LK^2 + (\partial F^3/\partial K^3)K^3 \\
[1(0) + 3(6) + 3(12) + 1(6)]L^2K^3 = 60L^2K^3 \\
= 5(4)(3)L^2K^3
\]

### 3.7 VES Production Function

We use standard notation to denote a general production technology as \( Y = F(K, L) \), where \( Y \), \( K \), and \( L \) stand for output, capital and labor, respectively. Following Revankar (1971), we consider the following specification:

\[
Y = AK^{\alpha \nu} [L + b\alpha K]^{(1-\alpha)\nu}. \quad (2.1)
\]

We mostly assume that the production function exhibits constant returns to scale, i.e., \( \nu = 1 \). This production function can be written in intensive form, \( y = f(k) \) where \( y \equiv Y/L \) and \( k \equiv K/L \), as

\[
y = Ak^\alpha [1 + b\alpha k]^{1-\alpha}. \quad (2.2)
\]

It follows that
Hence, this function satisfies standard properties of a production function, namely
\[ f(k) > 0, \quad f'(k) > 0 \text{ and } f''(k) < 0 \quad \forall k > 0, \quad \text{as long as} \]
\[ A > 0, \quad 0 < a \leq 1, \quad b > 1 \quad \text{and} \quad k^{-1} \geq -b. \]

Furthermore, it follows from (2.3) that
\[ \lim_{k \to 0} f(k) = 0, \quad \lim_{k \to \infty} f(k) = \infty \quad \text{if} \quad b > 0 \]
\[ \lim_{k \to -b^{-1}} f(k) = A(-b)^{-a}(1-a)^{1-a} > 0 \quad \text{if} \quad b < 0. \]
For this production function, the elasticity of substitution between capital and labor

\[ \sigma(x) = \frac{f'(x)}{xf(x)} \frac{f(x) - xf'(x)}{f''(x)} > 0 \]

or, using (2.8),

\[ \sigma(k) = 1 + bk > 0. \quad (2.8) \]

Hence, \( \sigma \) is if \( b \) is. Thus, the elasticity of substitution varies with the level of per capita capital, an index of economic development. Furthermore, \( \sigma \) plays an important role in the development process. To see why, note that (2.1) can be written as:

\[ Y = AK^{\alpha}L^{1-\alpha} \left[ 1 + b\alpha \frac{K}{L} \right]^{1-\alpha}, \]

or, using (2.8),

\[ Y = AK^{\alpha}L^{1-\alpha} \left[ 1 - \alpha + a\sigma(k) \right]^{1-\alpha}. \quad (2.9) \]

Hence, the production process can be decomposed into a Cobb-Douglas part, \( AK^{\alpha}L^{1-\alpha} \), and a part that depends on the (variable) elasticity of substitution,

\[ [1 - \alpha + a\sigma(k)]^{1-\alpha}. \]

Once again, if \( b = 0 \) then \( \sigma = 1 \) and

\[ Y = AK^{\alpha}L^{1-\alpha}, \]

which is the Cobb-Douglas production function. In intensive form (2.1) is written as

\[ y = Ak^{\alpha} [1 - \alpha + a\sigma(k)]^{1-\alpha}. \quad (2.10) \]

Some of the properties of the VES are also shared by the CES. Exceptions include the elasticity of substitution which for the CES production function is constant along an isoquant, while for the VES considered here it is constant only along a ray through the origin (see equation 2.8). Also, factor shares behave slightly differently, since for the CES \( \lim k \to 0 \) if \( \sigma > 1 \) and \( \lim k \to 0 \) if \( \sigma < 1 \).
3.8 TECHNICAL PROGRESS AND PRODUCTION FUNCTION

The technical progress function is a concept developed by Nicholas Kaldor to explain the rate of growth of labour productivity as a measure of technical progress:

The function is described by the following statements:

1. The larger the rate of growth of capital/input per worker, the larger the rate of growth of output per worker, of labour productivity. The rate of growth of labour productivity is thus explained by the rate of growth of capital intensity.
2. In equilibrium capital/input per worker and output per worker grow at the same rate, the equilibrium rate of growth.
3. At growth rates below the equilibrium rate of growth, the growth rate of output per worker is larger than the growth rate of capital/input per worker.
4. At growth rates above the equilibrium rate of growth it is the other way round, the rate of growth of output per worker is less than the rate of growth of capital/input per worker.

Adding Technical Progress

Recall that when we write our production function as \( Y = F(K, L) \), we are expressing output as a function of capital, labor and the production function's form itself, \( F(\cdot) \). If output is growing, then this can be due to labor growth (changes in \( L \)), capital growth (changes in \( K \)) and productivity growth/technical progress (changes in \( F(\cdot) \)). We have thus far ignored this last component. It is now time to consider it.

Technical progress swings the production function outwards. In a sense, all we need to do is simply add "time" into the production function so that:

\[ Y = F(K, L, t). \]

or, in intensive form:

\[ y = f(k, t) \]

The impact of technical progress on steady-state growth is depicted in Figure 1, where the production function \( f(\cdot, t) \) swings outwards from \( f(\cdot, 1) \) to \( f(\cdot, 2) \) to \( f(\cdot, 3) \) and so on, taking the steady-state capital ratio with it from \( k_1^* \) to \( k_2^* \) and then \( k_3^* \) respectively.

So, at \( t = 1 \), \( f(\cdot, 1) \) rules, so that beginning at \( k_0 \), the capital-labor ratio will rise, approaching the steady-state ratio \( k_1^* \). When technical progress happens at \( t = 2 \), then the production function swings to \( f(\cdot, 2) \), so the capital-labor ratio will continue increasing, this time towards \( k_2^* \). At \( t = 3 \), the third production function \( f(\cdot, 3) \) comes into force and thus \( k \) rises towards \( k_3^* \), etc. So, if technical progress is happens repeatedly over time, the capital-labor ratio will never actually settle down. It will continue to rise, implying all the while that that the growth rates of level variables (i.e. capital, output, etc.) are higher than the growth of population for a rather long period of time.
Before proceeding, the first thing that must be decided is whether this is a "punctuated" or "smooth" movement. Is technical progress a "sudden" thing that happens only intermittently (i.e. we swing the production function out brusquely and drastically and then let it rest), or is it something that is happening all the time (and so we swing the production function outwards slowly and steadily, without pause). Joseph Schumpeter (1912) certainly favored the exciting "punctuated" form of technical progress, but modern growth theorists have adhered almost exclusively to its boring, "smooth" version. In other words, most economists believe that $f(\cdot, t)$ varies continuously and smoothly with $t$.

The simple method of modeling production by merely adding time to the production function may not be very informative as it reveals very little about the nature and character of technical progress. Now, as discussed elsewhere, there are various types of "technical progress" in a production function. The one we shall consider here is Harrod-neutral or labor-augmenting technical progress. In fact, as Hirofumi Uzawa (1961) demonstrated, Harrod-neutral technical progress is the only type of technical progress consistent with a stable steady-state ratio $k^*$. This is because, as we prove elsewhere, only Harrod-neutral technical progress keeps the capital-output ratio, $v$, constant over time.

Formally, the easiest way to incorporate smooth Harrod-neutral technical progress is to add an "augmenting" factor to labor, explicitly:

$$Y = F(K, A(t) \cdot L)$$
where $A(t)$ is a shift factor which depends on time, where $A > 0$ and $\frac{dA}{dt} > 0$.

To simplify our exposition, we can actually think of $A(t) \cdot L$ as the amount of *effective labor* (i.e. labor units $L$ multiplied by the technical shift factor $A(t)$). So, output grows due not only to increases in capital and labor units ($K$ and $L$), but also by increasing the "effectiveness" of each labor unit ($A$). This is the simplest way of adding Harrod-neutral technical progress into our production function. Notice also what the real rate of return on capital and labor become: as $Y = F(K, A(t) \cdot L)$ then the rate of return on capital remains $r = F_K$, but the real wage is now $w = A(t) \cdot (\frac{\partial F}{\partial (A(t) \cdot L)}) = A \cdot F_{AL}$.

Modifying the Solow-Swan model to account for smooth Harrod-neutral technical progress is a simple matter of converting the system into "per effective labor unit" terms, i.e. whenever $L$ was present in the previous model, replace it now with effective labor, $A(t) \cdot L$ (henceforth shortened to $AL$). So, for instance, the new production function, divided by $AL$ becomes:

$$\frac{Y}{AL} = F\left(\frac{K}{AL}, 1\right)$$

so, in intensive form:

$$\frac{y^e}{k^e} = f\left(\frac{k^e}{AL}\right)$$

where $y^e$ and $k^e$ are the output-effective labor ratio and capital-effective labor ratio respectively. Notice that as $F(K, AL) = AL \cdot f\left(\frac{k^e}{AL}\right)$, then by marginal productivity pricing, the rate of return on capital is:

$$r = F_K = \frac{\partial (AL \cdot f\left(\frac{k^e}{AL}\right))}{\partial K}$$

But as $f\left(\frac{k^e}{AL}\right) = f\left(\frac{K}{AL}\right)$, then $\frac{\partial (AL \cdot f\left(\frac{k^e}{AL}\right))}{\partial K} = AL \cdot f\left(\frac{k^e}{AL}\right) \cdot (\frac{\partial k^e}{\partial K})$, and since $\frac{\partial k^e}{\partial K} = 1/AL$, then $\frac{\partial (AL \cdot f\left(\frac{k^e}{AL}\right))}{\partial K} = f\left(\frac{k^e}{AL}\right)$, i.e.

$$r = f\left(\frac{k^e}{AL}\right)$$

the slope of the intensive production function in per effective units terms is still the marginal product of capital.

What about the real wage? Well, continuing to let the marginal productivity theory rule, then notice that:

$$w = \frac{\partial (F(K, AL))}{\partial L}$$

$$= \frac{\partial (AL \cdot f\left(\frac{k^e}{AL}\right))}{\partial L}$$

$$= A \cdot f\left(\frac{k^e}{AL}\right) + AL \cdot f\left(\frac{k^e}{AL}\right) \cdot (\frac{\partial k^e}{\partial L})$$

as $\frac{dk^e}{dL} = -\frac{AK}{(AL)^2} = -\frac{K}{AL^2} = -\frac{k^e}{L}$ then:
\[ w = A \cdot f (k^e) - AL \cdot f' (k^e) \cdot k^e / L \]

or simply:
\[ w = A[f (k^e) - f' (k^e) \cdot k^e] \]

The macroeconomic equilibrium condition \( I = sY \), becomes:
\[ I/AL = s(Y/AL) \]
or:
\[ i^e = sy^e = sf (k^e) \]

where \( i^e \) is the investment-effective labor ratio.

Now, suppose the physical labor units, \( L \), grow at the population growth rate \( n \) (i.e. \( g_L = n \)) and labor-augmenting technical shift factor \( A \) grows at the rate \( \theta \) (i.e. \( g_A = \theta \) ), then effective labor grows at rate \( \theta + n \), i.e.:
\[ g_{AL} = g_A + g_L = \theta + n \]

Now, for steady-state growth, capital must grow at the same rate as effective labor grows, i.e. for \( k^e \) to be constant, then in steady state \( g_K = \theta + n \), or:
\[ I^i = dK/dt = (\theta + n)K \]

is the required investment level. Dividing through by \( AL \):
\[ i^c = (\theta + n)k^e \]

where \( i^c \) is the required rate of investment per unit of effective labor.

The resulting fundamental differential equation is:
\[ dk^e/dt = i^e - i^c \]
or:
\[ dk^e/dt = sf (k^e) - (n + \theta )k^e \]

which is virtually identical with the one we had before. The resulting diagram (Figure 2) will also be the same as the conventional one. The significant difference is that now the growth of the technical shift parameter, \( \theta \), is included into the required investment line and all ratios are expressed in terms of effective labor units.
Consequently, at steady state, \( \frac{d k^e}{dt} = 0 \), and we can define a steady-state capital-effective labor ratio \( k^e* \) which is constant and stable. All level terms -- output, \( Y \), consumption, \( C \), and capital, \( K \) -- grow at the rate \( n+\theta \).

If the end-result is virtually identical to before, what is the gain in adding Harrod-neutral technical progress? This should be obvious. While all the steady-state ratios -- output per effective capita, \( y^e* \), consumption per effective capita, \( c^e* \), and capital per effective capita, \( k^e* \) -- are constant, this is not informative of the welfare of the economy. It is people -- and not effective people -- that receive the income and consume. In other words, to assess the welfare of the economy, we want to look at output and consumption per physical labor unit.

Now, the physical population \( L \) is only growing at the rate \( n \), but output and consumption are growing at rate \( n + \theta \). Consequently, output per person, \( y = \frac{Y}{L} \), and consumption per person, \( c = \frac{C}{L} \), are not constant; they are growing at the steady rate \( \theta \), the rate of technical progress. Thus, although steady-state growth has effective ratios constant, actual ratios are increasing: actual people are getting richer and richer and consuming more and more even when the economy is experiencing steady-state growth.

Activity 3

1. Discuss the concept of Elasticity of Substitution? How it can be measured?
2. Give a brief note on Cobb-douglas Production function.
3. What do you understand by CES and VES production functions?
4. Discuss the applicability of Euler’s Theorem.
3.9 SUMMARY

Elasticity of substitution can be considered as responsiveness of the buyers of a good or service to the price changes in its substitutes. It is measured as the ratio of proportionate change in the relative demand for two goods to the proportionate change in their relative prices. Elasticity of substitution shows to what degree two goods or services can be substitutes for one another. Further in this chapter the Cobb Douglas Production function along with CES and VES production functions have been discussed in detail. Euler’s theorem has been presented with the formula. Finally the concept of Technical progress and its implication in production function was discussed.

3.10 FURTHER READINGS

M.A. PREVIOUS ECONOMICS

PAPER I

MICRO ECONOMIC ANALYSIS

BLOCK 3

PRICE AND OUTPUT DETERMINATION
PAPER I

MICRO ECONOMIC ANALYSIS

BLOCK 3

PRICE AND OUTPUT DETERMINATION

CONTENTS

<table>
<thead>
<tr>
<th>Unit</th>
<th>Page number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit 1 Introduction to Market Structures</td>
<td>4</td>
</tr>
<tr>
<td>Unit 2 Price and output determination under monopoly</td>
<td>32</td>
</tr>
<tr>
<td>Unit 3 Monopolistic competition, Oligopoly and other market structures</td>
<td>62</td>
</tr>
</tbody>
</table>
BLOCK 3 PRICE AND OUTPUT DETERMINATION

The three units contained in this block deal with price and output determination in various types of market structures.

The first unit examines the concept of market structures and introduces some important market structures namely Perfect competition, Imperfect competition, Monopoly, Oligopoly and Monopolistic competition.

The second unit deals with Price and output determination under monopoly. Monopoly equilibrium is discussed with marginal analysis, short run production analysis, profit maximization, and price discrimination.

The third unit discusses some other market structures along with their characteristics and price and output determination under these structures.
UNIT 1
INTRODUCTION TO MARKET STRUCTURES

Objectives

After completing this unit, you should be able to:

- Understand the meaning of market structures
- Become aware of the concept of perfect and imperfect competition
- Understand the approach to monopoly
- Determine the market structures called oligopoly and duopoly
- Know the basic concepts of monopolistic competition

Structure

1.1 Introduction
1.2 Perfect competition
1.3 Imperfect competition
1.4 Monopoly
1.5 Oligopoly
1.6 Monopolistic competition
1.7 Summary
1.8 Further readings

1.1 INTRODUCTION

Market structure can be described with reference to different characteristics of a market, including its size and value, the number of providers and their market share, consumer and business purchasing behavior, and growth forecasts. The description may also include a demographic and regional breakdown of providers and customers and an analysis of pricing structures, likely technological impacts, and domestic and overseas sales.

Interconnected characteristics of a market, such as the number and relative strength of buyers and sellers and degree of collusion among them, level and forms of competition, extent of product differentiation, and ease of entry into and exit from the market. Four basic types of market structure are (1) Perfect competition: many buyers and sellers, none being able to influence prices. (2) Oligopoly: several large sellers who have some control over the prices. (3) Monopoly: single seller with considerable control over supply and prices. (4) Monospony: single buyer with considerable control over demand and prices. These market structures and some other important ones are discussed in this unit.
1.2 PERFECT COMPETITION

In neoclassical economics and microeconomics, **perfect competition** describes the perfect being a market in which there are many small firms, all producing homogeneous goods. In the short term, such markets are productively inefficient as output will not occur where marginal cost is equal to average cost, but allocatively efficient, as output under perfect competition will always occur where marginal cost is equal to marginal revenue, and therefore where marginal cost equals average revenue. However, in the long term, such markets are both allocatively and productively efficient. In general a perfectly competitive market is characterized by the fact that no single firm has influence on the price of the product it sells. Because the conditions for perfect competition are very strict, there are few perfectly competitive markets.

A perfectly competitive market may have several distinguishing characteristics, including:

- **Many Buyers/Many Sellers** – Many consumers with the willingness and ability to buy the product at a certain price. Many producers with the willingness and ability to supply the product at a certain price.
- **Low-Entry/Exit Barriers** – It is relatively easy to enter or exit as a business in a perfectly competitive market.
- **Perfect Information** – Prices are assumed to be known to all consumers and producers.
- **Transactions are Costless** – Buyers and sellers incur no costs in making an exchange.
- **Firms Aim to Maximize Profits** – Firms aim to sell where marginal costs meet marginal revenue, where they generate the most profit.
- **Homogeneous Products** – The characteristics of any given market good or service do not vary across suppliers.

Some subset of these conditions is presented in most textbooks as defining perfect competition. More advanced textbooks (e.g. Mas-Colell et al. 1995 p. 315) try to reconcile these conditions with the definition of perfect competition as equilibrium price taking; that is whether or not firms treat price as a parameter or a choice variable. It is this distinction which differentiates perfectly competitive markets from imperfectly competitive ones. It should be noted that a general rigorous proof that the above conditions indeed suffice to guarantee price taking is still lacking (Kreps 1990, p. 265).

The importance of perfect competition derives from the fact that price taking by the firm guarantees that when firms maximize profits (by choosing quantity they wish to produce, and the combination of Factors of production to produce it with) the market price will be equal to marginal cost. An implication of this is that a factor's price (wage, rent, etc.) equals the factor's marginal revenue product. This allows for derivation of the supply curve on which the neoclassical approach is based (note that this is also the reason why "a monopoly does not have a supply curve"). The abandonment of price taking creates considerable difficulties to the demonstration of existence of a general equilibrium
Robertson and Sonnenschein 1977) except under other, very specific conditions such as that of monopolistic competition.

**Approaches and conditions**

Historically, in neoclassical economics there have been two strands of looking at what perfect competition is. The first emphasis is on the inability of any one agent to affect prices. This is usually justified by the fact that any one firm or consumer is so small relative to the whole market that their presence or absence leaves the equilibrium price very nearly unaffected. This assumption of negligible impact of each agent on the equilibrium price has been formalized by Aumann (1964) by postulating a continuum of infinitesimal agents. The difference between Aumann’s approach and that found in undergraduate textbooks is that in the first, agents have the power to choose their own prices but do not individually affect the market price, while in the second it is simply assumed that agents treat prices as parameters. Both approaches lead to the same result.

The second view of perfect competition conceives of it in terms of agents taking advantage of – and hence, eliminating – profitable exchange opportunities. The faster this arbitrage takes place, the more competitive a market is. The implication is that the more competitive a market is under this definition, the faster the average market price will adjust so as to equate supply and demand (and also equate price to marginal costs). In this view “perfect” competition means that this adjustment takes place instantaneously. This is usually modeled via the use of the Walrasian auctioneer (see article for more information). The widespread recourse to the auctioneer tale appears to have favored an interpretation of perfect competition as meaning price taking always, i.e. also at nonequilibrium prices; but this is rejected e.g. by Arrow (1959) or Mas-Colell et al. (1995, p. 315).

Thus nowadays the dominant intuitive idea of the conditions justifying price taking and thus rendering a market perfectly competitive is an amalgam of several different notions, not all present, nor given equal weight, in all treatments. Besides product homogeneity and absence of collusion, the notion more generally associated with perfect competition is the negligibility of the size of agents, which makes them believe that they can sell as much of the good as they wish at the equilibrium price but nothing at a higher price (in particular, firms are described as each one of them facing a horizontal demand curve). However, also widely accepted as part of the notion of perfectly competitive market are perfect information about price distribution and very quick adjustments (whose joint operation establish the law of one price), to the point sometimes of identifying perfect competition with an essentially instantaneous reaching of equilibrium between supply and demand. Finally, the idea of free entry with free access to technology is also often listed as a characteristic of perfectly competitive markets, probably owing to a difficulty with abandoning completely the older conception of free competition. In recent decades it has been rediscovered that free entry can be a foundation of absence of market power, alternative to negligibility of agents (Novshek and Sonnenschein 1987).
Free entry also makes it easier to justify the absence of collusion: any collusion by existing firms can be undermined by entry of new firms. The necessarily long-period nature of the analysis (entry requires time!) also allows a reconciliation of the horizontal demand curve facing each firm according to the theory, with the feeling of businessmen that "contrary to economic theory, sales are by no means unlimited at the current market price" (Arrow 1959 p. 49).

There are double the buyers in the market

![Figure 1](image)

**Results**

In the short-run, it is possible for an individual firm to make a profit. This situation is shown in this diagram, as the price or average revenue, denoted by $P$, is above the average cost denoted by $C$. 
However, in the long period, positive profit cannot be sustained. The arrival of new firms or expansion of existing firms (if returns to scale are constant) in the market causes the (horizontal) demand curve of each individual firm to shift downward, bringing down at the same time the price, the average revenue and marginal revenue curve. The final outcome is that, in the long run, the firm will make only normal profit (zero economic profit). Its horizontal demand curve will touch its average total cost curve at its lowest point. (See cost curve.)

In a perfectly competitive market, a firm's demand curve is perfectly elastic.

As mentioned above, the perfect competition model, if interpreted as applying also to short-period or very-short-period behaviour, is approximated only by markets of homogeneous products produced and purchased by very many sellers and buyers, usually organized markets for agricultural products or raw materials. In real-world markets, assumptions such as perfect information cannot be verified and are only approximated in organized double-auction markets where most agents wait and observe the behaviour of prices before deciding to exchange (but in the long-period interpretation perfect information is not necessary, the analysis only aims at determining the average around which market prices gravitate, and for gravitation to operate one does not need perfect information).

In the absence of externalities and public goods, perfectly competitive equilibria are Pareto-efficient, i.e. no improvement in the utility of a consumer is possible without a worsening of the utility of some other consumer. This is called the First Theorem of
Welfare Economics. The basic reason is that no productive factor with a non-zero marginal product is left unutilized, and the units of each factor are so allocated as to yield the same indirect marginal utility in all uses, a basic efficiency condition (if this indirect marginal utility were higher in one use than in other ones, a Pareto improvement could be achieved by transferring a small amount of the factor to the use where it yields a higher marginal utility).

A simple proof assuming differentiable utility functions and production functions is the following. Let $w_j$ be the 'price' (the rental) of a certain factor $j$, let $MP_{j1}$ and $MP_{j2}$ be its marginal product in the production of goods 1 and 2, and let $p_1$ and $p_2$ be these goods' prices. In equilibrium these prices must equal the respective marginal costs $MC_1$ and $MC_2$: remember that marginal cost equals factor 'price' divided by factor marginal productivity (because increasing the production of good $i$ by one very small unit through increase of the employment of factor $j$ requires increasing the factor employment by $1/MP_{ji}$ and thus increasing the cost by $w_j/MP_{ji}$, and through the condition of cost minimization that marginal products must be proportional to factor 'prices' it can be shown that the cost increase is the same if the output increase is obtained by optimally varying all factors). Optimal factor employment by a price-taking firm requires equality of factor rental and factor marginal revenue product, $w_j = p_i MP_{ji}$, so we obtain $p_1 = MC_1 = w_j/MP_{j1}$, $p_2 = MC_2 = w_j/MP_{j2}$.

Now choose any consumer purchasing both goods, and measure his utility in such units that in equilibrium his marginal utility of money (the increase in utility due to the last unit of money spent on each good), $MU_1/p_1 = MU_2/p_2$, is 1. Then $p_1 = MU_1$, $p_2 = MU_2$. The indirect marginal utility of the factor is the increase in the utility of our consumer achieved by an increase in the employment of the factor by one (very small) unit; this increase in utility through allocating the small increase in factor utilization to good 1 is $MP_{j1}MU_1 = MP_{j1}p_1 = w_j$, and through allocating it to good 2 it is $MP_{j2}MU_2 = MP_{j2}p_2 = w_j$ again. With our choice of units the marginal utility of the amount of the factor consumed directly by the optimizing consumer is again $w$, so the amount supplied of the factor too satisfies the condition of optimal allocation.

Monopoly violates this optimal allocation condition, because in a monopolized industry market price is above marginal cost, and this means that factors are underutilized in the monopolized industry, they have a higher indirect marginal utility than in their uses in competitive industries. Of course this theorem is considered irrelevant by economists who do not believe that general equilibrium theory correctly predicts the functioning of market economies; but it is given great importance by neoclassical economists and it is the theoretical reason given by them for combating monopolies and for antitrust legislation.

**Profit**

In contrast to a monopoly or oligopoly, it is impossible for a firm in perfect competition to earn economic profit in the long run, which is to say that a firm cannot make any more money than is necessary to cover its economic costs. In order not to misinterpret this
zero-long-run-profits thesis, it must be remembered that the term 'profit' is also used in other ways. Neoclassical theory defines profit as what is left of revenue after all costs have been subtracted, including normal interest on capital plus the normal excess over it required to cover risk, and normal salary for managerial activity. Classical economists on the contrary defined profit as what is left after subtracting costs except interest and risk coverage; thus, if one leaves aside risk coverage for simplicity, the neoclassical zero-long-run-profit thesis would be re-expressed in classical parlance as profits coinciding with interest in the long period, i.e. the rate of profit tending to coincide with the rate of interest. Profits in the classical meaning do not tend to disappear in the long period but tend to normal profit. With this terminology, if a firm is earning abnormal profit in the short term, this will act as a trigger for other firms to enter the market. They will compete with the first firm, driving the market price down until all firms are earning normal profit only.

It is important to note that perfect competition is a sufficient condition for allocative and productive efficiency, but it is not a necessary condition. Laboratory experiments in which participants have significant price setting power and little or no information about their counterparts consistently produce efficient results given the proper trading institutions (Smith, 1987, p. 245).

**The shutdown point**

When a firm is making a loss, it will have to decide whether to continue production or not. This decision will, in fact, depend on the different total costs levels and whether the firm is operating in the short run or in the long run.

If the firm is in the **short run**, and is making a loss whereby:

- Total costs (TC) is greater than total revenue (TR)
- and whereby total revenue is greater or equal to total variable cost (TVC)

it is advisable for the firm to continue production. If it fails to achieve these conditions, it is advised to close down so that the only costs the firm will have to pay will be the fixed costs.

Even if the firm stops producing, it will have to continue to meet the level of fixed costs. Since whether the firm produces or not, it will have to pay fixed costs, it is better for it to continue production in an attempt to decrease total costs and increase total revenue, thus making profits. This can be done by:

- **Increasing productivity.** The most obvious methods involve automation and computerization which minimize the tasks that must be performed by employees. All else constant, it benefits a business to improve productivity, which over time lowers cost and (hopefully) improves ability to compete and make profit.
- **Adopting new methods of production like Just In Time or lean manufacturing** in an attempt to reduce costs and wastages.
In the **long run**, the condition to continue producing requires the price $P$ to be higher than the ATC, i.e. the line representing market price should be above the minimum point of the ATC curve.

If $P$ is equal to ATC, the firm is indifferent between shutting down and continuing to produce. This case is different from the short run shut down case because in long run there's no longer a fixed cost (everything is variable).

**Examples**

Perhaps the closest thing to a perfectly competitive market would be a large auction of identical goods with all potential buyers and sellers present. By design, a stock exchange resembles this, not as a complete description (for no markets may satisfy all requirements of the model) but as an approximation. The flaw in considering the stock exchange as an example of Perfect Competition is the fact that large institutional investors (e.g. investment banks) may solely influence the market price. This, of course, violates the condition that "no one seller can influence market price".

Ebay auctions can be often seen as perfectly competitive. There are very low barriers to entry (anyone can sell a product, provided they have some knowledge of computers and the Internet), many sellers of common products and many potential buyers.

Moreover, the buyers and sellers on ebay have (nearly) perfect information about the market, since they can quickly search and view all the other equivalent products on sale, and have exact information on price and product across the entire market at once.

In the eBay market competitive advertising does not occur, because the products are homogeneous and this would be redundant. However, generic advertising (advertising which benefits the industry as a whole and does not mention any brand names) may occur.

Free software works along lines that approximate perfect competition. Anyone is free to enter and leave the market at no cost. All code is freely accessible and modifiable, and individuals are free to behave independently. Free software may be bought or sold at whatever price that the market may allow.

Some believe that one of the prime examples of a perfectly competitive market anywhere in the world is street food in developing countries. This is so since relatively few barriers to entry/exit exist for street vendors. Furthermore, there are often numerous buyers and sellers of a given street food, in addition to consumers/sellers possessing perfect information of the product in question. It is often the case that street vendors may serve a homogenous product, in which little to no variations in the product's nature exist.
Criticisms

The use of the assumption of perfect competition as the foundation of price theory for product markets is often criticized as representing all agents as passive, thus removing the active attempts to increase one's welfare or profits by price undercutting, product design, advertising, innovation, activities that - the critics argue - characterize most industries and markets. These criticisms point to the frequent lack of realism of the assumptions of product homogeneity and impossibility to differentiate it, but apart from this the accusation of passivity appears correct only for short-period or very-short-period analyses, in long-period analyses the inability of price to diverge from the natural or long-period price is due to active reactions of entry or exit.

A frequent criticism is that it is often not true that in the short run differences between supply and demand cause changes in price; especially in manufacturing, the more common behaviour is alteration of production without nearly any alteration of price (Lee 1998). Anyway, the critics of the assumption of perfect competition in product markets seldom question the basic neoclassical view of the working of market economies for this reason. The Neo-Austrian school insists strongly on this criticism, and yet the neoclassical view of the working of market economies as fundamentally efficient, reflecting consumer choices and assigning to each agent his/her contribution to social welfare, is esteemed to be fundamentally correct (Kirzner 1981). Some non-neoclassical schools, like Post-Keynesians, reject the neoclassical approach to value and distribution, but not because of their rejection of perfect competition as a reasonable approximation to the working of most product markets; the reasons for rejection of the neoclassical ‘vision’ are different views of the determinants of income distribution and of aggregate demand (Petri 2004).

In particular, the rejection of perfect competition does not generally entail the rejection of free competition as characterizing most product markets; indeed it has been argued (Clifton 1977) that competition is stronger nowadays than in 19th century capitalism, owing to the increasing capacity of big conglomerate firms to enter any industry: therefore the classical idea of a tendency toward a uniform rate of return on investment in all industries owing to free entry is even more valid to-day; and the reason why General Motors, Exxon or Nestle do not enter the computers or pharmaceutical industries is not insurmountable barriers to entry but rather that the rate of return in the latter industries is already sufficiently in line with the average rate of return elsewhere as not to justify entry. On this few economists, it would seem, would disagree, even among the neoclassical ones. Thus when the issue is normal, or long-period, product prices, differences on the validity of the perfect competition assumption do not appear to imply important differences on the existence or not of a tendency of rates of return toward uniformity as long as entry is possible, and what is found fundamentally lacking in the perfect competition model is the absence of marketing expenses and innovation as causes of costs that do enter normal average cost. The issue is different with respect to factor markets. Here the acceptance or denial of perfect competition in labour markets does make a big difference to the view of the working of market economies. One must distinguish neoclassical from non-neoclassical economists. For the former, absence of
perfect competition in labour markets, e.g. due to the existence of trade unions, impedes the smooth working of competition, which if left free to operate would cause a decrease of wages as long as there were unemployment, and would finally ensure the full employment of labour: labour unemployment is due to absence of perfect competition in labour markets. Most non-neoclassical economists deny that a full flexibility of wages would ensure the full employment of labour and find a stickiness of wages an indispensable component of a market economy, without which the economy would lack the regularity and persistence indispensable to its smooth working. This was, for example, Keynes's opinion. Particularly radical is the view of the Sraffian school on this issue: the labour demand curve cannot be determined hence a level of wages ensuring the equality between supply and demand for labour does not exist, and economics should resume the viewpoint of the classical economists, according to whom competition in labour markets does not and cannot mean indefinite price flexibility as long as supply and demand are unequal, it only means a tendency to equality of wages for similar work, but the level of wages is necessarily determined by complex sociopolitical elements; custom, feelings of justice, informal allegiances to classes, as well as overt coalitions such as trade unions, far from being impediments to a smooth working of labour markets that would be able to determine wages even without these elements, are on the contrary indispensable because without them there would be no way to determine wages (Garegnani 1990).

1.3 IMPERFECT COMPETITION

In economic theory, imperfect competition is the competitive situation in any market where the conditions necessary for perfect competition are not satisfied. It is a market structure that does not meet the conditions of perfect competition. [1]

Forms of imperfect competition include:

- **Monopoly**, in which there is only one seller of a good.
- **Oligopoly**, in which there is a small number of sellers.
- **Monopolistic competition**, in which there are many sellers producing highly differentiated goods.
- **Monopsony**, in which there is only one buyer of a good.
- **Oligopsony**, in which there is a small number of buyers.

There may also be imperfect competition in markets due to buyers or sellers lacking information about prices and the goods being traded.

There may also be imperfect competition due to a time lag in a market. An example is the “jobless recovery”. There are many growth opportunities available after a recession, but it takes time for employers to react, leading to high unemployment. High unemployment decreases wages, which makes hiring more attractive, but it takes time for new jobs to be created.
1.4 MONOPOLY

Monopoly is a market structure characterized by a single seller of a unique product with no close substitutes. This is one of four basic market structures. The other three are perfect competition, oligopoly, and monopolistic competition. As the single seller of a unique good with no close substitutes, a monopoly has no competition. The demand for output produced by a monopoly is THE market demand, which gives monopoly extensive market control. The inefficiency that results from market control also makes monopoly a key type of market failure.

Monopoly is a market in which a single firm is the only supplier of the good. Anyone seeking to buy the good must buy from the monopoly seller. This single-seller status gives monopoly extensive market control. It is a price maker. The market demand for the good sold by a monopoly is the demand facing the monopoly. Market control means that monopoly does not equate price with marginal cost and thus does not efficiently allocate resources.

Characteristics

The four key characteristics of monopoly are: (1) a single firm selling all output in a market, (2) a unique product, (3) restrictions on entry into the industry, and (4) specialized information about production techniques unavailable to other potential producers.

Single Supplier: First and foremost, a monopoly is a monopoly because it is the only seller in the market. The word monopoly actually translates as "one seller." As the only seller, a monopoly controls the supply-side of the market completely. If anyone wants to buy the good, they must buy from the monopoly.

Unique Product: A monopoly achieves single-seller status because the good supplied is unique. There are no close substitutes available for the good produced by a monopoly.

Barriers to Entry: A monopoly often acquires and generally maintains single seller status due to restrictions on the entry of other firms into the market. Some of the key barriers to entry are: (1) government license or franchise, (2) resource ownership, (3) patents and copyrights, (4) high start-up cost, and (5) decreasing average total cost. These restrictions might be imposed for efficiency reasons or simply for the benefit of the monopoly.

Specialized Information: A monopoly often possesses information not available to others. This specialized information comes in the form of legally-established patents, copyrights, or trademarks.

Reasons

Monopolies achieve their single-seller status for three interrelated reasons: (1) economies of scale, (2) government decree, and (3) resource ownership. While a monopoly can emerge and persist for any one of these reasons, most monopolies rely on two or all three.
**Economies of Scale**: Many real world monopolies emerge due to economies of scale and decreasing average cost. If average cost decreases over the entire range of demand, then a single seller can provide the good at lower per unit cost and more efficiently than multiple sellers. This often leads to what is termed a natural monopoly. The market might start with more than one seller, but it naturally ends up with a single seller that can best take advantage of decreasing average cost. Many public utilities (such as electricity distribution, natural gas distribution, garbage collection) have this natural monopoly inclination.

**Government Decree**: The monopoly status of a firm can be established by the mandate of government. Government simply gives one and only one firm the legal authority to supply a particular good. Such single seller legal status is usually justified on economic grounds, such as an electric company that naturally tends to monopolize a market. However, it might also result from political forces, such as mandating monopoly status to a firm controlled by a campaign donor or close political associate.

**Resource Ownership**: A monopoly is likely to arise if a firm has complete control over a key input or resource used in production. If the firm controls the input, then it controls the output. Monopolies have arisen over the years due to control over material resources (petroleum and bauxite ore), labor resources (talented entertainers and skilled athletes), or information resources (patents and copyrights).

**Demand and Revenue**

Single-seller status means that monopoly faces a negatively-sloped Demand Curve, demand curve, such as the one displayed in the exhibit to the right. In Monopoly fact, the demand curve facing the monopoly is the market demand curve for the product. The top curve in the exhibit is the demand curve (D) facing the monopoly. The lower curve is the marginal revenue curve (MR). Because a monopoly is a price maker with extensive market, it faces a negatively-sloped demand curve. To sell a larger quantity of output, it must lower the price. For example, the monopoly can sell 1 unit for $10. However, if it wants to sell 2 units, then it must lower the price to $9.50. For this reason, the marginal revenue generated from selling extra output is less than price. While the price of the second unit sold is $9.50, the marginal revenue generated by selling the second unit is only $9. While the $9.50 price means the monopoly gains $9.50 from selling the second unit, it loses $0.50 due to the lower price on the first unit ($10 to $9.50). The net gain in revenue, that is marginal revenue, is thus only $9 (= $9.50 - $0.50).
Laws pertaining to monopoly

The existence of a very high market share does not always mean consumers are paying excessive prices since the threat of new entrants to the market can restrain a high-market-share firm's price increases. Competition law does not make merely having a monopoly illegal, but rather abusing the power a monopoly may confer, for instance through exclusionary practices.

First it is necessary to determine whether a firm is dominant, or whether it behaves "to an appreciable extent independently of its competitors, customers and ultimately of its consumer." As with collusive conduct, market shares are determined with reference to the particular market in which the firm and product in question is sold.

Under EU law, very large market shares raise a presumption that a firm is dominant, which may be rebuttable. If a firm has a dominant position, then there is "a special responsibility not to allow its conduct to impair competition on the common market". The lowest yet market share of a firm considered "dominant" in the EU was 39.7%.

Certain categories of abusive conduct are usually prohibited under the country's legislation, though the lists are seldom closed. The main recognized categories are:

- Predatory pricing
- Tying (commerce) and product bundling
- Limiting supply
- Price discrimination
- Refusal to deal and exclusive dealing

Despite wide agreement that the above constitute abusive practices, there is some debate about whether there needs to be a causal connection between the dominant position of a company and its actual abusive conduct. Furthermore, there has been some consideration of what happens when a firm merely attempts to abuse its dominant position.

Historical monopolies
The term "monopoly" first appears in Aristotle's Politics, wherein Aristotle describes Thales of Miletus' cornering of the market in olive presses as a monopoly.

Common salt (sodium chloride) historically gave rise to natural monopolies. Until recently, a combination of strong sunshine and low humidity or an extension of peat marshes was necessary for winning salt from the sea, the most plentiful source. Changing sea levels periodically caused salt "famines" and communities were forced to depend upon those who controlled the scarce inland mines and salt springs, which were often in hostile areas (the Dead Sea, the Sahara desert) requiring well-organized security for transport, storage, and distribution. The "Gabelle", a notoriously high tax levied upon salt, played a role in the start of the French Revolution, when strict legal controls were in place over who was allowed to sell and distribute salt.

Robin Gollan argues in The Coalminers of New South Wales that anti-competitive practices developed in the Newcastle coal industry as a result of the business cycle. The monopoly was generated by formal meetings of the local management of coal companies agreeing to fix a minimum price for sale at dock. This collusion was known as "The Vend." The Vend collapsed and was reformed repeatedly throughout the late nineteenth century, cracking under recession in the business cycle. "The Vend" was able to maintain its monopoly due to trade union support, and material advantages (primarily coal geography). In the early twentieth century as a result of comparable monopolistic practices in the Australian coastal shipping business, the vend took on a new form as an informal and illegal collusion between the steamship owners and the coal industry, eventually going to the High Court as Adelaide Steamship Co. Ltd v. R. & AG.

Examples of alleged and legal monopolies

- The salt commission, a legal monopoly in China formed in 758.
- British East India Company; created as a legal trading monopoly in 1600.
- Standard Oil; broken up in 1911, two of its surviving "baby companies" are ExxonMobil and Chevron.
- Major League Baseball; survived U.S. anti-trust litigation in 1922, though its special status is still in dispute as of 2009.
- Microsoft; settled anti-trust litigation in the U.S. in 2001; fined by the European Commission in 2004 for 497 million Euros [1], which was upheld for the most part by the Court of First Instance of the European Communities in 2007. The fine was 1.35 Billion USD in 2008 for noncompliance with the 2004 rule.[15][16]
- Joint Commission; has a monopoly over whether or not US hospitals are able to participate in the Medicare and Medicaid programs.

**Short-Run Production**

The analysis of short-run production by a monopoly provides insight into efficiency (or lack thereof). The key assumption is that a monopoly, like any other firm, is motivated by profit maximization. The firm chooses to produce the quantity of output that
generates highest possible level of profit, given price, market demand, cost conditions, production technology, etc.

The short-run production decision for monopoly can be illustrated using the exhibit to the right. The top panel indicates the two sides of the profit decision—revenue and cost. The hump-shaped green line is total revenue (TR). Because price depends on quantity, the total revenue curve is a hump-shaped line. The curved red line is total cost (TC). The difference between total revenue and total cost is profit, which is illustrated by the lower panel as the brown line.

A firm maximizes profit by selecting the quantity of output that generates the greatest gap between the total revenue line and the total cost line in the upper panel, or at the peak of the profit curve in the lower panel. In this example, the profit maximizing output quantity is 6. Any other level of production generates less profit.

**Figure 3 Short-Run Production, Monopoly**

**A Few Problems**

Three problems often associated with a market controlled totally by a single firm are: (1) inefficiency, (2) income inequality, (3) political abuse.
Inefficiency: The most noted monopoly problem is inefficiency. Market control means that a monopoly charges a higher price and produces less output than would be achieved under perfect competition. In addition, and most indicative of inefficiency, the price charged by the monopoly is greater than the marginal cost of production.

Income Inequality: A lesser known problem with monopoly is an inequitable distribution of income. To the extent that monopoly earns economic profit, consumer surplus is transferred from buyers to the monopoly. Buyers end up with less income and the monopoly ends up with more. In addition, because price is greater than marginal cost and a monopoly receives economic profit, factor payments to some or all of the resources used by the monopoly are greater than their contributions to production. A portion of this economic profit is often "paid" to the owners of the labor, capital, or land.

Political Abuse: A third potential problem, one tied directly to the concentration of income by the monopoly resources, is the abuse of political power. The monopoly could use its economic profit to influence the political process, especially policies that might prevent potential competitors from entering the market.

1.5 OLIGOPOLY

Oligopoly is a market structure characterized by a small number of large firms that dominate the market, selling either identical or differentiated products, with significant barriers to entry into the industry. This is one of four basic market structures. The other three are perfect competition, monopoly, and monopolistic competition. Oligopoly dominates the modern economic landscape, accounting for about half of all output produced in the economy. Oligopolistic industries are as diverse as they are widespread, ranging from breakfast cereal to cars, from computers to aircraft, from television broadcasting to pharmaceuticals, from petroleum to detergent.

Oligopoly is a market structure characterized by a small number of relatively large firms that dominate an industry. The market can be dominated by as few as two firms or as many as twenty, and still be considered oligopoly. With fewer than two firms, the industry is monopoly. As the number of firms increase (but with no exact number) oligopoly becomes monopolistic competition.

Because an oligopolistic firm is relatively large compared to the overall market, it has a substantial degree of market control. It does not have the total control over the supply side as exhibited by monopoly, but its capital is significantly greater than that of a monopolistically competitive firm.

Relative size and extent of market control means that interdependence among firms in an industry is a key feature of oligopoly. The actions of one firm depend on and influence the actions of another. Such interdependence creates a number of interesting economic issues. One is the tendency for competing oligopolistic firms to turn into cooperating oligopolistic firms. When they do, inefficiency worsens, and they tend to come under the
scrutiny of government. Alternatively, oligopolistic firms tend to be a prime source of innovations, innovations that promote technological advances and economic growth.

Like much of the imperfection that makes up the real world, there is both good and bad with oligopoly. The challenge in economics is, of course, to promote the good and limit the bad.

**Characteristics**

The three most important characteristics of oligopoly are: (1) an industry dominated by a small number of large firms, (2) firms sell either identical or differentiated products, and (3) the industry has significant barriers to entry.

**Small Number of Large Firms:** An oligopolistic industry is dominated by a small number of large firms, each of which is relatively large compared to the overall size of the market. This generates substantial market control, the extent of market control depending on the number and size of the firms.

**Identical or Differentiated Products:** Some oligopolistic industries produce identical products, while others produce differentiated products. Identical product oligopolies tend to process raw materials or intermediate goods that are used as inputs by other industries. Notable examples are petroleum, steel, and aluminum. Differentiated product oligopolies tend to focus on consumer goods that satisfy the wide variety of consumer wants and needs. A few examples of differentiated oligopolistic industries include automobiles, household detergents, and computers.

**Barriers to Entry:** Firms in a oligopolistic industry attain and retain market control through barriers to entry. The most common barriers to entry include patents, resource ownership, government franchises, start-up cost, brand name recognition, and decreasing average cost. Each of these make it extremely difficult, if not impossible, for potential firms to enter an industry.

**Behavior**

Although oligopolistic industries tend to be diverse, they also tend to exhibit several behavioral tendencies: (1) interdependence, (2) rigid prices, (3) nonprice competition, (4) mergers, and (5) collusion.

- **Interdependence:** Each oligopolistic firm keeps a close eye on the activities of other firms in the industry. Decisions made by one firm invariably affect others and are invariably affected by others. Competition among interdependent oligopoly firms is comparable to a game or an athletic contest. One team's success depends not only on its own actions but on the actions of its competitor. Oligopolistic firms engage in competition among the few.
- **Rigid Prices:** Many oligopolistic industries (not all, but many) tend to keep prices relatively constant, preferring to compete in ways that do not involve changing the price. The prime reason for rigid prices is that competitors are likely to match
price decreases, but not price increases. As such, a firm has little to gain from changing prices.

- Nonprice Competition: Because oligopolistic firms have little to gain through price competition, they generally rely on nonprice methods of competition. Three of the more common methods of nonprice competition are: (a) advertising, (b) product differentiation, and (c) barriers to entry. The goal for most oligopolistic firms is to attract buyers and increase market share, while holding the line on price.

- Mergers: Oligopolistic firms perpetually balance competition against cooperation. One way to pursue cooperation is through merger—legally combining two separate firms into a single firm. Because oligopolistic industries have a small number of firms, the incentive to merge is quite high. Doing so then gives the resulting firm greater market control.

- Collusion: Another common method of cooperation is through collusion—two or more firms that secretly agree to control prices, production, or other aspects of the market. When done right, collusion means that the firms behave as if they are one firm, a monopoly. As such they can set a monopoly price, produce a monopoly quantity, and allocate resources as inefficiently as a monopoly. A formal method of collusion, usually found among international produces is a cartel.

### Kinked-Demand Curve

Short-run production activity of an oligopolistic firm is often illustrated by a kinked-demand curve, such as the one presented in the exhibit to the right. A kinked-demand curve has two distinct segments with different elasticities that join to form a kink. The primary use of the kinked-demand curve is to explain price rigidity in oligopoly. The two segments are: (1) a relatively more elastic segment for price increases and (2) a relatively less elastic segment for price decreases. The relative elasticities of these two segments is directly based on the interdependent decision-making of oligopolistic firms.

The kink of the demand curve exists at the current quantity (Qo) and price (Po).

Because competing firms ARE NOT likely to match the price increases of an oligopolistic firm, the firm is likely to lose customers and market share to the competition. Small price increases result in relatively large decreases in quantity demanded.

However, because competing firms ARE likely to match the price decreases of an oligopolistic firm, the firm is unlikely to gain customers and market share from
the competition. Large price decreases are needed to gain relatively small increases in quantity demanded.

Each segment of the demand curve has its own marginal revenue segment. This actually means that the marginal revenue curve facing an oligopolistic firm, labeled MR in the exhibit, contains three distinct segments.

Top Segment: The flatter top portion of the marginal revenue corresponds to the more elastic demand generated by price increases.

Bottom Segment: The steeper bottom portion of the marginal revenue corresponds to the less elastic demand generated by price decreases.

Middle Segment: The vertical middle segment connecting the top and bottom segments that occurs at the output quantity Qo corresponds with the kink of the curve.

The vertical segment is key to the analysis of short-run production by an oligopolistic firm. It means that the firm can equate marginal revenue with marginal cost, and thus maximize profit, even though marginal cost increases or decreases. If marginal cost increases a bit, the profit-maximizing price and quantity remain at Po and Qo. If marginal cost decreases a bit, the profit-maximizing price and quantity also remain at Po and Qo.

**Game Theory**

A technique often used to analyze interdependent behavior among oligopolistic firms is game theory. Game theory illustrates how the choices between two players affect the outcomes of a "game." This analysis illustrates two firms cooperating through collusion are better off than if they compete.

The exhibit to the right illustrates the alternative facing two oligopolistic firms, Juice-Up and Omni-Cola, as they ponder the prospects of advertising their products.

- In the top left quadrant, if Omni-Cola and Juice-Up BOTH decide to advertise, then each receives $200 million in profit.
- However, in the lower right quadrant, if NEITHER Omni-Cola nor Juice-Up decide to advertise, then each receives $250 million in profit. They receive more because they do not incur any advertising expense.
- Alternatively, as shown in the lower left quadrant, if Omni-Cola advertises but Juice-Up does not, then Omni-Cola receives $350 million in profit and Juice-Up receives only $100 in profit. Omni-Cola receives a big boost in profit because its advertising attracts customers away from Juice-Up.
But, as shown in the top right quadrant, if Juice-Up advertises and OmniCola does not, then Juice-Up receives $350 million in profit and OmniCola receives only $100 in profit. Juice-Up receives a big boost in profit because its advertising attracts customers away from OmniCola.

Game theory indicates that the best choice for OmniCola is to advertise, regardless of the choice made by Juice-Up. And Juice-Up faces EXACTLY the same choice. Regardless of the decision made by OmniCola, Juice-Up is wise to advertise.

The end result is that both firms decide to advertise. In so doing, they end up with less profit ($200 million each), than if they had colluded and jointly decided not to advertise ($250 million each).

**The Bad of Oligopoly**

Like much of life, oligopoly has both bad and good. The bads are that oligopoly: (1) tends to be inefficient in the allocation of resources and (2) promotes the concentration, and thus inequality, of income and wealth.

- **Inefficiency:** First and foremost, oligopoly does NOT efficiently allocate resources. Like any firm with market control, an oligopoly charges a higher price and produces less output than the efficiency benchmark of perfect competition. In fact, oligopoly tends to be the worst efficiency offender in the real world, because perfect competition does not exist, monopolistic competition inefficiency is minor, and monopoly inefficiency has the potential for being so bad that it is inevitably subject to corrective government regulation.

- **Concentration:** Another bad is that oligopoly tends to increase the concentration of wealth and income. This is not necessarily bad, but it can be self-reinforcing and inhibit pursuit of the microeconomic goal of equity. While the concentration of wealth is not bad unto itself, such wealth can then be used (or abused) to exert influence over the economy, the political system, and society, which might not be beneficial for society as a whole.

**The Good of Oligopoly**

With the bad comes a little good. The two most noted goods from oligopoly are (1) by developing product innovations and (2) taking advantage of economies of scale.

- **Innovations:** Of the four market structures, oligopoly is the one most likely to develop the innovations that advance the level of technology, expand production capabilities, promote economic growth, and lead to higher living standards. Oligopoly has both the motive and the opportunity to pursue innovation. Motive comes from interdependent competition and opportunity arises from access to abundant resources.

- **Economies of Scale:** Oligopoly firms are also able to take advantage of economies of scale that reduce production costs and prices. As large firms, they can "mass produce" at low average cost. Many modern goods—including cars, computers, aircraft, and assorted household products—would be significantly more expensive
if produced by a large number of small firms rather than a small number of large firms.

Examples

In industrialized countries barriers to entry have found oligopolies forming in many sectors of the economy:

Unprecedented levels of competition, fueled by increasing globalization, have resulted in oligopolies emerging in many market sectors. Market share in an oligopoly are typically determined on the basis of product development and advertising. There are now only a small number of manufacturers of civil passenger aircraft, though Brazil (Embraer) and Canada (Bombardier) have fielded entries into the smaller-market passenger aircraft market sector. A further instance arises in a heavily regulated market such as wireless communications. In some areas only two or three providers are licensed to operate.

Australia

- Most media outlets are owned either by News Corporation, Time Warner, or Fairfax Media.
- Retailing is dominated by Coles Group and Woolworths.

Canada

- Three companies (Rogers Wireless, Bell Mobility and Telus) constitute over a 94% share of Canada's wireless market.

United Kingdom

- Four companies (Tesco, Sainsbury's, Asda and Morrisons) share between them 74.4% of the grocery market
- Scottish & Newcastle, Molson Coors, and Inbev control two thirds of the beer brewing industry.
- The detergent market is dominated by two players Unilever and Procter & Gamble.

United States

- Anheuser-Busch and MillerCoors control about 80% of the beer industry.
- Many media industries today are essentially oligopolies. Six movie studios receive 90% of American film revenues, and four major music companies receive 80% of recording revenues. There are just six major book publishers, and the television industry was an oligopoly of three networks—ABC, CBS, and NBC—from the 1950s through the 1970s. Television has diversified since then,
especially because of cable, but today it is still mostly an oligopoly of five companies: Disney/ABC, CBS Corporation, NBC Universal, Time Warner, and News Corporation. See Concentration of media ownership.

Worldwide

- The accountancy market is controlled by PriceWaterhouseCoopers, KPMG, Deloitte Touche Tohmatsu, and Ernst & Young (commonly known as the Big Four)
- Three leading food processing companies, Kraft Foods, PepsiCo and Nestle, together form a large proportion of global processed food sales. These three companies are often used as an example of "The rule of 3", which states that markets often become an oligopoly of three large firms.
- Boeing and Airbus have a duopoly over the airliner market\[11\]

DUOPOLY

An oligopoly market structure containing exactly two firms. As an oligopoly, duopoly exhibits the oligopolistic characteristics and undertakes oligopolistic behavior, such as barriers to entry, interdependent actions, and nonprice competition. While duopoly, in its purest form of EXACTLY two firms in the industry, is seldom found in the real world, it does provide an excellent, easy to use illustration of oligopoly. In fact, most instructional analysis of oligopoly generally assumes a two-firm, duopoly market.

Duopoly is a special type of oligopoly market structure that contains only two firms, no more, no less. Duopoly is an ideal model for analyzing oligopoly behavior. With more than one firm, duopoly captures the essence of oligopoly, especially interdependent behavior, while keeping the analysis as simple as possible.

The duopoly model is commonly used to analyze collusion, which results when two firms join together to control the market like a monopoly. Duopoly is ideally suited for collusion analysis. It contains the minimum number of firms needed for an oligopoly and provides all of the insight that would be generated from analyzing three or more firms.

Another model using the duopoly market structure is game theory, which investigates the interdependent actions of two competing firms. Game theory is conceptually and analytically more difficult if more than two firms are included. The essential conclusion can be easily obtained using duopoly.

While the real world seldom contains a pure duopoly market structure, some industries come close. In particular, the duopoly model works well if a market is dominated by two large firms, even though it might contain other smaller ones.

One example that comes close to duopoly is the global market for passenger aircraft. This market is dominated by Boeing (from the United States) and Airbus (from Europe). The vast majority of all passenger airlines use planes manufactured by one of these two firms. Another example might be offered by a small town that contains two grocery stores.
1.6 MONOPOLISTIC COMPETITION

A market structure characterized by a large number of small firms, similar but not identical products sold by all firms, relative freedom of entry into and exit out of the industry, and extensive knowledge of prices and technology. This is one of four basic market structures. The other three are perfect competition, monopoly, and oligopoly. Monopolistic competition approximates most of the characteristics of perfect competition, but falls short of reaching the ideal benchmark that is perfect competition. It is the best approximation of perfect competition that the real world offers. Monopolistic competition is a market structure characterized by a large number of relatively small firms. While the goods produced by the firms in the industry are similar, slight differences often exist. As such, firms operating in monopolistic competition are extremely competitive but each has a small degree of market control.

In effect, monopolistic competition is something of a hybrid between perfect competition and monopoly. Comparable to perfect competition, monopolistic competition contains a large number of extremely competitive firms. However, comparable to monopoly, each firm has market control and faces a negatively-sloped demand curve.

The real world is widely populated by monopolistic competition. Perhaps half of the economy's total production comes from monopolistically competitive firms. The best examples of monopolistic competition come from retail trade, including restaurants, clothing stores, and convenience stores.

Characteristics

The four characteristics of monopolistic competition are: (1) large number of small firms, (2) similar, but not identical products, (3) relatively good, but not perfect resource mobility, and (4) extensive, but not perfect knowledge.

- **Large Number of Small Firms:** A monopolistically competitive industry contains a large number of small firms, each of which is relatively small compared to the overall size of the market. This ensures that all firms are relatively competitive with very little market control over price or quantity. In particular, each firm has hundreds or even thousands of potential competitors.

- **Similar Products:** Each firm in a monopolistically competitive market sells a similar, but not absolutely identical, product. The goods sold by the firms are close substitutes for one another, just not perfect substitutes. Most important, each good satisfies the same basic want or need. The goods might have subtle but actual physical differences or they might only be perceived different by the buyers. Whatever the reason, buyers treat the goods as similar, but different.

- **Relative Resource Mobility:** Monopolistically competitive firms are relatively free to enter and exit an industry. There might be a few restrictions, but not many. These firms are not "perfectly" mobile as with perfect competition, but they are largely unrestricted by government rules and regulations, start-up cost, or other substantial barriers to entry.
• **Extensive Knowledge**: In monopolistic competition, buyers do not know everything, but they have relatively complete information about alternative prices. They also have relatively complete information about product differences, brand names, etc. Each seller also has relatively complete information about production techniques and the prices charged by their competitors.

**Product Differentiation**

The goods produced by firms operating in a monopolistically competitive market are subject to **product differentiation**. The goods are essentially the same, but they have slight differences.

Product differentiation is usually achieved in one of three ways: (1) physical differences, (2) perceived differences, and (3) support services.

- **Physical Differences**: In some cases the product of one firm is physically different from the product of other firms. One good is chocolate, the other is vanilla. One good uses plastic, the other aluminum.
- **Perceived Differences**: In other cases goods are only perceived to be different by the buyers, even though no physical differences exist. Such differences are often created by brand names, where the only difference is the packaging.
- **Support Services**: In still other cases, products that are physically identical and perceived to be identical are differentiated by support services. Even though the products purchased are identical, one retail store might offer "service with a smile," while another provides express checkout.

Product differentiation is the primary reason that each firm operating in a monopolistically competitive market is able to create a little monopoly all to itself.

**Demand and Revenue**

The four characteristics of monopolistic competition mean that a monopolistically competitive firm faces a relatively elastic, but not perfectly elastic, demand curve, such as the one displayed in the exhibit to the right. Each firm in a monopolistically competitive market can sell a wide range of output within a relatively narrow range of prices.

Demand is relatively elastic in monopolistic competition because each firm faces competition from a large number of very, very close substitutes. However, demand is not perfectly elastic (as in perfect competition) because the output of each firm is slightly
different from that of other firms. Monopolistically competitive goods are close substitutes, but not perfect substitutes.

In the exhibit to the right, the monopolistically competitive firm can sell up to 10 units of output within the range of $5.50 to $6.50. Should the price go higher than $6.50, the quantity demanded drops to zero.

A monopolistically competitive firm is a price maker, with some degree of control over price. Once again, unlike perfect competition, a monopolistically competitive firm has the ability to raise or lower the price a little, not much, but a little. And like monopoly, the price received by a monopolistically competitive firm (which is also the firm’s average revenue) is greater than its marginal revenue.

In the exhibit to the right, the marginal revenue curve (MR) lies below the demand/average revenue curve (D = AR). While marginal revenue is less than price, because demand is relatively elastic, the difference tends to be relatively small. For example, 5 units of output correspond to a $5 price. The marginal revenue for the fifth unit is $4.80, less than price, but not by much.

**Short-Run Production**

The analysis of short-run production by a monopolistically competitive firm provides insight into market supply. The key assumption is that a monopolistically competitive firm, like any other firm, is motivated by profit maximization. The firm chooses to produce the quantity of output that generates the highest possible level of profit, given price, market demand, cost conditions, production technology, etc.

The short-run production decision for monopolistic competition can be illustrated using the exhibit to the right. The top panel indicates the two sides of the profit decision—revenue and cost. The slightly curved green line is total revenue. Because price depends on quantity, the total revenue curve is not a straight line. The curved red line is total cost. The difference between total revenue and total cost is profit, which is illustrated in the lower panel as the brown line.

A firm maximizes profit by selecting the quantity of output that generates the greatest gap between the total revenue line and the
total cost line in the upper panel, or at the peak of the profit curve in the lower panel. In this example, the profit maximizing output quantity is 6. Any other level of production generates less profit.

**Long-Run Production**

In the long run, with all inputs variable, a monopolistically competitive industry reaches equilibrium at an output that generates economies of scale or increasing returns to scale. At this level of output, the negatively-sloped demand curve is tangent to the negatively-sloped segment of the long run-average cost curve.

This is achieved through a two-fold adjustment process.

- The first of the folds is entry and exit of firms into and out of the industry. This ensures that firms earn zero economic profit and that price is equal to average cost.
- The second of the folds is the pursuit of profit maximization by each firm in the industry. This ensures that firms produce the quantity of output that equates marginal revenue with short-run and long-run marginal cost.

Because a monopolistically competitive firm has some market control and faces a negatively-sloped demand curve, the end result of this long-run adjustment is two equilibrium conditions:

\[ MR = MC = LRMC \]

\[ P = AR = ATC = LRAC \]

With marginal revenue equal to marginal cost, each firm is maximizing profit and has no reason to adjust the quantity of output or factory size. With price equal to average cost, each firm in the industry is earning only a normal profit. Economic profit is zero and there are no economic losses, meaning no firm is inclined to enter or exit the industry.

These conditions are satisfied separately. However, because price is not equal to marginal revenue, the two equations are not equal (unlike perfect competition). This further means that monopolistic competition does NOT achieve long-run equilibrium at the minimum efficient scale of production.

**Real World (In) Efficiency**

A monopolistically competitive firm generally produces less output and charges a higher price than would be the case for a perfectly competitive industry. In particular, the price charged by a monopolistically competitive firm is greater than its marginal cost.

The inequality of price and marginal cost violates the key condition for efficiency. Resources are NOT being used to generate the highest possible level of satisfaction.
The reason for this inefficiency is found with market control. Because a monopolistically competitive firm has control over a small slice of the market, it faces a negatively-sloped demand curve and price is greater than marginal revenue, which is set equal to marginal cost when maximizing profit.

While monopolistic competition is technically inefficient, it tends to be less inefficient than other market structures, especially monopoly. Even though price is greater than marginal revenue (and thus marginal cost), because the demand curve is relatively elastic, the difference is often relatively small.

For example, a monopoly that charges a $100 price while incurring a marginal cost of $20 creates a serious inefficiency problem. In contrast, the inefficiency created by a monopolistically competitive firm that charges a $50 price while incurring a marginal cost of $49.95 is substantially less.

The closer marginal revenue is to price, the closer a monopolistically competitive firm comes to allocating resources according to the efficiency benchmark established by perfect competition.

In the grand scheme of economic problems, the inefficiency created by monopolistic competition seldom warrants much attention... and deservedly so.

Activity 1

1. What do you understand by term ‘market structure’?
2. Differentiate between perfect and imperfect competition.
3. Give a brief note on monopolistic competition.
4. Explain how oligopoly is different from monopoly. Discuss main characteristics of both.

1.7 SUMMARY

In this unit we have discussed main market structures that are very important to be considered in every decision making pertaining to economic decisions. Perfect Competition is characterized by a large number of relatively small competitors, each with no market control. Perfect competition is an idealized market structure that provides a benchmark for efficiency. Monopoly is characterized by a single competitor and extensive market control. Monopoly contains a single seller of a unique product with no close substitutes. The demand for monopoly output is the market demand. Oligopoly is characterized by a small number of relatively large competitors, each with substantial market control. A substantial number of real world markets fit the characteristics of oligopoly.

Monopolistic Competition is residing closer to perfect competition, characterized by a large number of relatively small competitors, each with a modest degree of market control. A substantial number of real world markets fit the characteristics of
monopolistic competition. In depth discussion about all these market structures was given along with suitable examples.

1.8 FURTHER READINGS

UNIT 2

PRICE AND OUTPUT DETERMINATION UNDER MONOPOLY

Objectives

After reading this unit, you should be able to:

- Understand the concepts of price, output and equilibrium in context of monopoly
- Appreciate the relevance marginal analysis in monopoly
- Know the approaches to production analysis and profit maximization in monopoly
- Identify the concepts of price discrimination.
- Have deep understanding of practices to control and regulate monopolies across the globe.

Structure

2.1 Introduction
2.2 Monopoly equilibrium
2.3 Monopoly marginal analysis
2.4 Monopoly short run production analysis
2.5 Monopoly profit maximization
2.6 Price discrimination
2.7 Welfare aspects of monopoly
2.8 Monopoly and restrictive trade practices in India
2.9 Summary
2.10 Further readings

2.1 INTRODUCTION

The four key characteristics of monopoly as discussed earlier are: (1) a single firm selling all output in a market, (2) a unique product, (3) restrictions on entry into and exit out of the industry, and more often than not (4) specialized information about production techniques unavailable to other potential producers.

These four characteristics mean that a monopoly has extensive (boarding on complete) market control. Monopoly controls the selling side of the market. If anyone seeks to acquire the production sold by the monopoly, then they must buy from the monopoly. This means that the demand curve facing the monopoly is the market demand curve. They are one and the same.

The characteristics of monopoly are in direct contrast to those of perfect competition. A perfectly competitive industry has a large number of relatively small firms, each
producing identical products. Firms can freely move into and out of the industry and share the same information about prices and production techniques.

A monopolized industry, however, tends to fall far short of each perfectly competitive characteristic. There is one firm, not a lot of small firms. There is only one firm in the market because there are no close substitutes, let alone identical products produced by other firms. A monopoly often owes its monopoly status to the fact that other potential producers are prevented from entering the market. No freedom of entry here. Neither is there perfect information. A monopoly firm often has specialized information, such as patents or copyrights, that are not available to other potential producers.

2.2 MONOPOLY EQUILIBRIUM

In order to study equilibrium under monopoly let us draw the demand and supply or cost curves of a monopolist.

In Figure 1 AR and MR are the demand and marginal revenue curves of a monopolist. AC and MC are the respective cost or supply curves. The usual equilibrium of \( MR = MC \) is equally applicable to the monopolist. In the figure MR and MC have intersected at point \( e \) which is the equilibrium point. At this point the monopolist produces and supplies output quantity \( Q \). This is the only profit-maximizing condition for the monopolist. Under the given demand-cost structure no other level of output can help to enhance his profit.

In an equilibrium the monopolist charges price \( P \) which is determined by a corresponding point \( R \) on the average revenue curve. The total revenue of the monopolist is then,

\[
TR = OQ \times P = OQRP
\]
Similarly the total cost of the monopolist is governed by a point on the average cost curve. S or C is the average cost of producing output Q in which the total cost will be

$$TC = OQ \times AC = OQSC$$

The profits of the monopolist as the difference between TR and TC are,

$$\text{Profits} = TR - TC = OQRP - OQSC = CSRP$$

Hence CSRP are the monopoly profits. These profits look similar to Super Normal profits under competition. Monopoly profits differ in two respects:

i) Monopoly profits are permanent and enjoyed in the short as well as long run. There is no fear of monopoly profits being competed away.

ii) Monopoly profits arise out of control over conditions in the market. The monopolist follows restrictive policies and charges a higher price. This is the source of his profits. It is made clear by a downward sloping demand curve. Competitive Super Normal profits, on the other hand, are the result of more efficient and favorable conditions of production. Whether a monopolist will always earn extra profits or be satisfied with normal profits depends upon the technical cost conditions of a monopolist and the flexibility of the demand curve. By nature, a monopolist is not likely to allow his profits to fall. He will maintain some positive profits through restrictive practices.

**Monopoly Supply Curve**

The behavior of the monopoly demand curve is distinct from the demand curve under competition. The supply curve of a monopolist is similar to that of a competitive firm. Supply is governed by the technical conditions of production. There is no reason why these should be different for a monopolist. Hence supply curve of a monopolist depends upon the behavior of the usual average and marginal cost of production. With such cost curves and a downward sloping demand curve let us attempt an equilibrium analysis of a monopolist.

Market control means that monopoly does not have a supply relation between the quantity of output produced and the price. In contrast, the short-run supply curve a perfectly competitive is that portion of its marginal cost curve that lies above the minimum of the average variable cost curve. However, because monopoly does not set price equal to marginal revenue, it does NOT equate marginal cost and price. For this reason, a monopoly firm does not respond to price changes by moving along its marginal cost curve. A monopoly does not necessarily supply larger quantities at higher prices or smaller quantities at lower prices.

Monopoly maximizes profit by producing the quantity of output that equates marginal revenue and marginal cost. In that price (and average revenue) is greater than marginal revenue for a monopoly, price is also greater than marginal cost. Monopoly does not
produce output by moving up and down along its **marginal cost curve**. The marginal cost curve is thus not the **supply curve** for monopoly.

As a **price maker** that controls the market, monopoly reacts to demand conditions, especially the **price elasticity of demand**, when setting the price and corresponding quantity produced. While it is not out of the question that monopoly offers a larger quantity for sale at a higher price, it is also conceivable that it offers a smaller quantity at a higher price or a larger quantity at a lower price.

Consider the production and sale of Amblathan-Plus, the only cure for the deadly (but hypothetical) foot ailment known as amblathanitis. This drug is produced by the noted monopoly firm, Feet-First Pharmaceutical.

A typical profit-maximizing output determination using the marginal revenue and marginal cost approach is presented in this diagram. Feet-First Pharmaceutical maximizes profit by producing output that equates marginal revenue and marginal cost, which is 6 ounces of Amblathan-Plus in this example. The corresponding price charged is $7.50.

What might happen, however, with a change in demand—in particular, a demand that leads to a higher price? For a perfectly competitive market, a higher price induces a larger quantity supplied as the market moves from one **equilibrium** to another, along a positively-sloped supply curve. The supply curve is positively sloped because it is comprised of positively-sloped marginal cost curves for every firm in the market.

Does this also happen for a monopoly? It might. But then again, it might not. Click the [Demand Shift] button to illustrate this shift. Note that the demand curve does not simply shift, slope (and elasticity) also change. This is the key.

Check out the new equilibrium. The monopoly seeks the production level that maximizes profit, just as it did before, by equating marginal revenue and marginal cost. This production level is 5 ounces, which is LESS than the original 6 ounce production level. The subsequent price charged by Feet-First Pharmaceutical is $9, obviously greater than the original $7.50 price.
2.3 MONOPOLY, MARGINAL ANALYSIS

A monopoly produces the profit-maximizing quantity of output that equates marginal revenue and marginal cost. This marginal approach is one of three methods that used to determine the profit-maximizing quantity of output. The other two methods involve the direct analysis of economic profit or a comparison of total revenue and total cost. Monopoly is a market in which a single firm is the only supplier of the good. Anyone seeking to buy the good must buy from the monopoly seller. This single-seller status gives monopoly extensive market control—a price maker that faces a negatively-sloped demand curve. With this negatively-sloped demand curve, marginal revenue is less than average revenue and price.

Comparable to any profit-maximizing firm, a monopoly produces the quantity of output in the short run that generates the maximum difference between total revenue and total cost, which is economic profit. This profit maximizing level of production is also achieved by the equality between marginal revenue and marginal cost. At this production level, the firm cannot increase profit by changing the level of production. The analysis of marginal revenue and marginal cost can be achieved through a table of numbers or with marginal revenue and marginal cost curves.

A monopoly is presumed to produce the quantity of output that maximizes economic profit—the difference between total revenue and total cost. This decision can be analyzed using the exhibit below. This table presents revenue and cost information for Feet-First Pharmaceutical, a hypothetical example of a monopoly, for the production and sale of Amblathan-Plus, the only cure for the deadly (but hypothetical) foot ailment known as amblathanitis.

Because Feet-First Pharmaceutical produces a unique product it has extensive market control and sells its Amblathan-Plus according to the market demand. To sell a larger quantity, it must lower the price. Feet-First Pharmaceutical's status as a monopoly firm is reflected in this table.

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Price</th>
<th>Total Revenue</th>
<th>Marginal Revenue</th>
<th>Total Cost</th>
<th>Marginal Cost</th>
<th>Profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10.50</td>
<td>0</td>
<td>10</td>
<td>-10</td>
<td>-10</td>
<td>-10</td>
</tr>
<tr>
<td>1</td>
<td>10.00</td>
<td>10</td>
<td>17</td>
<td>-7</td>
<td>-7</td>
<td>-7</td>
</tr>
<tr>
<td>2</td>
<td>9.50</td>
<td>19</td>
<td>22</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
</tr>
<tr>
<td>3</td>
<td>9.00</td>
<td>27</td>
<td>26</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>8.50</td>
<td>34</td>
<td>29.5</td>
<td>4.5</td>
<td>4.5</td>
<td>4.5</td>
</tr>
<tr>
<td>5</td>
<td>8.00</td>
<td>40</td>
<td>33</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>7.50</td>
<td>45</td>
<td>37</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>7</td>
<td>7.00</td>
<td>49</td>
<td>42</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>6.50</td>
<td>52</td>
<td>49</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>6.00</td>
<td>54</td>
<td>59</td>
<td>-5</td>
<td>-5</td>
<td>-5</td>
</tr>
<tr>
<td>10</td>
<td>5.50</td>
<td>55</td>
<td>73</td>
<td>-18</td>
<td>-18</td>
<td>-18</td>
</tr>
<tr>
<td>11</td>
<td>5.00</td>
<td>55</td>
<td>92</td>
<td>-37</td>
<td>-37</td>
<td>-37</td>
</tr>
<tr>
<td>12</td>
<td>4.50</td>
<td>54</td>
<td>117</td>
<td>-63</td>
<td>-63</td>
<td>-63</td>
</tr>
</tbody>
</table>
could produce more than 12 ounces, this range is sufficient for the present analysis.

- **Price:** The second column presents the price received by Feet-First Pharmaceutical for selling Amblathan-Plus. As a price maker, the first and second columns represent the market demand for Amblathan-Plus. The price Feet-First Pharmaceutical faces ranges from a high of $10.50 per ounce for a zero quantity to a low of $4.50 per ounce for 12 ounces. Feet-First Pharmaceutical can sell a larger quantity of Amblathan-Plus, but only by reducing the price. Feet-First Pharmaceutical is a price maker.

- **Total Revenue:** Total revenue is presented in the third column. This indicates the revenue Feet-First Pharmaceutical receives at each level of Amblathan-Plus production. It is derived as the quantity in the first column multiplied by the price in the second column. Total revenue ranges from $0 if no output is sold to a high of $55 for selling 10 or 11 ounces of Amblathan-Plus. For example, selling 4 ounces of Amblathan-Plus generates $34 of revenue and selling 7 ounces leads to $49 of revenue.

- **Total Cost:** The fifth column presents the total cost incurred by Feet-First Pharmaceutical in the production of Amblathan-Plus, ranging from a low of $10 for zero output (which is fixed cost) to a high of $117 for 12 ounces. Total cost continues to rise beyond 12 ounces, but this information is not needed. Producing 1 ounce of Amblathan-Plus incurs a total cost of $17. Producing 2 ounces of Amblathan-Plus incurs a total cost of $22. Total cost rises as Feet-First Pharmaceutical produces more.

- **Profit:** The seventh column at the far right of the table displays economic profit, the difference between total revenue in the third column and total cost in the fifth column. It starts at -$10, rises to $8, then falls to -$63.

The task is to determine which Amblathan-Plus production level provides the maximum profit using marginal revenue and marginal cost. The process goes something like this:

1. Marginal revenue indicates how much total revenue changes by producing one more or one less unit of output.
2. Marginal cost indicates how much total cost changes by producing one more or one less unit of output.
3. Profit increases if marginal revenue is greater than marginal cost and profit decreases if marginal revenue is less than marginal cost.
4. Profit neither increases nor decreases if marginal revenue is equal to marginal cost.
5. As such, the production level that equates marginal revenue and marginal cost is profit maximization.

The fourth and sixth columns are reserved for these two marginal measures. To display the numbers, click the [Marginals] button.

- **Marginal Revenue:** The fourth column displays marginal revenue, which declines from a high of $10 per ounce for the first ounce of Amblathan-Plus to a low of -
$1 per ounce for the twelfth ounce. Because Feet-First Pharmaceutical is a monopoly with market control, marginal revenue is less than price.

- **Marginal Cost:** The sixth column presents the marginal cost that Feet-First Pharmaceutical incurs in the short run for the production of Amblathan-Plus. It starts at $7, declines to a low of $3.50, then rises to $25. The declining values are the result of increasing marginal returns and the rising values are due to decreasing marginal returns and the law of diminishing marginal returns.

How can these marginals be used to identify the profit-maximizing output level?

- First, a quick look at the profit column indicates that the profit-maximizing production is 6 ounces of Amblathan-Plus, which generates a profit of $8. Click the [Profit Max] button to highlight this result.
- Second, with this outcome highlighted, note the corresponding marginal revenue and marginal cost values. Marginal revenue is $5 for the sixth ounce of Amblathan-Plus production and marginal cost is $4.

Feet-First Pharmaceutical increases production from 5 ounces to 6 ounces because doing so generates $5 of extra revenue and incurs only $4 of extra cost, meaning profit increases by $1 over the production of 5 ounces. Feet-First Pharmaceutical does not increase production from 6 ounces to 7 ounces because doing so generates only $4 of extra revenue but incurs $5 of extra cost, meaning profit decreases by $1.

As such, Feet-First Pharmaceutical settles in with the production and 6 ounces of Amblathan-Plus and achieves maximum profit. It cannot increase profit by changing production.

- Third, while marginal revenue and marginal cost might not appear to be equal for the profit-maximizing 6 ounces of Amblathan-Plus production ($5 versus $4), they really are. The reason is that the marginal numbers in the table actually represent discrete changes from one ounce to the next. Reducing the size of the discrete change, say from 5.9999 ounces to 6 ounces, results in marginal revenue and marginal cost that are actually closer to $4.50. At the limit of an infinitesimally small change, marginal revenue and marginal cost are exactly $4.50.

While the equality between marginal revenue and marginal cost shows up better in a graph, the best practical method of identifying similar results, with a table of numbers, is to average the discrete changes on either side of the quantity. For example, the marginal cost AT the sixth ounce of Amblathan-Plus production is the average of the change from 5 to 6 ($4) and from 6 to 7 ($5), which is $4.50. Marginal revenue AT the sixth ounce is the average of the change from 5 to 6 ($5) and from 6 to 7 ($4), which is also $4.50.
The short-run production decision for a monopoly can be graphically illustrated using marginal revenue and marginal cost curves. The exhibit to the right is standing poised to display these curves.

- Average Revenue: First up is the average revenue curve, which can be seen with a click of the [Average Revenue] button. Because Feet-First Pharmaceutical is a monopoly, this average revenue curve is the market demand curve for Amblathan-Plus, which is negatively-sloped due to the law of demand.
- Marginal Revenue: A click of the [Marginal Revenue] button reveals the green line labeled MR that depicts the marginal revenue Feet-First Pharmaceutical receives from Amblathan-Plus production. Because Feet-First Pharmaceutical is a price maker, this marginal revenue curve is also a negatively-sloped line, and it lies beneath the average revenue (market demand) curve.
- Marginal Cost: A click of the [Marginal Cost] button reveals a red U-shaped curve labeled MC that represents the marginal cost Feet-First Pharmaceutical incurs in the production of Amblathan-Plus. The shape is based on increasing, then decreasing marginal returns.

The key for Feet-First Pharmaceutical is to identify the production level that gives the greatest level of economic profit. Profit is maximized at the quantity of output found at the intersection of the marginal revenue and marginal cost curves, which is 6 ounces of Amblathan-Plus. Click the [Profit Max] button to highlight this production level.

To demonstrate why the equality between marginal revenue and marginal cost is the profit-maximizing production level, consider what results if marginal revenue is not equal to marginal cost:

- If marginal revenue is greater than marginal cost, as is the case for small quantities of output, then the firm can increase profit by increasing production. Extra production adds more to revenue than to cost, so profit increases.
- If marginal revenue is less than marginal cost, as is the case for large quantities of output, then the firm can increase profit by decreasing production. Reducing production reduces revenue less than it reduces cost, so profit increases.
- If marginal revenue is equal to marginal cost, then the firm cannot increase profit by producing more or less output. Profit is maximized.
Once the profit maximizing output is revealed, the last step is to identify the price charged by the monopoly. This is easily accomplished by clicking the [Price] button. Price is found by extending the 6-ounce quantity upward to the average revenue curve, which is the market demand. Buyers are willing to pay $7.50 per ounce for Amblathan-Plus if 6 ounces are sold.

Note that this $7.50 price is greater than the $4.50 marginal cost, which indicates that monopoly does not achieve the price-equals-marginal-cost condition for efficiency.

### 2.4 MONOPOLY, SHORT-RUN PRODUCTION ANALYSIS

A monopoly produces the profit-maximizing quantity of output that equates marginal revenue and marginal cost. This production level can be identified using total revenue and cost, marginal revenue and cost, or profit. Because a monopoly faces a negatively-sloped demand curve, it does not efficiently allocate resources by equating price and marginal cost.

Monopoly is a market structure characterized by a single seller of a unique product that has no close substitutes. These conditions mean a monopoly has complete control of the supply side of the market, which also means that the negatively-sloped market demand curve is the demand curve facing the monopoly. With this demand curve, marginal revenue is less than both average revenue and price.

Comparable to any profit-maximizing firm, a monopoly produces the quantity of output in the short run that equates marginal revenue with marginal cost. At this production level, the firm cannot increase profit by changing the level of production. Because price is not equal to marginal revenue, a monopoly does not produce the quantity of output that equates price and marginal cost, which means it is not efficiently allocating resources.

**Determining Output**

A monopoly is presumed to produce the quantity of output that maximizes economic profit—the difference between total revenue and total cost. This production decision can be analyzed in three different, but interrelated ways.

- **Profit:** The first is directly through the analysis of economic profit, especially using a profit curve that traces the level of economic profit for different levels of output.
- **Total Revenue and Cost:** A second is by comparing total revenue and total cost, commonly accomplished with total revenue and total cost curves.
- **Marginal Revenue and Cost:** The third, and perhaps most noted, way is by comparing marginal revenue and marginal cost, similarly achieved with marginal revenue and marginal cost curves.

The profit-maximization production decision facing a monopoly can be analyzed using the exhibit below. This table presents revenue and cost information for Feet-First Pharmaceutical, a hypothetical example of a monopoly by virtue of its exclusive control.
over the supply of Amblathan-Plus, the only cure for the deadly (but hypothetical) foot ailment known as amblathanitis.

- **Quantity:** The first column presents the quantity of output produced, ranging from 0 to 12 ounces of Amblathan-Plus.

- **Price:** The second column presents the price received by Feet-First Pharmaceutical for selling Amblathan-Plus.

As a price maker, the first and second columns represent the market demand for Amblathan-Plus. The price Feet-First Pharmaceutical faces ranges from a high of $10.50 per ounce for a zero quantity to a low of $4.50 per ounce for 12 ounces.

- **Total Revenue:** The third column presents total revenue, ranging from a low of $0 for no output to a high of $55 for 10 and 11 ounces, before declining for 12 ounces.

- **Marginal Revenue:** The fourth column presents marginal revenue, the change in total revenue due to a change in the quantity, which declines from a high of $10 to a low of -$1.

- **Total Cost:** The fifth column presents the total cost incurred by Feet-First Pharmaceutical in the production of this Amblathan-Plus, ranging from a low of $10 for zero output (which is also fixed cost) to a high of $117 for 12 ounces of Amblathan-Plus.

- **Marginal Cost:** The sixth column presents the marginal cost, the change in total cost due to a change in the quantity, which declines from $7, reaches a low of $3.50, then rises until reaching $25.

- **Profit:** The seventh column at the far right of the table is profit, the difference between total revenue in the second column and total cost in the third column. It begins at -$11, rises to $8, then falls to -$63.

The choice facing Feet-First Pharmaceutical is to produce the quantity of Amblathan-Plus that maximizes profit. This can be easily identified using the far right "Profit" column. The quantity that generates the greatest level of economic profit is 6 ounces of Amblathan-Plus. This alternative can be highlighted by clicking the [Profit Max] button.

The production of 6 ounces of Amblathan-Plus results in $45 of total revenue and $37 of total cost, a difference of $8. No other production level generates a greater economic profit. Producing one more ounce of Amblathan-Plus or one less ounce of Amblathan-Plus reduces profit to $7. The price that achieves this profit-maximizing quantity is $7.50.

**Table 2**

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Price</th>
<th>Total Revenue</th>
<th>Marginal Revenue</th>
<th>Total Cost</th>
<th>Marginal Cost</th>
<th>Profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10.50</td>
<td>0</td>
<td>10</td>
<td>-10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>10.00</td>
<td>10</td>
<td>10</td>
<td>17</td>
<td>7</td>
<td>-7</td>
</tr>
<tr>
<td>2</td>
<td>9.50</td>
<td>19</td>
<td>9</td>
<td>22</td>
<td>5</td>
<td>-3</td>
</tr>
<tr>
<td>3</td>
<td>9.00</td>
<td>27</td>
<td>8</td>
<td>26</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>8.50</td>
<td>34</td>
<td>7</td>
<td>29.5</td>
<td>3.5</td>
<td>4.5</td>
</tr>
<tr>
<td>5</td>
<td>8.00</td>
<td>42</td>
<td>6</td>
<td>33</td>
<td>3.5</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>7.50</td>
<td>45</td>
<td>5</td>
<td>37</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>7.00</td>
<td>49</td>
<td>4</td>
<td>42</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>6.50</td>
<td>52</td>
<td>3</td>
<td>49</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>6.00</td>
<td>54</td>
<td>2</td>
<td>59</td>
<td>10</td>
<td>-5</td>
</tr>
<tr>
<td>10</td>
<td>5.50</td>
<td>55</td>
<td>1</td>
<td>73</td>
<td>14</td>
<td>-18</td>
</tr>
<tr>
<td>11</td>
<td>5.00</td>
<td>55</td>
<td>0</td>
<td>92</td>
<td>19</td>
<td>-37</td>
</tr>
<tr>
<td>12</td>
<td>4.50</td>
<td>54</td>
<td>-1</td>
<td>117</td>
<td>25</td>
<td>-63</td>
</tr>
</tbody>
</table>
While marginal revenue and marginal cost might not appear to be equal for the profit-maximizing 6 ounces of Amblathan-Plus production ($5 versus $4), they really are. The reason is that the marginal numbers in the table actually represent discrete changes from one ounce to the next. Reducing the size of the discrete change, say from 5.9999 ounces to 6 ounces, results in a marginal cost that is closer to $4.50. The same is true for marginal cost. At the limit of an infinitesimally small change, both marginal cost and marginal revenue are exactly $4.50.

**Short-Run Supply Curve?**

Market control means that monopoly does necessarily not have a supply relation between the quantity of output produced and the price. By way of comparison a perfectly competitive firm has a short-run supply curve based on the positively sloped marginal cost curve. A perfectly competitive firm's supply curve is that portion of the marginal cost curve that lies above the minimum of the average variable cost curve. A perfectly competitive firm maximizes profit by producing the quantity of output that equates price and marginal cost. As such, the firm moves along the marginal cost curve in response to alternative prices.

However, market control means that price is NOT equal to marginal revenue, and thus monopoly does NOT equate marginal cost and price. As such, a monopoly firm does not move along the marginal cost curve. A monopoly does not necessarily supply larger quantities at higher prices or smaller quantities at lower prices.

As a price maker that controls the market, a monopoly reacts to demand conditions, especially the price elasticity of demand, when setting the price and corresponding quantity produced. While it is not out of the question for a monopoly to supply a larger quantity at a higher price, it is also conceivable that a monopoly produces a smaller quantity at a higher price or a larger quantity at a lower price.

The bottom line is that monopoly does not necessarily have a short-run supply curve relation.
(In)Efficiency
Because a monopoly charges a price that exceeds marginal cost, it does not efficiently allocate resources. Price represents the value of goods produced and marginal cost represents the value of goods not produced. When both are equal, satisfaction cannot be increased by changing production.

However, because price exceeds marginal cost, the economy gives up less satisfaction from other goods not produced than it receives from the good that is produced. The economy can gain satisfaction by producing more of the good.

This exhibit once again displays the profit-maximizing production output of Feet-First Pharmaceutical. Feet-First Pharmaceutical produces the quantity (6 ounces) of Amblathan-Plus that equates marginal revenue and marginal cost. The resulting price charge is $7.50 per ounce.

This price exceeds marginal cost, meaning Feet-First Pharmaceutical is inefficient. The price is higher and the quantity produced is less than what would be achieved with an efficient production.

To see why, consider how an efficient allocation would look. This alternative can be displayed with a click of the [Efficient] button. This efficient production level occurs at the intersection of the marginal cost curve (MC) and the demand/average revenue curve (AR). The efficient price is just under $7 per ounce and the efficient quantity produced is almost 7.5 ounces. At this production level, price is equal to marginal cost.

In comparison, the price charged by the monopoly is higher and the quantity produced is less.

2.5 MONOPOLY, PROFIT MAXIMIZATION

A monopoly is presumed to produce the quantity of output that maximizes economic profit—the difference between total revenue and total cost. This production decision can be analyzed directly with economic profit, by identifying the greatest difference between total revenue and total cost, or by the equality between marginal revenue and marginal cost.
The profit-maximizing level of output is a production level that achieves the greatest level of economic profit given existing market conditions and production cost. For a monopoly, this entails adjusting the price and corresponding production level to achieved the desired match between total revenue and total cost.

**Three Views**
Profit-maximizing output can be identified in one of three ways--directly with economic profit, with a comparison of total revenue and total cost, and with comparison of marginal revenue and marginal cost.

This exhibit illustrates how it can be identified for a monopoly, such as that operated by Feet-First Pharmaceutical, a well-known monopoly supplier of Amblathan-Plus, the only cure for the deadly (but hypothetical) foot ailment known as amblathanitis. Feet-First Pharmaceutical is the exclusive producer of Amblathan-Plus, meaning that it is a price maker and the demand curve it faces is the market demand curve.

The top panel presents the profit curve. The middle panel presents total revenue and total cost curves. The bottom panel presents average revenue and average total cost curves. In all three panels, Feet-First Pharmaceutical maximizes when producing 6 ounces of Amblathan-Plus.

- **Profit**: First, profit maximization can be illustrated with a direct evaluation of profit. If the profit curve is at its peak, then profit is maximized. In the top panel, the profit curve achieves its highest level at 6 ounces of Amblathan-Plus. At other output levels, profit is less.
- **Total Revenue and Total Cost**: Second, profit maximization can be identified by a comparison of total revenue and total cost. The quantity of output that achieves
the greatest difference of total revenue over total cost is profit maximization. In the middle panel, the vertical gap between the total revenue and total cost curves is the greatest at 6 ounces of Amblathan-Plus. For smaller or larger output levels, the gap is either less or the total cost curve lies above the total revenue curve.

- Marginal Revenue and Marginal Cost: Third, profit maximization can be identified by a comparison of marginal revenue and marginal cost. If marginal revenue is equal to marginal cost, then profit cannot be increased by changing the level of production. Increasing production adds more to cost than revenue, meaning profit declines. Decreasing production subtracts more from revenue than from cost, meaning profit also declines. In the bottom panel, the marginal revenue and marginal cost curves intersect at 6 ounces of Amblathan-Plus. At larger or smaller output levels, marginal cost exceeds marginal revenue or marginal revenue exceeds marginal cost.

### 2.6 PRICE DISCRIMINATION

The act of selling the same good to different buyers for different prices that are not justified by different production costs. This is practiced by suppliers who have achieved some degree of market control, especially monopoly. Common examples of price discrimination are electricity rates, long-distance telephone charges, movie ticket prices, airplane ticket prices, and assorted child or senior citizen discounts. Price discrimination takes the form of one of three degrees: (1) first degree, in which each price is the maximum price that buyers are willing and able to pay, (2) second degree, in which price is based on the quantity sold, and (3) third degree, in which prices are based on an easily identifiable characteristic of the buyer.

Price discrimination is a technique used by sellers to extract the maximum price possible from buyers, which has the obvious and direct implications of increasing the revenue and profit received by the seller. This is accomplished by transferring consumer surplus from buyers to sellers.

#### Three Conditions

To be a successful price discriminator a seller must satisfy three things: (1) to have market control and be a price maker, (2) to identify two or more groups that are willing to pay different prices, and (3) to keep the buyers in one group from reselling the good to another group. In this way, a seller is able to charge each group what they, and they alone, are willing to pay.

- **Market Control:** First and foremost, a seller must be able to control the price. Monopoly is quite adept at price discrimination because it is a price maker, it can set the price of the good. Oligopoly and monopolistic competition can undertake price discrimination to the extent that they are able to control the price. Perfect competition, with no market control, does not do well in the price discrimination arena.

- **Different Buyers:** The second condition is that a seller must be able to identify different groups of buyers, and each group must have a different price elasticity of demand. The different price elasticity means that buyers are willing and able to
pay different prices for the same good. If buyers have the same elasticity and are willing to pay the same price, then price discrimination is pointless. The price charged to each group is the same in this case.

- Segmented Buyers: Lastly, price discrimination requires that each group of buyers be segmented and sealed into distinct markets. Segmentation means that the buyers in one market cannot resell the good to the buyers in another market. Price discriminate is not effective if trade among groups is possible. Those buyers charged a higher price cannot purchase the good from those paying the lower price instead of from the seller.

Three Degrees
Price discrimination can take one of three forms (or degrees):

- First-Degree Price Discrimination: Also termed perfect price discrimination, this form exists when a seller is able to sell each quantity of a good for the highest possible price that buyers are willing and able to pay. In other words, ALL consumer surplus is transferred from buyers to the seller.

- Second-Degree Price Discrimination: Also termed block pricing, this form occurs when a seller charges different prices for different quantities of a good. Such discrimination is possible because the different quantities are purchased by different types of buyers with different demand elasticities. Block pricing of electricity, in which electricity prices depend on the amount used, is the most common example. The key is that regular households tend to use very little electricity compared to retail stores, which uses less compared to large manufacturing firms.

- Third-Degree Price Discrimination: This is the most common of price discrimination. It occurs when the seller is able to separate buyers based on an easily identifiable characteristic, such as age, location, gender, and ethnic group. Senior citizen discounts are a common example. Higher gasoline prices near highways versus inside cities is another.

Two Groups, Two Prices
The general process of price discrimination can be illustrated using the hypothetical Shady Valley Cinemaplex, which has 20 movie screens and is the only movie theater in the greater metropolitan Shady Valley community. This gives the Cinemaplex extensive market control. It also provides the opportunity to practice price discrimination.

With market control, the Cinemaplex has the ability charge different prices to different "types" of movie goers. Suppose the Cinemaplex decision makers have evidence that two distinct groups of ticket buyers have distinctively different price elasticities of demand. For the sake of simplicity, other groups of movie goers are ignored in this analysis.

- Youthful movie patrons, those in the 13 to 18 year age group, are not very responsive to ticket prices. In other words, they have a less elastic demand and are willing to pay relatively high prices to see a movie.
By way of contrast, older folks, those in the 50-plus age range, are more selective in their ticket purchases and thus have a relatively elastic demand. They are not as inclined to pay high ticket prices.

Age is the key factor that lets the Cinemaplex segment and seals each market demand. Youngsters can be charged one price and given a ticket that only admits youthful individuals. Oldsters can be charged another price and given tickets that only admits elderly individuals. Tickets cannot be traded between the groups because only those with the proper aged-based tickets are allowed entry into the Cinemaplex. The goal of the Cinemaplex, like any firm, is to maximize profit. It does this by equating marginal revenue and marginal cost. The curve labeled MC in the exhibit to the right is the marginal cost curve for the production of this good. This is one side of the profit-maximizing decision. However, because the Cinemaplex can identify two groups of movie patrons, each with a different price elasticity of demand, there are two different marginal revenue curves (and associated demand curves) on the other side of the decision.

Each of the two groups--the youngsters and the oldsters--has its own demand curve. And with each demand curve comes a corresponding marginal revenue curve. Because the Cinemaplex has the power to control prices and can segment and seal each demand, price discrimination is bound to arise.

Consider how the Cinemaplex sets the price for the two groups.

- First the Youngsters: Click the [Youngsters Demand] button to reveal the demand curve and the corresponding marginal revenue curve for movie patrons between the ages of 13 and 18. The relatively steep demand curve indicates a relatively low elasticity. This group is not very sensitive to price. The Cinemaplex maximizes profit from this group by equating marginal revenue and marginal cost, then charging the corresponding price that the youngsters are willing to pay. Click the [Youngsters Price] to reveal this profit-maximizing solution. The price charged the youngsters is $9 per ticket.

- Next the Oldsters: Click the [Oldsters Demand] button to reveal the demand curve and the corresponding marginal revenue curve for movie patrons over the age of

![Figure 6 Price Discrimination](image-url)
50. The relatively flat demand curve indicates a relatively high elasticity. This group is very sensitive to price.

The Cinemaplex maximizes profit from this group by equating marginal revenue and marginal cost, then charging the corresponding price that the oldsters are willing to pay. Click the [Oldsters Price] to reveal this profit-maximizing solution. The price charged the oldsters is $5.85 per ticket.

The obvious conclusion from this analysis is that the price charged to each group is different. The oldsters with the more elastic demand are charged a lower price and the youngsters with the less elastic demand are charged a higher price. This, of course, is possible because the two groups have different price elasticities of demand. If the two groups have the same elasticity, then they are charged the same price.

If the Cinemaplex can identify other groups, each with a different price elasticity of demand, each that can be separated by an easily identifiable characteristics (gender, height, hair color), then it can extend this price discrimination... and in all likelihood increase its profit.

2.7 WELFARE ASPECTS OF MONOPOLY

The debate about monopoly will never be settled! The consensus seems to be that the economic case for and against monopoly needs to be judged on a case by case basis - particularly when assessing the impact on economic welfare.

The standard economic case against monopoly is that, with the same cost structure, a monopoly supplier will produce at a lower output and charge a higher price than a competitive industry. This leads to a net loss of economic welfare and efficiency because price is driven above marginal cost - leading to allocative inefficiency.

The diagram below shows how price and output differ between a competitive and a monopolistic industry. We have assumed that the cost structure for both the competitive firm and the monopoly is the same - indeed we have assumed that output can be supplied at a constant marginal and average cost.
Assuming that the monopolist seeks to maximise profits and that they take the whole of the market demand curve, then the price under monopoly will be higher and the output lower than the competitive market equilibrium. This leads to a deadweight loss of consumer surplus and therefore a loss of static economic efficiency.

### 2.7.1 CAN MONOPOLY BE DEFENDED?

#### Monopoly and Economies of Scale
Because monopoly producers are often supplying goods and services on a very large scale, they may be better placed to take advantage of economies of scale - leading to a fall in the average total costs of production. These reductions in costs will lead to an increase in monopoly profits but some of the gains in productive efficiency might be passed onto consumers in the form of lower prices. The effect of economies of scale is shown in the diagram above.

Economies of scale provide potential gains in economic welfare for both producers and consumers.

2.7.2 Regulation of monopoly

Because of the potential economic welfare loss arising from the exploitation of monopoly power, the Government regulates some monopolies. Regulators can control annual price increases and introduce fresh competition into particular industries.

Monopoly and Innovation (Research and Development)

How are the supernormal profits of monopoly used? Is consumer surplus of equal value to producer surplus?

Are large-scale firms required to create a comparative advantage in global markets? Some economists argue that large-scale firms are required to be competitive in international markets.

An important issue is what happens to the monopoly profits both in the short run and the long run. Undoubtedly some of the profits will be distributed to shareholders as dividends. This raises questions of equity. Some low income consumers might be exploited by the monopolist because of higher prices. And, some of their purchasing power might be transferred via dividends to shareholders in the higher income brackets - thus making the overall distribution of income more unequal.

However some of the supernormal profits might be used to invest in research and development programmes that have the potential to bring dynamic efficiency gains to consumers in the markets. There is a continuing debate about whether competitive or monopolistic markets provide the best environment for high levels of research spending.

Price Discrimination

Are there potential welfare improvements from price discrimination? Some forms of price discrimination benefit certain consumers.

Domestic monopoly but international competition

A firm may have substantial domestic monopoly power but face intensive competition from overseas producers. This limits their market power and helps keep prices down for
consumers. A good example to use here would be the domestic steel industry. **Corus** produces most of the steel manufactured inside the UK but faces intensive competition from overseas steel producers.

**Contestable markets**

Contestable market theory predicts that monopolists may still be competitive even if they enjoy a dominant position in their market. Their price and output decisions will be affected by the threat of "hit and run entry" from other firms if they allow their costs to rise and inefficiencies to develop.

### 2.7.3 Practices to control a monopoly on international platforms

From society's point of view, monopoly presents two problems. First, the outcome is Pareto inefficient (not enough is produced); and second, the distribution of welfare is biased in favor of the monopolist.

If there is only one firm in the industry because of some artificial restriction, then one possibility is to remove that restriction---if the outcome that results after the restriction is lifted is likely to be better than the monopoly outcome. But if the industry is a natural monopoly then lifting a restriction to entry will not in fact induce entry---in this case, an entrant cannot make a profit (the market is not big enough to support more than one firm).

Three ways in which the monopolist might be controlled are considered in the following sections.

**a) Price ceilings**

Suppose a price ceiling is imposed. How does this affect the monopolist's revenue curves?

Suppose the monopolist is not allowed to charge a price above \( p_0 \). Then if it sells less than is demanded at \( p_0 \) it must do so at the price \( p_0 \) (rather than at a higher price), and so its marginal revenue is \( p_0 \). If it sells more than is demanded at the price \( p_0 \) then the price is the same as it is in the absence of any restriction, and hence its marginal revenue is the same as it was originally. Thus its marginal revenue has a discontinuity, as in the following figure.
In the presence of the restriction, the firm's optimal output is $y_0$ (note that its profit at this output is positive), where the marginal revenue has a discontinuity: for smaller outputs MR exceeds MC, and for larger outputs MR is less than MC. The outcome of the regulation is thus that the price falls to $p_0$ (from its original value $p^*$) and output increases from $y^*$ to $y_0$.

If the regulated price ceiling is exactly the competitive price then *if the firm makes a profit*, it is induced to produce the efficient output. If it does not make a profit at this output, then a subsidy, in addition to a price ceiling, is necessary to induce the firm to produce the efficient output.

Notice that in order to set the price ceiling exactly at the level that induces the firm to produce the efficient output, the regulator has to know the firm's marginal cost curve. (And if the firm knows how the information about its MC is going to be used by the regulator, obviously it has an incentive not to reveal what it is.)

**b) Average cost pricing**

Suppose that the government requires the monopoly to set a price equal to average cost. That is, it requires the firm to choose an (output, price) pair for which AC is equal to AR. This regulation eliminates profit, but does not necessarily lead to an efficient outcome.

The outcome that this regulation produces is illustrated in the figure. In the left panel the output $y_0$ under the regulation is larger than the efficient output $y^*$, while in the right panel it is less than the efficient output. (In both cases the curves are drawn so that there is only one output at which AC and AR are equal; in other cases there may be more than one such output.)
C) Rate-of-return regulation

A common form of control in the US and Canada is rate-of-return regulation: the rate of return on invested capital is capped. This regulation results in a monopolist using more capital than it would if it were unregulated, given its output.

To see this, suppose that a firm uses capital (input 2) and another input (1). The return on capital is

$$TR(y) = w_1 z_1,$$

where $z_1$ is the amount of input 1. Rate-of-return regulation specifies the rate of return $r$ and requires the firm to choose an output $y$ and inputs $(z_1, z_2)$ such that

$$TR(y) = w_1 z_1 \leq rz_2,$$

where $z_2$ is the amount of input 2 (capital).

If $r = w_2$ then rate-of-return regulation is equivalent to average cost pricing: it requires the firm to make zero profit. Assume that $r > w_2$. Then under rate-of-return regulation the firm's profit-maximization problem is:

$$\max_{y, z_1, z_2} [TR(y) - w_1 z_1 - w_2 z_2] \text{ subject to } F(z_1, z_2) \geq y \text{ and } TR(y) \leq w_1 z_1 + rz_2.$$

Let $(y^*, z_1^*, z_2^*)$ be a solution of this problem. Consider the $y^*$-isoquant. The regulation requires that the firm use a pair of inputs satisfying the regulatory constraint $TR(y) \leq w_1 z_1 + rz_2$; that is, a pair $(z_1, z_2)$ above the line $TR(y) = w_1 z_1 + rz_2$. The slope of this line is $-w_1/r$ (write $z_2$ as a function of $z_1$), so since $r > w_2$, the line is not as steep as an isocost line. Note that as $r$ changes, the line pivots around the point it intersects the $z_1$ axis.
The problem of the firm is represented in the following figure. The firm has to choose an input bundle \((z_1, z_2)\) on or above the \(y^*\)-isoquant and on or above the regulatory constraint (i.e. in the area shaded light blue); it wants to choose the one that is on the lowest possible isocost line. The point in the figure that satisfies this condition is \((z_1^*, z_2^*)\). In the absence of regulation the firm would use the input bundle \((z_1', z_2')\).

![Figure 11](image)

Thus we see that the firm uses more of input 2 than if would if it were unregulated: the output is not produced at minimal cost. If the input prices reflect their social costs (as they do in a competitive equilibrium of the input markets) then the outcome is inefficient. Of course, this inefficiency has to be balanced against any change in the output of the monopolist in the direction of the efficient output, and also against any change in the distribution of income that the policy induces.

As \(r\) changes, the regulatory constraint pivots around the point it intersects the \(z_1\) axis. Thus when \(r\) is very large, the regulatory constraint has no effect on the firm. When \(r\) is not so large, as in the figure, the firm uses less of input 1 and more of input 2 than it does to the output \(y\) at minimal cost.

### 2.7.4 Efficient regulatory mechanisms

Is there any regulation that induces the efficient outcome? Obviously if the regulator knows the cost function and demand function, it can just tell the firm what to do. But it is not likely to have sufficiently detailed information about costs and demand.

Suppose there are many demanders, each of whom either buys 0 or 1 units of the good; buyer \(i\) has the reservation price \(R_i\), with \(R_i > R_{i+1} > \ldots > R_n > 0\). Then the following mechanism induces a monopolist to produce efficiently and requires the regulator to know only the demand function.

**Mechanism:** If the monopolist charges the price \(p\) and sells \(m\) units it gets the subsidy
$R_1 + R_2 + \ldots + R_m = mp$.  
How should a monopolist behave when confronted with this scheme? Its revenue from selling $m$ units at the price $p$ is 

$$mp + (R_1 + R_2 + \ldots + R_m - mp) = R_1 + R_2 + \ldots + R_m.$$  
That is, its revenue is exactly the revenue that a perfectly discriminating monopolist obtains when it maximizes its profit. Since the output chosen by a profit-maximizing perfectly discriminating monopolist is Pareto efficient, this subsidy scheme induces the monopolist to produce the Pareto efficient output.

The calculation that leads the monopolist to do so is the following. Suppose it is now selling $m$ units and considers selling another unit. Its additional revenue is $p + (R_m + 1 - p) = R_m + 1$, which is exactly the AR at $m$ units. Thus the monopolist chooses to sell exactly the number for which AR is equal to MC, which is the efficient amount.

The point is that when the monopolist decides to sell another unit, the price on the units that have "already" been sold is not lowered---the monopolist gets the full value of $R_m + 1$.

Obviously the scheme has the disadvantage that the monopolist gets the entire surplus: the distribution of income induced by the scheme is very inequitable, unless the monopolist is particularly deserving. But as we know, a proportional tax on profit does not affect the monopolist's behavior; it would allow the surplus to be distributed to other members of the economy.

### 2.8 THE MONOPOLISTIC AND RESTRICTIVE TRADE PRACTICES IN INDIA

The Monopolistic and Restrictive Trade Practices Act, 1969, was enacted

- To ensure that the operation of the economic system does not result in the concentration of economic power in hands of few,
- To provide for the control of monopolies, and
- To prohibit monopolistic and restrictive trade practices.

The MRTP Act extends to the whole of India except Jammu and Kashmir.

Unless the Central Government otherwise directs, this act shall not apply to:

a. Any undertaking owned or controlled by the Government Company,

b. Any undertaking owned or controlled by the Government,

c. Any undertaking owned or controlled by a corporation (not being a company established by or under any Central, Provincial or State Act,

d. Any trade union or other association of workmen or employees formed for their own reasonable protection as such workmen or employees,
e. Any undertaking engaged in an industry, the management of which has been taken over by any person or body of persons under powers by the Central Government,

f. Any undertaking owned by a co-operative society formed and registered under any Central, Provincial or state Act,

g. Any financial institution.

2.8.1 MONOPOLISTIC TRADE PRACTICES

A monopolistic trade practice is one, which has or is likely to have the effect of:

i. maintaining the prices of goods or charges for the services at an unreasonable level by limiting, reducing or otherwise controlling the production, supply or distribution of goods or services;

ii. unreasonably preventing or lessening competition in the production, supply or distribution of any goods or services whether or not by adopting unfair method or fair or deceptive practices;

iii. limiting technical development or capital investment to the common detriment;

iv. deteriorating the quality of any goods produced, supplied or distribute; and

v. increasing unreasonably -
   a. the cost of production of any good; or
   b. charges for the provision, or maintenance, of any services; or
   c. the prices for sale or resale of goods; or
   d. the profits derived from the production, supply or distribution of any goods or services.

A monopolistic trade practice is deemed to be prejudicial to the public interest, unless it is expressly authorized under any law or the Central Government permits to carry on any such practice.

INQUIRY INTO MONOPOLISTIC TRADE PRACTICES

The Commission may inquire into

Any monopolistic trade practice,

- Upon a reference made to it by the Central Government or
- Upon an application made to it by the Director General or
- Upon it own knowledge or information
2.8.2 UNFAIR TRADE PRACTICE

An unfair trade practice means a trade practice, which, for the purpose of promoting any sale, use or supply of any goods or services, adopts unfair method, or unfair or deceptive practice.

Unfair practices may be categorised as under:

1. FALSE REPRESENTATION

The practice of making any oral or written statement or representation which:

- Falsely suggests that the goods are of a particular **standard quality**, quantity, grade, composition, style or model;
- Falsely suggests that the services are of a particular **standard, quantity or grade**;
- Falsely suggests any re-built, second-hand renovated, reconditioned or **old goods as new goods**;
- Represents that the goods or services have sponsorship, approval, performance, characteristics, accessories, uses or benefits which they do not have;
- Represents that the seller or the supplier has a sponsorship or approval or affiliation which he does not have;
- Makes a false or misleading representation concerning the need for, or the usefulness of, any goods or services;
- Gives any warranty or **guarantee** of the performance, efficacy or length of life of the goods, that is **not based on an adequate or proper test**;
- Makes to the public a **representation** in the form that purports to be-
  - a warranty or guarantee of the goods or services,
  - a promise to replace, maintain or repair the goods until it has achieved a specified result,

  if such representation is **materially misleading** or there is no reasonable prospect that such warranty, guarantee or promise will be fulfilled
- Materially **misleads about the prices** at which such goods or services are available in the market; or
- Gives false or misleading facts disparaging the goods, services or trade of another person.
2. FALSE OFFER OF BARGAIN PRICE-

Where an advertisement is published in a newspaper or otherwise, whereby goods or services are offered at a bargain price when in fact there is no intention that the same may be offered at that price, for a reasonable period or reasonable quantity, it shall amount to an unfair trade practice.

The ‘bargain price’, for this purpose means-

a. the price stated in the advertisement in such manner as suggests that it is lesser than the ordinary price, or

b. the price which any person coming across the advertisement would believe to be better than the price at which such goods are ordinarily sold.

3. FREE GIFTS OFFER AND PRIZE SCHEMES

The unfair trade practices under this category are:

- Offering any gifts, prizes or other items along with the goods when the real intention is different, or
- Creating impression that something is being offered free along with the goods, when in fact the price is wholly or partly covered by the price of the article sold, or
- Offering some prizes to the buyers by the conduct of any contest, lottery or game of chance or skill, with real intention to promote sales or business.

4. NON-COMPLIANCE OF PRESCRIBED STANDARDS

Any sale or supply of goods, for use by consumers, knowing or having reason to believe that the goods do not comply with the standards prescribed by some competent authority, in relation to their performance, composition, contents, design, construction, finishing or packing, as are necessary to prevent or reduce the risk of injury to the person using such goods, shall amount to an unfair trade practice.

5. HOARDING, DESTRUCTION, ETC.

Any practice that permits the hoarding or destruction of goods, or refusal to sell the goods or provide any services, with an intention to raise the cost of those or other similar goods or services, shall be an unfair trade practice.

INQUIRY INTO UNFAIR TRADE PRACTICES

The Commission may inquire into

Any unfair trade practice

- Upon receiving a complaint from any trade association, consumer or a registered consumer association, or
• Upon reference made to it by the Central Government or State Government
• Upon an application to it by the Director General or
• Upon its own knowledge or information.

2.8.3 RESTRICTIVE TRADE PRACTICE
A restrictive trade practice is a trade practice, which
• Prevents, distorts or restricts competition in any manner; or
• **Obstructs** the flow of capital or resources into the stream of production; or
• Which tends to bring about **manipulation of prices or conditions** of delivery or affected the flow of supplies in the market of any goods or services, imposing on the consumers unjustified cost or restrictions.

INQUIRY INTO RESTRICTIVE PRACTICES
The Commission may inquire into any restrictive trade practice
• Upon receiving a complaint from any trade association, consumer or a registered consumer association, or
• Upon a reference made to it by the Central or State Government or
• Upon its own knowledge or information

2.8.4 POWERS OF THE COMMISSION
The MRTP Commission has the following powers:
1. Power of Civil Court under the Code of Civil Procedure, with respect to:
   a. Summoning and enforcing the attendance of any witness and examining him on oath;
   b. Discovery and production of any document or other material object producible as evidence;
   c. Reception of evidence on affidavits;
   d. Requisition of any public record from any court or office.
   e. Issuing any commission for examination of witness; and
   f. Appearance of parties and consequence of non-appearance.
2. Proceedings before the commission are deemed as judicial proceedings with in the meaning of sections 193 and 228 of the Indian Penal Code.
3. To require any person to produce before it and to examine and keep any books of accounts or other documents relating to the trade practice, in its custody.

4. To require any person to furnish such information as respects the trade practice as may be required or such other information as may be in his possession in relation to the trade carried on by any other person.

5. To authorise any of its officers to enter and search any undertaking or seize any books or papers, relating to an undertaking, in relation to which the inquiry is being made, if the commission suspects tat such books or papers are being or may be destroyed, mutilated, altered, falsified or secreted.

PRELIMINARY INVESTIGATION
Before making an inquiry, the Commission may order the Director General to make a preliminary investigation into the complaint, so as to satisfy itself that the complaint is genuine and deserves to be inquired into.

2.8.5 REMEDIES UNDER THE ACT
The remedies available under this act are –

TEMPORARY INJUNCTION
Where, during any inquiry, the commission is satisfied that any undertaking or any person is carrying on, or is about to carry on, any monopolistic, restrictive or unfair trade practice, which is a pre-judicial to the public interest or the interest of any trader or class of traders generally, or of any consumer or class of consumers, or consumers generally, the commission may grant a temporary injunction restraining such undertaking or person form carrying on such practice until the conclusion of inquiry or until further orders.

COMPENSATION
Where any monopolistic, restrictive or unfair trade practice has caused damage to any Government, or trader or consumer, an application may be made to the Commission asking for compensation, and the Commission may award appropriate compensation.

Where any such loss or damage is caused to a number of persons having the same interest, compensation can be claimed with the permission of the commission, by any of them on behalf of all of them.

Activity 2

1. Discuss the concept of monopoly equilibrium. What do you understand by marginal analysis in context of monopoly?

2. How production can be analyzed in monopoly firm? Discuss the approach to profit maximization.

3. What do you understand by term ‘price discrimination’?
4. Give a brief note on welfare aspects of monopoly in Indian and international perspectives?

2.9 SUMMARY

Monopoly is a market in which a single firm is the only supplier of the good. Anyone seeking to buy the good must buy from the monopoly seller. This single-seller status gives monopoly extensive market control. It is a price maker. The market demand for the good sold by a monopoly is the demand facing the monopoly. Market control means that monopoly does not equate price with marginal cost and thus does not efficiently allocate resources. Other related concepts to price and output determination have been discussed in detail. Further, welfare aspects of monopoly including control over monopoly practices on international level were discussed.

2.10 FURTHER READINGS

- Guy Ankerl, Beyond Monopoly Capitalism and Monopoly Socialism. Cambridge,
- The revolution in monopoly theory, by Glyn Davies and John Davies. Lloyds Bank Review, July 1984
UNIT 3

MONOPOLISTIC COMPETETION, OLIGOPOLY AND OTHER MARKET STRUCTURES

Objectives

After reading this unit, you should be able to:

- Understand the concepts of price differentiation, selling costs, excess capacity etc. with reference to monopolistic competition
- Explain the approaches to general and chamberlain equilibrium in monopolistic competition.
- Use the approach of kinked demand curve in decisions pertaining to oligopoly.
- Identify situations under monopsony.
- Find or assess the price, outputs and profits under Bilateral monopoly

Structure

3.1 Introduction
3.2 Monopolistic competition
3.3 General equilibrium theory
3.4 Chamberlain approach to monopolistic competition
3.5 Criticisms to monopolistic competition
3.6 Oligopoly
3.7 Kinked demand curve under oligopoly
3.8 Monopsony
3.9 Bilateral Monopoly
3.10 Summary
3.11 Further readings

3.1 INTRODUCTION

Monopolistic competition is a common market structure where many competing producers sell products that are differentiated from one another (i.e. the products are substitutes, but are not exactly alike). Similarly an oligopoly is a market form in which a market or industry is dominated by a small number of sellers ( oligopolists). There are various ways to determine price, output and even profits of the firms in these kind of market structures. Some recent concepts of bilateral monopoly and monopsony are also discussed in this chapter.
3.2 MONOPOLISTIC COMPETITION

The five key characteristics of monopolistic competition are: (1) large number of small firms, (2) similar but not identical products sold by the firms, (3) relative freedom of entry into and exit out of the industry, and (4) extensive knowledge of prices and technology (5) selling costs. These four characteristics mean that a given monopolistically competitive firm has a little bit of control over its small corner of the market. The large number of small firms, all producing nearly identical products, mean that a large (very, very large) number of close substitutes exists for the output produced by any given firm. This makes the demand curve for that firm's output relatively elastic.

Freedom of entry into and exit out of the industry means that capital and other resources are highly mobile and that any barriers to entry that might exist are minimal. Entry barriers allow real world firms to acquire and maintain above normal economic profit. Extensive knowledge means that all firms operate on the same footing, that buyers know a lot about possible substitutes for a given good, and that firms are aware of essentially the same production techniques.

LARGE NUMBER OF SMALL FIRMS

A monopolistically competitive industry contains a large number of small firms, each of which is relatively small compared to the overall size of the market. This ensures that all firms are relatively competitive with very little market control over price or quantity. In particular, each firm has hundreds or even thousands of potential competitors.

This number-of-firms characteristic is a sliding scale. The extent to which an industry has a large number of small firms, the more it is monopolistic competition. Industries with a smaller number of larger firms then tend to be more oligopolistic. There is no clear-cut dividing line that separates monopolistic competition from the more concentrated market structure--oligopoly. A three-firm industry is most assuredly an oligopoly. A 3,000 firm industry is most assuredly monopolistic competition. But, an industry with 30 firms could be oligopoly or monopolistic competition.

Consider this example of monopolistic competition from the Shady Valley restaurant market. Manny Mustard's House of Sandwich is one of 2,000 eateries scattered throughout Shady Valley. Each restaurant is in competition with every other. With 2,000 restaurants in total, none has a great deal of market control. If the going price of lunch is about $5, then each restaurant charges "about" $5 for lunch. Should one try to charge $10 or $20, then it will literally price itself out of the market as the quantity demanded drops to zero and customers switch to competitors.
PRODUCT DIFFERENTIATION

Each firm in a monopolistically competitive market sells a similar product. Yet each product is slightly different from the others. The term used to describe this is product differentiation. Product differentiation is responsible for giving each monopolistically competitive a little bit of a monopoly, and hence a negatively-sloped demand curve. Differences among products generally fall into one of three categories: (1) physical difference, (2) perceived difference, and (3) difference in support services.

- **Physical Difference:** This means that the product of one firm is physically different from the product of other firms. The most popular product of Manny Mustard's House of Sandwich is, for example, the Deluxe Club Sandwich. While many restaurants sell club sandwiches, Manny makes his with barbecue sauce rather than mayonnaise. It is similar to other club sandwiches, but slightly different.

- **Perceived Difference:** Product differentiation can also result from differences perceived by buyers, even though no actual physical differences exist. For example, OmniGuzzle gasoline is chemically identical to Bargain Discount Fuel gasoline. However, many buyers are absolutely convinced that OmniGuzzle is a "higher quality" gasoline. This could be due to years of intense OmniGuzzle advertising that has burned the OmniGuzzle brand name into heads of the consuming public. Brand names, in fact, are a common method of creating the perception of differences among products when none physically exist. However, perceived differences work just as well for monopolistic competition as actual differences. In the minds of the buyers, it matters not whether the differences are real or perceived.

- **Support Service Difference:** Products that are physically identical and perceived to be identical, can also be differentiated by support services. This is quite common in retail trade. For example, several independent stores might sell Master Foot brand athletic shoes. Buyers know that Master Foot shoes are the same regardless of who does the selling. No physical nor perceived differences exist. However, Bobby's Bunyon-Free Footware provides individual service, money-back guarantees, extended warranties, and service with a smile. Bobby's Bunyon-Free Footware sells buyers the perfect Master Foot brand athletic shoe that fits an individual's lifestyles. Mega-Mart Discount Warehouse Super Center, in contrast, has self-service shelves filled with Master Foot brand athletic shoes. Buyers must find their own sizes. Bobby's Bunyon-Free Footware is thus able to differentiate its Master Foot brand athletic shoe from those sold by Mega-Mart Discount Warehouse Super Center.

This characteristic means that every monopolistically competitive firm produces a good that is a close, but not a perfect substitute for the good produced by every other firm in the market. As such, different firms can charge slightly different prices. Manny Mustard can charge a slightly higher price for his club sandwich. OmniGuzzle can charge a slightly higher price for its gasoline. And Bobby's Bunyon-Free Footware can charge a slightly higher price for its Master Foot brand of athletic shoes.
RESOURCES MOBILITY

Monopolistically competitive firms, like perfectly competitive firms, are free to enter and exit an industry. The resources might not be as "perfectly" mobile as in perfect competition, but they are relatively unrestricted by government rules and regulations, start-up cost, or other substantial barriers to entry. While some firms incur high start-up cost or need government permits to enter an industry, this is not the case for monopolistically competitive firms. Likewise, a monopolistically competitive firm is not prevented from leaving an industry as is the case for government-regulated public utilities.

Most important, monopolistically competitive firms can acquire whatever labor, capital, and other resources that they need with relative ease. There is no racial, ethnic, or sexual discrimination.

For example, if Manny Mustard wants to leave the restaurant industry and enter the retail shoe sales industry, he can do that without restriction. Likewise if the Bobby (of Bobby's Bunyon-Free Footware) wants to leave the retail shoe industry and enter the restaurant industry, he can do so without restraint. Manny Mustard and Bobby are not faced with heavy up-front investment costs, such as the construction of a multi-million dollar factory, that would prevent them from entering a monopolistically competitive industry and competing on nearly equal ground with existing firms.

Consider the story of Manny Mustard, the second of seven children born to an immigrant butcher and his lovely wife, Ivana. Assuming the butchering duties upon his father's passing, Manny aspired to be the best butcher ever to slice sirloin in Shady Valley. The invasion of OmniFoods Discount Food Stores and several similar low-priced national food chains forced Manny to reevaluate his profession. With economic profit dropping into the negative range and normal profit on the verge of vanishing entirely, Manny transformed the Shady Valley Old Country Butcher's Shop into Manny Mustard's House of Sandwich. The slabs of raw meat were replaced by Deluxe Club Sandwich (with barbecue sauce). Manny's entry into the restaurant industry was relatively easy, and relatively straightforward.

EXTENSIVE KNOWLEDGE

In monopolistic competition, buyers do not know everything, but they have relatively complete information about alternative prices. They also have relatively complete information about product differences, brand names, etc. Moreover, each seller also has relatively complete information about the prices charged by other sellers so that they do not inadvertently charge less than the going market price.

Manny Mustard, for example, knows that the going price of club sandwiches in Shady Valley is about $5.50, give or take a little. All of the sandwich buyers know that the going Shady Valley price of club sandwiches is about $5.50, give or take a little.
Extensive knowledge also applies to technology and production techniques. Every monopolistically competitive firm has access to essentially the same production technology. Some firms might have a few special production tricks (a way to slice lettuce, a secret recipe, etc.), but these differences are not major. None of the firms has a patent on a perpetual motion machine or has uncovered the secret to time travel.

Manny Mustard, for example, has all of the information needed to make his club sandwiches. This is essentially the same information available to every sandwich-slicing supplier in the industry. Manny Mustard knows that club sandwich production involves a few slices of meat, lettuce, tomato layered between slices of bread, lubricated with sauce, and toasted to a golden brown.

**SELLING (ADVERTISING) COST**

Selling Cost (SC) is another outstanding feature of a monopolistic competitive market. This in the form of advertisement expenditure. Selling Cost and Product Differentiation together enable the producer to maintain some control over market conditions and influence the shape of the demand curve. Both features are interdependent. Whenever a product is differentiated it is necessary to inform buyers; and advertisement is the only medium through which buyers can be told about superiority of that product. Selling Cost by itself is apparent product differentiation. When a product does not contain any genuine qualitative difference, buyers can be made to treat a product differently through advertisements. So whenever products are differentiated and advertised, the market becomes a monopolistic competition. These are the hallmarks of this form of market. The presence of selling cost increases the firm’s cost of production. In order to recover it, firms have to charge a higher price. The net effect of a monopolistic competitive market is pricing goods at a higher rate. Consumers have to bear this extra expenditure.

**3.3 GENERAL EQUILIBRIUM THEORY**

General equilibrium theory is a key branch of theoretical neoclassical economics. It seeks to explain the behavior of supply, demand and prices in a whole economy with several or many markets, by assuming that equilibrium prices for goods exist and that all prices are at equilibrium, hence general equilibrium, in contrast to partial equilibrium. As with all models, this is an abstraction from a real economy, but is proposed as being a useful model, both by considering equilibrium prices as long-term prices, and by considering actual prices as deviations from equilibrium.

General equilibrium theory both studies economies using the model of equilibrium pricing, and seeks to determine in which circumstances the assumptions of general equilibrium will hold. The theory dates to the 1870s, particularly the work of French economist Léon Walras.

It is often assumed that agents are price takers and in that setting two common notions of equilibrium exist: Walrasian (or competitive) equilibrium, and its generalization; a price equilibrium with transfers.
Broadly speaking, general equilibrium tries to give an understanding of the whole economy using a "bottom-up" approach, starting with individual markets and agents. Macroeconomics, as developed by the Keynesian economists, focused on a "top-down" approach, where the analysis starts with larger aggregates, the "big picture". Therefore general equilibrium theory has traditionally been classed as part of microeconomics.

The difference is not as clear as it used to be, however, since much of modern macroeconomics has emphasized microeconomic foundations, and has constructed general equilibrium models of macroeconomic fluctuations. But general equilibrium macroeconomic models usually have a simplified structure that only incorporates a few markets, like a "goods market" and a "financial market". In contrast, general equilibrium models in the microeconomic tradition typically involve a multitude of different goods markets. They are usually complex and require computers to help with numerical solutions.

In a market system, the prices and production of all goods, including the price of money and interest, are interrelated. A change in the price of one good -- say, bread -- may affect another price, such as bakers’ wages. If bakers differ in tastes from others, the demand for bread might be affected by a change in bakers’ wages, with a consequent effect on the price of bread. Calculating the equilibrium price of just one good, in theory, requires an analysis that accounts for all of the millions of different goods that are available.

**PROPERTIES AND CHARACTERIZATION OF GENERAL EQUILIBRIUM**

Basic questions in general equilibrium analysis are concerned with the conditions under which an equilibrium will be efficient, which efficient equilibria can be achieved, when an equilibrium is guaranteed to exist and when the equilibrium will be unique and stable.

**First Fundamental Theorem of Welfare Economics**

The first fundamental welfare theorem asserts that market equilibria are Pareto efficient. In a pure exchange economy, a sufficient condition for the first welfare theorem to hold is that preferences be locally nonsatiated. The first welfare theorem also holds for economies with production regardless of the properties of the production function. Implicitly, the theorem assumes complete markets and perfect information. In an economy with externalities, for example, it is possible for equilibria to arise that are not efficient.

The first welfare theorem is informative in the sense that it points to the sources of inefficiency in markets. Under the assumptions above, any market equilibrium is tautologically efficient. Therefore, when equilibria arise that are not efficient, the market system itself is not to blame, but rather some sort of market failure.
Second Fundamental Theorem of Welfare Economics

While every equilibrium is efficient, it is clearly not true that every efficient allocation of resources will be an equilibrium. However, the Second Theorem states that every efficient allocation can be supported by some set of prices. In other words all that is required to reach a particular outcome is a redistribution of initial endowments of the agents after which the market can be left alone to do its work. This suggests that the issues of efficiency and equity can be separated and need not involve a trade off. However, the conditions for the Second Theorem are stronger than those for the First, as now we need consumers’ preferences to be convex (convexity roughly corresponds to the idea of diminishing rates of marginal substitution, or to preferences where "averages are better than extrema").

Existence

Even though every equilibrium is efficient, neither of the above two theorems say anything about the equilibrium existing in the first place. To guarantee that an equilibrium exists we once again need consumer preferences to be convex (although with enough consumers this assumption can be relaxed both for existence and the Second Welfare Theorem). Similarly, but less plausibly, feasible production sets must be convex, excluding the possibility of economies of scale.

Proofs of the existence of equilibrium generally rely on fixed point theorems such as Brouwer fixed point theorem or its generalization, the Kakutani fixed point theorem. In fact, one can quickly pass from a general theorem on the existence of equilibrium to Brouwer’s fixed point theorem. For this reason many mathematical economists consider proving existence a deeper result than proving the two Fundamental Theorems.

Uniqueness

Although generally (assuming convexity) an equilibrium will exist and will be efficient the conditions under which it will be unique are much stronger. While the issues are fairly technical the basic intuition is that the presence of wealth effects (which is the feature that most clearly delineates general equilibrium analysis from partial equilibrium) generates the possibility of multiple equilibria. When a price of a particular good changes there are two effects. First, the relative attractiveness of various commodities changes, and second, the wealth distribution of individual agents is altered. These two effects can offset or reinforce each other in ways that make it possible for more than one set of prices to constitute an equilibrium.

A result known as the Sonnenschein-Mantel-Debreu Theorem states that the aggregate (excess) demand function inherits only certain properties of individual's demand functions, and that these (Continuity, Homogeneity of degree zero, Walras' law, and boundary behavior when prices are near zero) are not sufficient to restrict the admissible aggregate excess demand function in a way which would ensure uniqueness of equilibrium.
There has been much research on conditions when the equilibrium will be unique, or which at least will limit the number of equilibria. One result states that under mild assumptions the number of equilibria will be finite (see Regular economy) and odd (see Index Theorem). Furthermore if an economy as a whole, as characterized by an aggregate excess demand function, has the revealed preference property (which is a much stronger condition than revealed preferences for a single individual) or the gross substitute property then likewise the equilibrium will be unique. All methods of establishing uniqueness can be thought of as establishing that each equilibrium has the same positive local index, in which case by the index theorem there can be but one such equilibrium.

**Determinacy**

Given that equilibria may not be unique, it is of some interest to ask whether any particular equilibrium is at least locally unique. If so, then comparative statics can be applied as long as the shocks to the system are not too large. As stated above, in a Regular economy equilibria will be finite, hence locally unique. One reassuring result, due to Debreu, is that "most" economies are regular. However recent work by Michael Mandler (1999) has challenged this claim. The Arrow-Debreu-McKenzie model is neutral between models of production functions as continuously differentiable and as formed from (linear combinations of) fixed coefficient processes. Mandler accepts that, under either model of production, the initial endowments will not be consistent with a continuum of equilibria, except for a set of Lebesgue measure zero. However, endowments change with time in the model and this evolution of endowments is determined by the decisions of agents (e.g., firms) in the model. Agents in the model have an interest in equilibria being indeterminate:

"Indeterminacy, moreover, is not just a technical nuisance; it undermines the price-taking assumption of competitive models. Since arbitrary small manipulations of factor supplies can dramatically increase a factor's price, factor owners will not take prices to be parametric." (Mandler 1999, p. 17)

When technology is modeled by (linear combinations) of fixed coefficient processes, optimizing agents will drive endowments to be such that a continuum of equilibria exist:

"The endowments where indeterminacy occurs systematically arise through time and therefore cannot be dismissed; the Arrow-Debreu-McKenzie model is thus fully subject to the dilemmas of factor price theory." (Mandler 1999, p. 19)

Critics of the general equilibrium approach have questioned its practical applicability based on the possibility of non-uniqueness of equilibria. Supporters have pointed out that this aspect is in fact a reflection of the complexity of the real world and hence an attractive realistic feature of the model.

**Stability**

In a typical general equilibrium model the prices that prevail "when the dust settles" are simply those that coordinate the demands of various consumers for various goods. But
this raises the question of how these prices and allocations have been arrived at and whether any (temporary) shock to the economy will cause it to converge back to the same outcome that prevailed before the shock. This is the question of stability of the equilibrium, and it can be readily seen that it is related to the question of uniqueness. If there are multiple equilibria, then some of them will be unstable. Then, if an equilibrium is unstable and there is a shock, the economy will wind up at a different set of allocations and prices once the convergence process terminates. However stability depends not only on the number of equilibria but also on the type of the process that guides price changes (for a specific type of price adjustment process see Tatonnement). Consequently some researchers have focused on plausible adjustment processes that guarantee system stability, i.e., that guarantee convergence of prices and allocations to some equilibrium. However, when more than one stable equilibrium exists, where one ends up will depend on where one begins.

3.4 CHAMBERLIN APPROACH TO MONOPOLISTIC COMPETITION

Edward Hastings Chamberlin (b. 1899) in 1933 published The Theory of Monopolistic Competition as a reorientation of the theory of value, designed to base it on a synthesis of monopolistic and competitive theories. He argues that the old idea of monopoly and competition as alternative is wrong; and that most situations are composites in which elements of both monopoly and competition are combined. But he asserts that the correct procedure is to start from the theory of monopoly. This, he thinks, has the merit of eliminating none of the competitive elements, since these operate through the demand for the monopolist's product; while on the contrary the alternative assumption of competition rules out the monopoly elements.

Thus, in taking monopoly as a starting point, Chamberlin's approach is similar to that of Cournot.

But, while with Cournot the transition to perfect competition takes place only on a scale of numbers of competitors, with Chamberlin it takes place also on a scale of substitution of products. Any producer whose product is significantly different from the products of others has some monopoly of it, subject to the competition of substitutes. He considers each producer in an industry as having some monopoly in his own product. If he be the sole seller of a unique product, he has a pure monopoly.1 If there be two sellers of similar products, the situation is one of "duopoly." If there be several, an "oligopoly" exists. The condition may range through various degrees of oligopoly to pure competition, under which there are so many sellers of a highly standardized product that any one could sell all his product without affecting the demand. Pure competition is found only under the dual condition of (a) a large number, and (b) a perfectly standardized product. The usual condition Chamberlin considers to be in the intermediate area, in which some element of "monopoly" exists, and which he calls "monopolistic competition."

Economic inertia and friction are "imperfections" which he does not consider as part of "monopolistic competition."
Thus Chamberlin's thought centers on the product. Each producer, under "monopolistic competition," faces competition from "substitute" products which are not identical and which are sold by other concerns with various price policies, and sales expenses. These merely limit his "monopoly" of his own product.

The individual demand curve (or sales) for one seller's product is then regarded as affected by the market policies of other individual sellers whose products are partial substitutes. Total sales of the partly competing group of substitute products are treated as limiting the sales of the product of any one seller. Under "pure" competition (many sellers and a completely standardized product) a horizontal demand curve (average revenue) would exist for each individual competitor's product. This would mean identical prices. Chamberlin argues that "pure" competition would force all individual competitors to treat differential advantages, or rents, as costs, the same as other costs.

Chamberlin emphasizes the effect of judgments by one seller concerning his rivals' policies, possible retaliation, etc. He also argues that selling costs such as advertising are not part of the cost of production, but are incurred to increase the sales of the given product; and thus they affect the demand curve. Throughout, his basic idea is that, no matter how slight, any differentiation of a seller's product gives him to that extent a monopoly. And all these conditions, commonly found in competitive markets, are either "impurities" in the nature of monopoly elements, or are associated with such elements. They make "pure" competition impossible.

To Chamberlin, actual "competition"1 includes the effort of competitors to increase their monopoly powers.

![figure 1](image-url)
DD' = demand curve (avg. rev.)
PP' = avg. cost of production, including a "minimum profit" (charge required to attract capital and enterprise) and all "rents"
FF' = avg. total cost, including fixed uniform selling costs
AR = price
EHRR' = profit (above "minimum")
OA = quantity sold
pp' = marginal cost of production
dd' = marginal revenue
Q = intersection of pp' and dd'

And the essence of "monopoly," and therefore of "monopolistic competition," is seen as lying in differences — (1) differences in price policy, (2) differences in nature of product, and (3) differences in such sales effort as advertising outlays. It is a contribution of Chamberlin's to have developed the second and third of these variables as arising out of the mixture of monopoly and competition.

Chamberlin starts with a single firm and develops the idea of monopoly price and competitive prices as determined by the intersection of revenue or sales curves with expense curves. Either the marginal revenue curve, or the average revenue curve (from which it is derived), may be used to determine the monopoly output and price, the former by intersecting the rising marginal cost curve, the latter by the familiar Marshallian method of fitting the maximum profit area between it and the average cost curve, which includes rents or differentials and thus equals the average price.

The analysis with respect to all three variables then is extended beyond the firm to groups of sellers, which may be taken as corresponding to conventional "industries," depending on how broadly a "class of product" is conceived in a particular case. The group is analyzed, first under the assumption of symmetry (all its members assumed to have uniform cost and demand curves). Then some consideration is given to what might happen if a "diversity of conditions" existed. If selling costs are not great, and if they reduce the slope of the sellers' demand curves, increasing them may result in a lower price. Variations in product may lead to either smaller or larger outputs. Group equilibrium (with "alert" competitors) must result in the optimum with respect to all the variables, and no profits above a necessary minimum for every producer.

The conclusion is drawn that under monopolistic competition the equilibrium price is higher, and the volume of output probably (not necessarily) lower, than under pure competition. The net profits of enterprise, however, may or may not be higher than under pure competition because of the expense which is required to maintain the monopoly elements and which is often increased by a multiplication of substitute products surrounding the monopolist. Chamberlin argues that monopolistic competition need not bring higher profits to the marginal firm in a given industry. Instead it may allow the existence of a larger number of firms making normal profits.
MONOPOLISTIC COMPETITION, LONG-RUN EQUILIBRIUM CONDITIONS

The long-run equilibrium of monopolistically competitive industry generates six specific equilibrium conditions: (1) economic inefficiency (P > MC), (2) profit maximization (MR = MC), (3) market control (P = AR > MR), (4) breakeven output (P = AR = ATC), (5) excess capacity (ATC > MC), and (6) economies of scale (LRAC > LRMC).

A monopolistically competitive industry achieves long-run equilibrium through the adjustment of the market price, the number of firms in the industry, and the scale of production of each firm. These adjustments mean that each firm produces at a point of tangency between its negatively-sloped average revenue (demand) curve and its long-run average cost curve. This occurs in the economies of scale portion of the long-run average cost curve. With this production two equilibrium conditions are achieved:

\[\text{MR = MC = LRMC}\]
\[\text{P = AR = ATC = LRAC}\]

The first condition, marginal revenue (MR) equal to marginal cost (MC and LRMC), means that each firm maximizes profit and has no reason to adjust its quantity of output or plant size. The second condition, price (P) equal to average cost (ATC and LRAC), means that each firm in the industry is earning only a normal profit. Economic profit is zero and there is no economic loss.

These two equilibrium conditions can be divided into the six specific conditions: (1) economic inefficiency (P > MC), (2) profit maximization (MR = MC), (3) market control (P = AR > MR), (4) breakeven output (P = AR = ATC), (5) excess capacity (ATC > MC), and (6) economies of scale (LRAC > LRMC).

For a closer look at these six conditions, consider the hypothetical monopolistically competitive Shady Valley restaurant industry. The hypothetical Shady Valley restaurant industry contains a large number of relatively small firms (thousands of meal makers, each producing a handful of meals each day), similar but not identical products (each produces food, but the meals differ from restaurant to restaurant), relative ease of entry and exit (anyone can set up a restaurant with little or no upfront cost and few legal restrictions), and extensive knowledge of prices and technology (ever restaurant knows how to prepare meals and they are aware of relevant prices).

ECONOMIC INEFFECTIVENESS

P > MC

The condition that price is greater marginal cost (P = MC) means that production does NOT achieve economic efficiency. This means that resources are not being used to produce goods that generate the greatest possible level of satisfaction.

If the price that a hypothetical restauranteer named Manny (as well as every other monopolistically competitive restauranteer) receives for his meals is greater than the marginal cost of producing a luncheon meal such as sandwiches, then it is possible to produce more sandwiches and improve society's overall satisfaction.
Suppose, for example, that in long-run equilibrium the sandwich price is $4.95 but the marginal cost of sandwich production is $4.65.

- From the buyers viewpoint, this $4.95 price means that they receive $4.95 worth of satisfaction from consuming a sandwich. If buyers did not enjoy $4.95 worth of satisfaction, then they are not willing to pay $4.95. As such, the good produced by the monopolistically competitive restauranteers (that would be sandwiches) generates $4.95 of satisfaction.
- From the sellers viewpoint, the marginal cost is the opportunity cost of producing sandwiches. This is the value of other goods NOT produced when resources are used to produce sandwiches. If the marginal cost of producing a sandwich is $4.65, then the resources that Manny uses to produce sandwiches could have been used to produce another good, such as macrame plant holders or paperback books. And the value of the other good NOT produced is $4.65. In other words, macrame plant holder buyers are willing to pay $4.65 for the macrame plant holders that could have been produced with the resources used to produce the sandwich.

Inefficiency exists because the value of the good produced is greater than the value of the good NOT produced. As such, it is possible to increase total satisfaction by producing more sandwiches and fewer macrame plant holders.

The only way to achieve economic efficiency is to satisfy the condition \( P = MC \). This condition is NOT satisfied by a monopolistically competitive industry in the long run.

**PROFIT MAXIMIZATION**

\[ \text{MR} = \text{MC} \]

The condition that marginal revenue equals marginal cost \( (\text{MR} = \text{MC}) \) is the standard condition for profit-maximization by a firm. This condition means that, given existing price and cost conditions, a firm is producing the quantity of output that generates the highest possible level of economic profit.

Manny the hypothetical restauranteer maximizes his economic profit by producing the quantity of meals that equates the marginal revenue received for selling meals ($4.65) with the marginal cost of producing meals (also $4.65). It is not possible for Manny to generate any greater economic profit by producing more or fewer meals.

Marginal revenue is the extra revenue that Manny receives for producing meals. Marginal cost is the extra cost Manny incurs when producing meals. When the production of a sandwich changes revenue by exactly the same as it changes cost, economic profit does not change, profit has reached its maximum.

To see why, consider how profit is affected if marginal revenue and marginal cost are NOT equal.

- Suppose, for example, that marginal revenue (at $4.65) is greater than marginal cost (only $3.65). When Manny receives $4.65 for producing an extra sandwich
that incurs a cost of only $3.65, then his economic profit rises by $1. Any time Manny can sell sandwiches for more than the cost, profit goes up. But if profit can be increased by producing and selling more sandwiches, it must not be maximized.

- Alternatively, suppose that marginal revenue (at $4.65) is less than marginal cost ($5.65). When Manny receives $4.65 for producing an extra sandwich that incurs a cost of $5.65, then his economic profit falls by $1. Any time Manny sells sandwiches for less than the cost, profit goes down. But if profit goes down by producing MORE, it goes up by producing LESS. And if profit can be changed by changing production, it must not be maximized.

Only by satisfying the condition \(\text{MR} = \text{MC}\) is profit maximized. This condition is satisfied by a monopolistically competitive firm in the long run.

**MONOPOLISTIC COMPETITION**

\[ P = \text{AR} > \text{MR} \]

The condition that price equals average revenue but is greater than marginal revenue \(\left(P = \text{AR} > \text{MR}\right)\) is the standard condition for a monopolistically competitive firm. This condition means that a firm is a price maker with some degree of market control and faces a negatively-sloped demand curve.

The key to this condition is that a monopolistically competitive firm has some market control. The firm does not merely accept the going market price. It can increase the quantity sold by lowering the price.

Manny, for example, can sell his sandwiches for $4.95 each. If he wants to sell more sandwiches, then he must lower the price. If he raises his price, then he sells fewer sandwiches.

This condition also means that the price Manny charges is his average revenue. In fact, average revenue and price are really just two terms for the same thing. Average revenue is the revenue Manny receives per sandwich. Price is the revenue Manny receives per sandwich. Price is almost always equal to average revenue for any firm regardless of market structure.

Monopolistic competition, however, is indicated because marginal revenue is less than price and average revenue. Because Manny is a monopolistically competitive restauranteur who has some degree of market control and faces a negatively-sloped demand curve, each EXTRA sandwich he sells generates less EXTRA revenue than the price.

**BREAKEVEN OUTPUT**

\[ P = \text{ATC} = \text{LRAC} \]

The condition that price equals both short-run average total cost and long-run average cost \(\left(P = \text{ATC} = \text{LRAC}\right)\) indicates that a firm is producing breakeven output, earning
exactly a normal profit. Manny, the monopolistically competitive firm, is not receiving an economic profit nor incurring an economic loss.

This condition further means that firms have no incentive to enter or exit the industry. If no firms IN the monopolistically competitive industry receive above-normal economic profit, then there is no incentive for other firms to enter the industry. If no firms IN the monopolistically competitive industry incur economic loss or receive below-normal profit, then there is no incentive for any firms to exit the industry.

Consider the $4.95 price that Manny receives for his sandwiches. Because this price is equal to the short-run average total cost and the long-run average cost of producing meals, Manny earns exactly a normal profit. Note that normal profit is included as a cost of production. Because Manny is earning a normal profit, he has no incentive to switch from sandwich production to an alternative industry, such as macrame plant holder production.

Normal profit is the profit that Manny could be earning in another activity, such as macrame plant holder production. Because this is equal to the profit he is earning in sandwich production, there is no reason to change. There is no reason to leave the sandwich industry in search of greater profit on the other side of the shopping mall. Felicity the macrame plant holder maker reaches the same conclusion. Her macrame plant holder profit is the same as Manny’s sandwich profit. She has no incentive to leave the macrame plant holder industry and enter the sandwich industry.

What would happen, however, if price is not equal to average cost.

- Suppose, for example, that the sandwich price ($4.95) is greater than the average total cost of producing sandwiches (say $3.95). In this case, Manny receives $1 of economic profit for each sandwich sold. Because this exceeds the zero economic profit earned by Felicity the macrame plant holder maker, Felicity is induced to leave macrame plant holder production and take up sandwich production.
- Alternatively, if the sandwich price ($4.95) is less than the average total cost of producing sandwiches (say $5.95), then Manny incurs an economic loss of $1 for each sandwich sold. Because he much prefers NOT to incur an economic loss, he is attracted to the zero economic profit (that is, normal profit) earned by Felicity the macrame plant holder maker. Manny is induced to leave sandwich production and take up macrame plant holder production.

Only when the condition (P = ATC = LRAC) is satisfied do firms earn exactly a normal profit, receiving neither an economic profit nor incurring an economic loss. And only when this condition is satisfied is there no incentive for firms to enter or exit an industry. This condition is satisfied by a monopolistically competitive industry in the long run.

**EXCESS CAPACITY**

ATC > MC
The condition that short-run average total cost exceeds short-run marginal cost equals (ATC > MC) means that a firm is NOT operating at the minimum point of its short-run average total cost curve. In fact, this condition means that the firm is producing a smaller quantity than that achieved at this minimum point. Moreover, this means that a firm is NOT producing output at the lowest possible per unit cost and that the capital (or factory) is NOT being used in the most technically efficient manner possible.

Because the average cost of Manny's sandwich production is greater than marginal cost, Manny is operating to the left of the minimum point on his short-run average total cost curve. Manny can actually produce sandwiches at a lower per unit cost, given his existing capital (the current size of his restaurant and his array of kitchen tools), by increasing production. And in so doing, the average cost of production would decline.

Long-run equilibrium for a monopolistically competitive industry achieves the condition that ATC > MC. This means that each firm produces less output than could be achieved by fully using the available capacity of the plant size.

ECONOMIES OF SCALE

LRAC > LRMC

The condition that long-run average cost is greater than long-run marginal cost (LRAC > LRMC) means that a firm is operating to the left of the minimum point of its long-run average cost curve (the minimum efficient scale of production). This is the economies of scale range of production, characterized by a negatively-sloped long-run average cost curve. This condition means a firm has NOT constructed the most technically efficient factory. The firm is NOT producing output at the lowest possible long-run per unit cost. By increasing production, long-run average cost decreases.

When the long-run average cost exceeds long-run marginal cost, Manny's sandwich production is not at the minimum point on his long-run average cost curve. Manny can produce meals at a lower per unit cost in the long run by taking advantage of economies of scale, such as volume resource price discounts, input specialization, etc.

Long-run equilibrium for a monopolistically competitive industry achieves the condition (LRMC > LRAC), which means that firms are not producing output at the lowest possible per unit cost.

MONOPOLISTIC COMPETITION, LONG-RUN PRODUCTION ANALYSIS

In the long run, a monopolistically competitive firm adjusts plant size, or the quantity of capital, to maximize long-run profit. In addition, the entry and exit of firms into and out of a monopolistically competitive market eliminates economic profit and guarantees that each monopolistically competitive firm earns nothing more or less than a normal profit. Monopolistic competition is a market structure characterized by a large number of small firms producing similar but not identical products with relatively good resource mobility
and extensive knowledge. These conditions mean that each firm has some degree of market control and faces a negatively-sloped demand curve. As such, the entry and exit of firms into and out of the industry eliminates economic profit. Moreover, the pursuit of profit maximization by individual firms facing negatively-sloped demand curves results in economic inefficiency.

**Long-Run Adjustment**

The two adjustments undertaken by a monopolistically competitive industry in the pursuit of long-run equilibrium are:

- **Firm Adjustment:** Each firm in the monopolistically competitive industry adjusts short-run production and long-run plant size to achieve profit maximization. This adjustment entails producing the quantity that equates marginal revenue to short-run marginal cost for a given plant size as well as selecting the plant size that equates marginal revenue to long-run marginal cost.

- **Industry Adjustment:** Firms enter and exit a monopolistically competitive industry in response to economic profit and loss. If firms in the industry earn above-normal profit or receive economic profit, then other firms are induced to enter. If firms in the industry receive below-normal profit or incur economic loss, then existing firms are induced to exit. The entry and exit of firms causes the market price to change, which eliminates economic profit and loss, and leads to exactly normal profit.

**Long-Run Equilibrium Conditions**

The combination of firm and industry adjustment results in two equilibrium conditions. The profit-maximizing condition is that marginal revenue is equal to marginal cost (both short run and long run). The zero economic profit condition is that price (and average revenue) is equal to average cost (both short run and long run).

\[
\text{MR} = \text{MC} = \text{LRMC} \\
\text{P} = \text{AR} = \text{ATC} = \text{LRAC}
\]

With marginal revenue (MR) equal to marginal cost (MC and LRMC), each firm maximizes profit and has no reason to adjust its quantity of output or plant size. With price (P) equal to average cost (ATC and LRAC), each firm in the industry is earning only a normal profit. Economic profit is zero and there is no economic loss.
The six specific equilibrium conditions achieved by long-run equilibrium of monopolistically competitive industry are: (1) economic inefficiency (P > MC), (2) profit maximization (MR = MC), (3) market control (P = AR > MR), (4) breakeven output (P = AR = ATC), (5) excess capacity (ATC > MC), and (6) economies of scale (LRAC > LRMC).

These conditions are only satisfied by the tangency of the negatively-sloped demand (average revenue) curve facing a monopolistically competitive industry and the economies of scale portion of the long-run average cost curve. This means that a monopolistically competitive firm does not achieve long-run economic efficiency.

Key to these conditions is that they are NOT equal. Because price is not equal to marginal revenue in monopolistic competition average cost is not equal to marginal cost. The only production level in which average cost is equal to marginal cost (both short run and long run) is at the minimum efficient scale, the bottom of the long-run average cost curve. The only way to achieve this production level is the equality between price and marginal revenue. This equality is only achieved by perfect competition.

### 3.5 CRITICISM TO MONOPOLISTIC COMPETITION

While monopolistically competitive firms are inefficient, it is usually the case that the costs of regulating prices for every product that is sold in monopolistic competition by far exceed the benefits; the government would have to regulate all firms that sold heterogeneous products—an impossible proposition in a market economy. A monopolistically competitive firm might be said to be marginally inefficient because the firm produces at an output where average total cost is not a minimum. A monopolistically competitive market might be said to be a marginally inefficient market structure because marginal cost is less than price in the long run.

Another concern of critics of monopolistic competition is that it fosters advertising and the creation of brand names. Critics argue that advertising induces customers into spending more on products because of the name associated with them rather than because of rational factors. This is disputed by defenders of advertising who argue that (1) brand names can represent a guarantee of quality, and (2) advertising helps reduce the cost to consumers of weighing the tradeoffs of numerous competing brands. There are unique information and information processing costs associated with selecting a brand in a monopolistically competitive environment. In a monopoly industry, the consumer is faced with a single brand and so information gathering is relatively inexpensive. In a perfectly competitive industry, the consumer is faced with many brands. However, because the brands are virtually identical, again information gathering is relatively inexpensive. Faced with a monopolistically competitive industry, to select the best out of many brands the consumer must collect and process information on a large number of different brands. In many cases, the cost of gathering information necessary to selecting the best brand can exceed the benefit of consuming the best brand (versus a randomly selected brand).
Evidence suggests that consumers use information obtained from advertising not only to assess the single brand advertised, but also to infer the possible existence of brands that the consumer has, heretofore, not observed, as well as to infer consumer satisfaction with brands similar to the advertised brand.

3.6 OLIGOPOLY

Oligopolistic industries share several behavioral tendencies, including: (1) interdependence, (2) rigid prices, (3) nonprice competition, (4) mergers, and (5) collusion. In other words, each oligopolistic firm keeps a close eye on the decisions made by other firms in the industry (interdependence), are reluctant to change prices (rigid prices), but instead try to attract customers from the competition using incentives other than prices (nonprice competition), and when they get tired of competing with their competitors they are inclined to cooperate formally and legally (mergers) or informally and illegally (collusion).

Oligopolistic industries are nothing if not diverse. Some sell identical products, others differentiated products. Some have three or four firms of nearly equal size, others have one large dominate firm (a clear industry leader) and a handful of smaller firms (that follow the leader). Some sell intermediate goods to other producers others sell consumer goods directly to the public.

However, through this diversity, all oligopolistic industries engage in similar types of behavior. The most noted behavior tendencies are: (1) interdependent decision making, (2) relatively constant prices, (3) competition in ways that do not involve prices, (4) the legal merger of two or more firms, and (5) the illegal collusion among firms to control price and production.

INTERDEPENDENCE

Each firm in an oligopolistic industry keeps a close eye on the activities of other firms in the industry. Because oligopolistic firms engage in competition among the few, decisions made by one firm invariably affect others. Competition among interdependent oligopoly firms is comparable to a game or an athletic contest. One team's success depends not only on its own actions but the actions of its competitors. Chip Merthington might win a foot race not just because he runs really fast, but because his competition (Edgar Millbottom) runs really slow.

In a game of chess, Chip captures Edgar's knight with his rook. Edgar then counters by capturing Chip's rook with his queen. The key point is that Edgar would not have taken Chip's rook if Chip had not captured Edgar's knight. This is how oligopolies behave. An action by one firm motivates a counter action by another firm.

Consider, for example, the hypothetical oligopolistic athletic footwear industry, dominated by two companies OmniRun, Inc. and The Master Foot Company. If OmniRun introduces the OmniFast 9000, a new running shoe with ankle stabilizers and an extra thick cushioned insole, then The Master Foot Company needs to introduce a comparable shoe to keep pace with the competition because its existing model, the Fleet Foot 30, does not have ankle stabilizers nor an extra thick cushioned insole. If The
Master Foot Company does not counter the action by OmniRun, then buyers will likely choose the new OmniFast 9000 over the older Fleet Foot 30.

As such, The Master Foot Company will probably introduce something like the Fleet Foot 40 with flexible ankle stabilizers, a double extra thick cushioned insole, and metallic heal reflectors. And when it does, it is also likely to launch a massive advertising campaign to promote the new shoe, using the well-known, and wildly popular baseball superstar, Harold "Hair Doo" Dueterman as a spokesperson. This is likely to prompt OmniRun, Inc. to launch its own advertising blitz for the OmniFast 9000 featuring motion picture box office mega-star, Brace Brickhead.

And on it goes... each firm taking action to counter that of the other firm, which then takes further action, which then prompts more action.

**RIGID PRICES**

Oligopolistic industries tend to keep prices relatively constant, preferring to compete in ways that do not involve changing the price. The prime reason for rigid prices rests with the interdependence among oligopolistic firms.

- Because competing firms ARE NOT likely to match the price increases of an oligopolistic firm, the firm is likely to lose customers and market share to the competition should it charge a higher price. As such, it has little motive to increase its price.
- Because competing firms ARE likely to match the price decreases of an oligopolistic firm, the firm is unlikely to gain customers and market share from the competition should it charge a lower price. As such, it has little motive to decrease its price.

Consider, once again, the oligopolistic athlete shoe industry. OmniRun, Inc. sells its OmniFast 9000 shoe for $100. Likewise, The Master Foot Company sells its Fleet Foot 40 running shoe for $100.

OmniRun could reduce the price of its OmniFast 9000 to $95, thinking buyers will select it over the more expensive Fleet Foot 40. But Master Foot is not likely to sit idly by as OmniRun dominates the market by virtue of a lower price. Master Foot will reduce the price of the Fleet Foot 40 to $95 as well.

The net result of this joint price reduction is that each firm retains the same market share, but sells its shoe for $5 less. While the lower overall shoe price might increase the overall quantity demanded in the market (due to the law of demand), neither firm gains a competitive advantage over the other. Both maintain the same market share at the $95 price as they had with the $100 price.

To the extent that OmniRun realizes Master Foot will match any price reduction, it has little motivation to reduce prices.
Master Foot also has little motivation to pursue a price increase of its Fleet Foot 40 to $105. OmniRun is unlikely to match this higher price. If the higher Fleet Foot 40 price is $5 more, the OmniFast 9000 is $5 cheaper. Buyers will select the less expensive OmniFast 9000 over the now more expensive Fleet Foot 40. As such, The Master Foot Company has nothing to gain with a higher price, but it is likely to lose market share to OmniRun.

The net result is that neither firm can gain a competitive advantage by changing the price. As such, the seek to compete in ways that do not involve price changes.

However, this does not mean prices in oligopolistic industries NEVER change. Should industry-wide conditions change, such as higher input prices, regulatory changes, or technological advances, conditions that affect all firms, then all firms are likely respond in the same manner. They are likely to raise or lower prices together.

Should the Athletic Shoe Workers Union negotiate an across-the-board 10 percent wage increase, then OmniRun and Master Foot are both inclined to raise shoe prices to the same degree. Should Professor Magnaminious, the leading expert on athletic shoe fabrication, design a new athletic shoe assembly machine that is twice as productive as the old assembly method, then OmniRun and Master Foot are both inclined to reduce shoe prices to the same degree.

**NONPRICE COMPETITION**

Because oligopolistic firms realize that price competition is ineffective, they generally rely on nonprice methods of competition. Three of the more common methods of nonprice competition are: (1) advertising, (2) product differentiation, and (3) barriers to entry. The key for a firm is to attract buyers and increase market share, while holding the line on price.

- **Advertising:** A large share of commercial advertising, especially at the national level, is designed as nonprice competition among oligopolistic firms. The Master Foot Company, as a hypothetical example, promotes its Fleet Foot 40 running shoe using the baseball superstar, Harold "Hair Doo" Dueterman, as a spokesperson. OmniRun, Inc. counters with advertising for the OmniFast 9000 featuring motion picture mega-star, Brace Brickhead. Each firm engages in advertising is an attempt either: (a) to attract customers from its competition or (b) to prevent the competition from attracting its customers.

- **Production Differentiation:** Another common method of nonprice competition among oligopolistic firms is product differentiation. Such firms often compete by offering a bigger, better, faster, cleaner, and newer product—and especially one that is different from the competition. This is the reason why OmniRun might introduce its OmniFast 9000, with ankle stabilizers and an extra thick cushioned insole. This is also the reason why The Master Foot Company might introduce its Fleet Foot 40 with flexible ankle stabilizers, a double extra thick cushioned insole, and metallic heal reflectors. Each firm seeks to differentiate its product and to
give customers a reason (other than price differences) to select its product over the competition.

- Barriers to Entry: Oligopolistic firms also frequently "compete" by preventing the competition from entering the industry. Master Foot, for example, has a patent on the design of its innovative Fleet Foot 40 shoe which, for obvious reasons, it does not care to share with any potential competitors. Alternatively, OmniRun has acquired exclusive ownership of the world's supply of plaviminium (the material used to make the extra thick cushioned insole of the OmniFast 9000), which it is not inclined to sell to potential competitors. While assorted entry barriers exist, a popular form is government restrictions, especially if the competition happens to reside in another country.

**MERGERS**

Interdependence means that oligopolistic firms perpetually balance the need for competition against the benefits of cooperation. OmniRun and Master Foot are competitors in the market for athletic shoes. The profitability of OmniRun depends on the actions of Master Foot and vice versa. Such competition is inherent in an industry with a small number of large firms.

However, oligopolistic firms also realize that cooperation is often more beneficial than competition. One common method of cooperation is through a merger, that is, the legally combination of two firms into a single firm. OmniRun, for example, can eliminate its number one competitor, Master Foot, by merging with it and forming a new, larger company (MasterRun or perhaps OmniFoot). With such a merger, OmniRun now has one less competitor to worry about. If Master Foot is OmniRun's only competitor, then this merger gives OmniRun a **monopoly** in the athletic shoe market. In general, as the number of competitors in an industry declines, then market control of the remaining firms is enhanced.

Because oligopoly has a small number of firms, the incentive to cooperate through mergers is quite high. The large number of firms in monopolistic competition, by contrast, provides very few merger benefits. The merger of two monopolistically competitive firms, each with one-thousandth of the overall market, does not enhanced market control much at all. However, the merger of two oligopolistic firms, each with one-third of the market, greatly enhances the market control of the new firm.

**COLLABORATION**

The incentive among oligopolistic firms to cooperate also takes the form of collusion. With collusion, oligopolistic firms remain legally independent and autonomous, but they enjoy the benefits of cooperation. Collusion occurs when two or more firms secretly agree to control prices, production, or other aspects of the market. For example, OmniRun and Master Foot might secretly agree to raise their shoe prices to $150 a pair. If these are the only two firms in the market for athletic shoes, then buyers have no choice but to pay the $150 price. In effect, the two firms operate as if they were one firm, a monopoly.
By acting like a monopoly, the colluding firms can set a monopoly price, produce a monopoly quantity, generate monopoly profit; and allocate resources as inefficiently as a monopoly.

Collusion can take one of two forms. Explicit collusion results when two or more firms reach a formal agreement. Implicit collusion results when two or more firms informally control the market with necessarily reaching a formal agreement.

Given that collusion is usually illegal, especially within the United States, it is invariably kept secret. Some collusive agreements, however, are anything but secret. The most well-known example is the Organization of Petroleum Exporting Countries (OPEC). OPEC is an open, formal collusive agreement among petroleum producing countries to control prices and production.

**Two Types of Collusion**

Collusion can take one of two forms--explicit collusion and implicit collusion.

- **Explicit Collusion**: Also termed overt collusion, this occurs when two or more firms in the same industry formally agree to control the market. Admittedly, because collusion in the United States and most industrialized countries is illegal, such a formal agreement is likely to be highly secret and unlikely to be documented in any way. It might involve nothing more than a "casual" lunch among company presidents, a "chance" meeting at a conference of industry executives, or company decision-makers skulking around back alleys in the dead of the night discussing price charges.

- **Implicit Collusion**: Also termed tacit collusion, this occurs when two or more firms in the same industry informally agree to control the market, often through nothing more than interdependent actions. A prime example of implicit collusion is **price leadership**. In this case, one firm takes the lead of setting a price that will boost profits for the entire industry. Other firms then go along with this price, knowing that they stand to benefit by doing so.

**PRICE LEADERSHIP**

A method used by a group of firms in the same market (typically oligopoly firms) in which one firm takes the lead in setting or changing prices, with other firms then following behind. The lead firm is often the largest firm in the industry, but it could be a smaller firm that has just historically assumed the role of price leader perhaps because it is more aware of changing market conditions. While price leadership is totally legal, it could be a sign of collusion, particular implicit collusion, in which the firms have effectively monopolized the market.

**CARTEL**

One of the most noted examples of explicit collusion is a cartel. While the term cartel can be used to mean any type of explicit collusion, it is often reserved for international
agreements, such as the Organization of Petroleum Exporting Countries (better known as OPEC).

OPEC is perhaps the most famous international cartel, which exerts control over the world petroleum market. International cartels, more often than not, officially are political treaties among countries. However, when the countries also control the production of a good like petroleum, and when the treaty is primarily designed as a means of influencing the global market for this good, then the treaty also becomes a formal economic arrangement and an example of explicit collusion.

**ACTING LIKE MONOPOLY**

In general, collusion among oligopolistic firms means that two or more firms decide to act like a monopoly. Rather than maximizing profit for each individual firm, the firms maximize total industry profit just as if a monopoly controlled the industry. Motivation behind collusion is relatively straightforward. Total profit is greater when firms collude than when they compete. Cooperating firms can agree to charge a higher price and produce less output—just like a monopoly.

A side benefit for the colluding firms is that non-colluding firms can be driven from the market. For example, the top two soft drink firms that control a sizeable share of the hypothetical Shady Valley soft drink market might decide to do a little colluding. While their ultimate goal is to raise the price, they might first agree to lower the price. By so doing, they can force other smaller firms out of the market. Once these other firms have left, then their market control increases, making their collusion more effective. They can then raise the price and more effectively act like a monopoly, without concern that the smaller firms will undercut their collusion price.

![Collusion Production](image)

**Figure 3 Collusion Production**

**Collusion Production**
The exhibit above summarizes collusion among two firms, OmniCola and Juice-Up, in the hypothetical oligopolistic Shady Valley soft drink industry. Here is an overview of the analysis:

- **Two Firms:** The cost curves for the two oligopoly firms in this analysis are presented in the far left (for OmniCola) and middle (Juice-Up) panels. Juice-Up has higher costs than OmniCola.
- **Industry Marginal Cost:** The marginal cost curves for the two firms are combined in an industry marginal cost curve, labeled MC\text{m}, presented in the far right panel. This curve indicates the change in cost, using the production plants of both firms, that is incurred by producing one more unit of output.
- **Demand and Marginal Revenue:** The far right panel also presents the demand curve for the soft drink market, labeled D, and what would be the marginal revenue curve, labeled MR, for a monopoly seller, which is what the two firms have become.
- **Profit Maximization:** Total industry profit is maximized by equating the industry marginal cost curve (MC\text{m}) with the monopoly marginal revenue curve (MR). This is achieved at a quantity of 16,000 cans.
- **Setting Price:** Once the profit-maximizing quantity is identified, the colluding firms act just like a monopoly to set the price. They determine the demand price that buyers are willing to pay for the 16,000 cans quantity, which is $1 per can.
- **Dividing Production:** The 16,000 cans of total production is divided between OmniCola and Juice-Up based on their individual marginal costs. OmniCola produces 10,000 cans and Juice-Up produces 6,000 cans.
- **Profit to Each:** The profit received by each firm is then the difference between revenue generated at the $1 price and the total cost each firm incurs when producing its quota of soft drinks. This profit is indicated by the yellow areas.

The key to this collusion is that the yellow areas, in total, are larger with collusion than with competition.

### 3.7 KINKED DEMAND CURVE UNDER OLIGOPOLY

The kinked demand curve theory is an economic theory regarding oligopoly and monopolistic competition. When it was created, the idea fundamentally challenged classical economic tenets such as efficient markets and rapidly-changing prices, ideas that underly basic supply and demand models. Kinked demand was an initial attempt to explain sticky prices.

An oligopolist faces a downward sloping demand curve but the elasticity may depend on the reaction of rivals to changes in price and output. Assuming that firms are attempting to maintain a high level of profits and their market share it may be the case that:

(a) rivals will not follow a price increase by one firm - therefore demand will be relatively elastic and a rise in price would lead to a fall in the total revenue of the firm
(b) rivals are more likely to match a price fall by one firm to avoid a loss of market share. If this happens demand will be more inelastic and a fall in price will also lead to a fall in total revenue.

The kink in the demand curve at price $P$ and output $Q$ means that there is a discontinuity in the firm’s marginal revenue curve. If we assume that the marginal cost curve is cutting the MR curve then the firm is maximising profits at this point.
In the bottom diagram, we see that a rise in marginal costs will not necessarily lead to higher prices providing that the new MC curve (MC2) cuts the MR curve at the same output. The kinked demand curve theory suggests that there will be price stickiness in these markets and that firms will rely more on non-price competition to boost sales, revenue and profits.

**Figure 6**

**OLIGOPOLY, REALISM**

Real world markets are heavily populated by oligopoly. About half of all output produced in the U.S. economy each year is done so by oligopoly firms. Other industrialized nations can make a similar claim. Oligopoly markets arise in a wide assortment different industries, ranging from manufacturing to retail trade to resource extraction to financial services.

Oligopoly markets provide a veritable who's who of business firms in the United States and throughout the global economy. Any listing of oligopolistic industries, let alone specific oligopolistic firms, is bound to be incomplete. However, here is a partial list.

**Automobiles**

At the top of the list is the market for cars, which has been one of the most important industries in this country for decades. A handful of firms--especially General Motors, Ford, Chrysler, Honda, and Toyota--account for over 90 percent of the cars, trucks, vans, and sport utility vehicles sold in the United States. The need for large factories, a nationwide network of dealerships, and brand name recognition create entry barriers that limit production to a few large firms.
Petroleum
An industry closely tied to the market for cars is the extraction and refinement of petroleum. A few representatives in this market include ExxonMobile, ConocoPhillips, Gulf, and Shell. While the petroleum industry contains hundreds, if not thousands, of smaller firms, the biggest ones tend to dominate the market. In addition, another major player on the international scene is the Organization of Petroleum Exporting Countries (OPEC), which is an international cartel representing several petroleum-rich countries especially in the Middle East. Ownership and control of petroleum resources is a prime factor in the creation of an oligopolistic industry.

Tires
Another industry closely connected to the automobile industry is the manufacture of tires. Every car needs tires. This industry is also dominated by a small number of familiar firms as well, including Goodyear, Firestone, Goodrich, Uniroyal, and Michelin. The number of firms in this industry is also limited by the need for large factories.

Computers
An increasingly important market is that for personal computers. The manufacture of computers tends to be dominated by a small number of firms, including Dell, Hewlett-Packard, Gateway, Apple, and IBM. In the early years of the computer revolution (1970s and 1980s), numerous firms entered the market (Kaypro, Osborne, Packard Bell, Compaq, Texas Instruments), then went bankrupt, merged with other firms, or simply stopped offering personal computers. Although entry barriers are not insurmountable, brand name recognition and the need for manufacturing facilities tend to limit the entry of other firms.

Banking
A market that is becoming increasingly oligopolistic is banking. While the United States has a total of approximately 20,000 banks, a small contingent of firms tends to dominate the national market. Names include Citibank, Bank of America, Wells Fargo, Bank One, and MBNA. Moreover, most banking is done at the local level and most cities generally have no more than a handful of firms, including these national firms, that provide banking services. The key entry barrier that limits the number of firms in the banking industry is government authorization. Before a firm can provide banking services, it must obtain a government charter.

Wireless Telephone
Throughout much of the 1900s, the only company providing telephone services was AT&T. Technological advances and regulatory changes enabled the development of an oligopolistic market for wireless telephone services (cell phones). AT&T has been joined by a small number of other companies, including Verizon, Cingular, Sprint, T-Mobile. The need for a nation-wide network of relay towers makes this industry well suited for a small number of large companies.
**Television**

While television sets are filled with hundreds of television channels, only a handful of companies dominate the market. The big players, including Disney (ABC, ESPN, Disney), Viacom (CBS, UPN, MTV), General Electric (NBC, Bravo, CNBC), Time-Warner (HBO, WB, CNN, TBS), and News Corp. (Fox, FX, Fox News), own and control many of the channels. (These companies also play major roles in related markets—including motion pictures, cable systems, radio and television stations, and newspapers.) Domination by a few firms arises due to the upfront costs of producing programming and acquiring satellite relay access.

**Airlines**

The airline industry has long been dominated by a small number of firms. Throughout the middle part of the 1900s, seven firms dominated the U.S. market—American, United, TWA, PanAm, Continental, Braniff, and Eastern. However, deregulation in the 1980s lead to decades of changes. Some airlines folded. New airlines emerged. Even though changes continue, the industry remains dominated by a small number of competitors—American, United, Southwest, Delta. Heavy expenses needed to purchase planes, establish flight routes, and acquire terminal space tends to limit the entry of new firms.

### 3.8 MONOPSONY

A market characterized by a single buyer of a product. Monopsony is the buying-side equivalent of a selling-side monopoly. Much as a monopoly is the only seller in a market, monopsony is the only buyer. While monopsony could be analyzed for any type of market it tends to be most relevant for factor markets in which a single firm is the only buyer of a factor. Two related buying side market structures are oligopsony and monopsonistic competition.

Monopsony is a market in which a single buyer completely controls the demand for a good. While the market for any type of good, service, resource, or commodity could, in principle, function as monopsony, this form of market structure tends to be most pronounced for the exchange of factor services.

While the real world does not contain monopsony in its absolute purest form, labor markets in which a single large factory is the dominate employer in a small community comes as close as any.

Like a monopoly seller, a monopsony buyer is a price maker with complete market control. Monopsony is also comparable to monopoly in terms of inefficiency. Monopsony does not generate an efficient allocation of resources. The price paid by a monopsony is lower and the quantity exchanged is less than with the benchmark of perfect competition.

**Characteristics**

The three key characteristics of monopsony are: (1) a single firm buying all output in a market, (2) no alternative buyers, and (3) restrictions on entry into the industry.
• Single Buyer: First and foremost, a monopsony is a monopsony because it is the only buyer in the market. The word monopsony actually translates as "one buyer." As the only buyer, a monopsony controls the demand-side of the market completely. If anyone wants to sell the good, they must sell to the monopoly.

• No Alternatives: A monopsony achieves single-buyer status because sellers have no alternative buyers for their goods. This is the key characteristics that usually prevents monopsony from existing in the real world in its pure, ideal form. Sellers almost always have alternatives.

• Barriers to Entry: A monopsony often acquires and generally maintains single buyer status due to restrictions on the entry of other buyers into the market. The key barriers to entry are much the same as those that exist for monopoly: (1) government license or franchise, (2) resource ownership, (3) patents and copyrights, (4) high start-up cost, and (5) decreasing average total cost.

A Hypothetical Example
One example of a monopsony factor market is the hypothetical Natural Ned Lumber Company, which is a lumbering operation in the isolated Jagged Mountains region north of the greater Shady Valley metropolitan area. The Natural Ned Lumber Company is an expansive operation employing several thousand workers, all of whom reside in Lumber Town, which is adjacent to the Natural Ned Lumber Company lumbering operations. In fact, everyone living in Lumber Town works for the Natural Ned Lumber Company.

This makes the Natural Ned Lumber Company a monopsony employer. If anyone in Lumber Town seeks employment, then they must seek it with the Natural Ned Lumber Company. As such, the Natural Ned Lumber Company is a price maker when it comes to buying labor services. The Natural Ned Lumber Company can the determine of labor services desired, then charge the minimum factor price that sellers are willing and able to receive.

While the Natural Ned Lumber Company and Lumber Town is obviously a fictitious example of a monopsony, it does illustrate one of the more prevalent categories of monopsony that existed in the early history of the U.S. economy--the company town. During the early days of the U.S. industrial revolution, the late 1800s through the early 1900s, it was quite common for a large industrial facility (factory, mining operation, lumber company) to dominate employment in a given area. In some cases, the company literally built and owned the town in which the workers lived. Even those people who did not work directly in the primary activity (mining, lumber, etc.) worked in the company-owned store, hospital, school, or theater. Hence the term company town.

MODERN MONOPSONY
Like other extreme market structures (perfect competition and monopoly) monopsony is only approximated in the real world. Achieving the status of THE ONLY BUYER is not easy. Few if any buyers actually achieve this status. However, several have come close. In modern times a few examples of markets that come very close to monopsony come from the world of sports.
Should a talented quarterback wish to obtain a job as a professional football player, then THE employer is the National Football League (NFL). Of course, the NFL is not absolutely the ONLY employer. Employment as a professional football player can also be found with the Canadian Football League (CFL). However, sufficient difference exists between these two employers to give the NFL significant monopsony control.

Similar near monopsony status exists for other professional sports. A professional baseball player seeks employment with Major League Baseball (MLB), with minimal competition from Japan. A profession basketball player seeks employment with the National Basketball Association (NBA), with minimal competition from Europe. A profession hockey player seeks employment with the National Hockey League (NHL), again with some competition from Europe.

Other modern markets that exhibit varying degrees of monopsony status can be found in collegiate sports and the National Collegiate Athletic Association (NCAA) and the medical profession and the American Medical Association (AMA). Much like a professional football, baseball, basketball, or hockey player seeks employment in the "big leagues," a collegiate athlete seeks "employment" with a college affiliated with the NCAA. In a similar manner, a physician seeks "employment" through the AMA.

The reasons for quotation marks around employment for these two examples is that the monopsony employer does not technically employ these workers in a traditional sense. Monopsony status, however, is attributable to the ability to influence the factor market. In other words, a collegiate athlete who does not satisfy NCAA guidelines has difficulty "working" for a university, that is providing athletic entertainment services through the college in return for a scholarship. A physician who does not satisfy AMA guidelines also has difficulty working at a hospital or in private practice.

**Supply and Cost**

Single-buyer status means that monopsony faces a positively-sloped supply curve, such as the one displayed in the exhibit to the right. In fact, the supply curve facing the monopsony is the market supply curve for the product.

The far right curve in the exhibit is the red supply curve (S) facing the monopsony. The far left curve is the brown marginal factor cost curve (MFC). The marginal factor cost curve indicates the change in total factor cost incurred due to buying one additional unit of the good.

Because a monopsony is a price maker with extensive market control, it faces a positively-sloped supply curve. To buy a larger quantity of output, it must pay a higher price. For example, the monopsony can hire 10,000 workers...
for a wage of $5. However, if it wants to hire 20,000 workers, then it must raise the wage to $6.10.

For this reason, the marginal factor cost incurred from hiring extra workers is greater than the wage, or factor price. Suppose for example that the factor price needed to hire ten workers is $5 and the factor price needed to hire eleven workers is $5.10. The marginal factor cost incurred due to hiring the eleventh unit is $6.10. While the $6 factor price means the monopsony incurs a $5.10 factor cost from hiring this worker, this cost is compounded by an extra cost of $1 due to the higher wage paid to the first ten workers. The overall increase in cost, that is marginal factor cost, is thus $6.10 (= $5.10 + $1).

**PROFIT MAXIMIZING EMPLOYMENT**

The exhibit to the right illustrates the profit-maximizing employment of a monopsony. This firm faces a positively-sloped supply curve, represented by the red supply curve (S). Because higher wages are needed to attract more labor, the positively-slope brown curve is the marginal factor cost curve (MFC). The third curve displayed in the exhibit is the green negatively-sloped marginal revenue product curve (MRP), which indicates the value of the extra production generated by each worker.

As a profit-maximizing firm, monopsony hires the quantity that equates marginal factor cost and marginal revenue product found at the intersection of the MFC and MRP curves. This quantity is 37,000 workers. The monopsony then pays each worker $8.40.

This price and quantity maximizes profit because the revenue generated from hiring the last worker (marginal revenue product) is exactly equal to the cost incurred from hiring the last worker (marginal factor cost). Because the extra revenue generated equals the extra cost incurred, there is no way to increase profit by hiring more or less of this input.

### 3.9 BILATERAL MONOPOLY

A market containing a single buyer and a single seller, or the combination of a monopoly market and a monopsony market. A market dominated by a profit-maximizing monopoly tends to charge a higher price. A market dominated by a profit-maximizing monopsony tends to pay a lower price. When combined into a bilateral monopoly, the buyer and seller both cannot maximize profit simultaneously and are forced to negotiate a price and quantity. Then resulting price could be anywhere between the higher monopoly price and the lower monopsony price. Where the price ends ups depends on the relative negotiating power of each side.
The bilateral monopoly model, with a single buyer and a single seller, can be used to analyze many types of markets, but it is most relevant for factor markets, especially those for labor services.

The bilateral monopoly model was developed to explain assorted labor markets operating in the early days of the U.S. industrial revolution, the late 1800s and early 1900s. During this period, large industrial activities (factories, mines, lumber operations) commonly created monopsony markets by dominating the labor market of a given community (a so-called company town). The expected monopsony outcome, especially low wages, inevitably resulted.

The workers sought to counter these less than desirable situations, by forming labor unions. The expressed goal of most unions was to monopolize the selling side of a labor market AND balance the monopsony power of the employer. This resulted in a bilateral monopoly.

In fact, a bilateral monopoly emerges by combining the Natural Ned Lumber Company monopsony with the United Tree Choppers Union monopoly. Because the profit-maximizing price charged by the United Tree Choppers Union monopoly ($15) is higher than the profit-maximizing price paid by the Natural Ned Lumber Company monopsony ($8.50), both sides cannot maximize profit. The price ultimately achieved in this market depends on the relatively negotiating power of each side.

This exhibit summarizes the bilateral monopoly for the employment of tree choppers. A word or two about the four main curves in the exhibit is needed.

- Marginal Revenue Product and Demand: The first of the two negatively-sloped curves (D/MRP) is the marginal revenue product curve for employment by the Natural Ned Lumber Company monopsony, the MRP part. This curve, however, is also the demand curve for labor services facing the United Tree Choppers Union monopoly, the D part.
- Marginal Revenue: The second negatively-sloped curve (MR) is the marginal revenue curve when the D/MRP curve is viewed as the demand curve facing for the United Tree Choppers Union monopoly. This marginal revenue curve indicates how much additional revenue the monopoly union generates when an extra worker is employed.
Before moving on to the other two curves, make certain the distinction between these two marginal revenue curves is clear. The MRP curve is based on the revenue received by the Natural Ned Lumber Company when it hires labor to produce lumber.

When it is viewed as the demand curve for labor services, it is also the average revenue curve for the United Tree Choppers Union monopoly. And like any demand curve that is also an average revenue curve, it has a corresponding marginal revenue curve, which is the second curve, labeled MR. The MR curve thus represents the additional revenue received by the United Tree Choppers Union for selling additional units of its "output," which is labor services.

- Marginal Cost and Supply: The first of two positively-sloped curves (S/MC) is the marginal cost curve, which is the marginal cost incurred by the United Tree Choppers Union for selling labor services, the S part. This curve is also the supply curve for labor services facing the Natural Ned Lumber Company monopsony employer, the MC part.
- Marginal Factor Cost: The second positively-sloped curve (MFC) is the marginal factor cost curve when the S/MC curve is viewed as the supply curve facing the Natural Ned Lumber Company monopsony. This marginal factor cost curve indicates how much additional cost the monopsony incurs when an extra worker is employed.

The distinction between these two marginal cost curves is similar to the distinction between the two marginal revenue curves. The MC curve is based on the cost incurred by the United Tree Choppers Union when selling labor to the lumber company.

When it is viewed as the supply curve for labor services, it is also the average factor cost curve for the Natural Ned Lumber Company monopsony buying the labor. And like any supply curve that is also an average factor cost curve, it has a corresponding marginal factor cost curve, which is the second curve, labeled MFC. The MFC curve thus represents the additional cost incurred by the monopsony resort for buying additional units of labor services.

**Activity 3**

1. Discuss the main characteristics of monopolistic competition with special reference to product differentiation.
2. Give a brief note on comparison between general and chamberlain equilibrium approaches to monopolistic competition.
3. What do you understand by kinked demand curve in context of oligopoly?
4. Give your views on monopsony and bilateral monopoly with reference to real world examples.
3.10 SUMMARY
Monopolistic competition contains a large number of small firms, each with some, but not a lot of market control. Oligopoly contains a few large firms that dominate a market. However, monopolistic competition and oligoplies are actually the heart and soul of the market structure continuum. Further concepts related to general equilibrium theory and chamberlain approach were discussed in detail. Followed by criticism to monopolistic competition the oligopoly behavior with kinked demand curve was enumerated. Recent concepts of monopsony which is a market form in which only one buyer faces many sellers along with bilateral monopoly were explained with suitable examples.

3.11 FURTHER READINGS

- Eaton, Eaton and Allen, "Intermediate Microeconomics"
- Managerial Economics. "G S Gupta"
- Robinson, J. (1933) the Economics of Imperfect Competition London: Macmillan.
M.A. PREVIOUS ECONOMICS

PAPER I

MICRO ECONOMIC ANALYSIS

BLOCK 4

DISTRIBUTION AND PIGOVIAN WELFARE ECONOMICS
PAPER I

MICRO ECONOMIC ANALYSIS

BLOCK 4

DISTRIBUTION AND PIGOVIAN WELFARE ECONOMICS

CONTENTS

<table>
<thead>
<tr>
<th>Unit</th>
<th>Title</th>
<th>Page number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit 1</td>
<td>The Classical and Neoclassical theories of distribution</td>
<td>4</td>
</tr>
<tr>
<td>Unit 2</td>
<td>Distribution and related aspects</td>
<td>21</td>
</tr>
<tr>
<td>Unit 3</td>
<td>Fundamental theorems of welfare economics</td>
<td>49</td>
</tr>
<tr>
<td>Unit 4</td>
<td>Basic concepts of welfare economics</td>
<td>63</td>
</tr>
<tr>
<td>Unit 5</td>
<td>The second best theory and impossibility theorem</td>
<td>80</td>
</tr>
</tbody>
</table>
This block presents to you various concepts related to distributions and pigovian welfare economics. Unit 1 explains the Classical and Neoclassical theories of distribution. Ricardian theory and approach of Karl marx are discussed in detail.

Unit 2 throws light on Distribution and related aspects. Marginal productivity theory of distribution has been discussed with factor payment and concept of rent. Substitution and distribution and the technical progress are other areas of concern.

Unit 3 describes Fundamental theorems of welfare economics. Various approaches related to these theorems are discussed in detail to give readers a broad spectrum.

Unit 4 identifies Basic concepts of welfare economics in which Pareto optimal conditions, Value judgment, Social welfare function and Compensation principle are specified with market imperfections, incomplete markets and market failure. A brief review of optimum welfare is also given.

At last unit 5 discusses the second best theory and impossibility theorem. Arrow’s impossibility theorem; The voting paradox and Arrow Debreu model are discussed with suitable examples.
UNIT 1

THE CLASSICAL AND NEOCLASSICAL THEORIES OF DISTRIBUTION

Objectives

On successful completion of this unit, you should be able to:

- Appreciate the classical theory of distribution.
- Identify the aspects of neoclassical theory of distribution.
- Know the Ricardian contribution to distribution theory.
- Recognize the Karl Marx’s approach of value in perspective of distribution theory.

Structure

1.1 Introduction
1.2 The classical theory of distribution
1.3 Neo classical distribution theory
1.4 Ricardian theory of rent, land and labor
1.5 Contributions of Karl Marx to distribution theory
1.6 Summary
1.7 Further readings

1.1 INTRODUCTION

Distribution in economics refers to the way total output or income is distributed among individuals or among the factors of production (labor, land, and capital) (Samuelson and Nordhaus, 2001, p. 762). In general theory and the national income and product accounts, each unit of output corresponds to a unit of income. One use of national accounts is for classifying factor incomes and measuring their respective shares, as in National Income. But, where focus is on income of persons or households, adjustments to the national accounts or other data sources are frequently used. Here, interest is often on the fraction of income going to the top (or bottom) \( x \) percent of households, the next \( y \) percent, and so forth (say in quintiles), and on the factors that might affect them (globalization, tax policy, technology, etc.).
1.2 THE CLASSICAL THEORY OF DISTRIBUTION

Classical theory of distribution in economics is widely regarded as the first modern school of economic thought. It is associated with the idea that free markets can regulate themselves. Its major developers include Adam Smith, David Ricardo, Thomas Malthus and John Stuart Mill. Sometimes the definition of classical economics is expanded to include William Petty, Johann Heinrich von Thünen.

Adam Smith's *The Wealth of Nations* in 1776 is usually considered to mark the beginning of classical economics. The school was active into the mid 19th century and was followed by neoclassical economics in Britain beginning around 1870.

They produced their "magnificent dynamics" during a period in which capitalism was emerging from a past feudal society and in which the industrial revolution was leading to vast changes in society. These changes also raised the question of how a society could be organized around a system in which every individual sought his or her own (monetary) gain.

Classical economists and their immediate predecessor reoriented economics away from an analysis of the ruler's personal interests to broader national interests. Physiocrat Francois Quesnay and Adam Smith, for example, identified the wealth of a nation with the yearly national income, instead of the king's treasury. Smith saw this income as produced by labor, land, and capital equipment. With property rights to land and capital by individuals, the national income is divided up between laborers, landlords, and capitalists in the form of wages, rent, and interest or profits.

**Value Theory**

Classical economists developed a theory of value, or price, to investigate economic dynamics. Petty introduced a fundamental distinction between market price and natural price to facilitate the portrayal of regularities in prices. Market prices are jostled by many transient influences that are difficult to theorize about at any abstract level. Natural prices, according to Petty, Smith, and Ricardo, for example, capture systematic and persistent forces operating at a point in time. Market prices always tend toward natural prices in a process that Smith described as somewhat similar to gravitational attraction.

The theory of what determined natural prices varied within the Classical school. Petty tried to develop a par between land and labor and had what might be called a land-and-labor theory of value. Smith confined the labor theory of value to a mythical pre-capitalist past. He stated that natural prices were the sum of natural rates of wages, profits (including interest on capital and wages of superintendence) and rent. Ricardo also had what might be described as a cost of production theory of value. He criticized Smith for describing rent as price-determining, instead of price-determined, and saw the labor theory of value as a good approximation.
Some historians of economic thought, in particular, Sraffian economists, see the classical theory of prices as determined from three givens:

1. The level of outputs at the level of Smith's "effectual demand",
2. technology, and
3. wages.

From these givens, one can rigorously derive a theory of value. But neither Ricardo nor Marx, the most rigorous investigators of the theory of value during the Classical period, developed this theory fully. Those who reconstruct the theory of value in this manner see the determinants of natural prices as being explained by the Classical economists from within the theory of economics, albeit at a lower level of abstraction. For example, the theory of wages was closely connected to the theory of population. The Classical economists took the theory of the determinants of the level and growth of population as part of Political Economy. Since then, the theory of population has been seen as part of some other discipline than economics. In contrast to the Classical theory, the determinants of the neoclassical theory value:

1. tastes
2. technology, and
3. endowments

are seen as exogenous to neoclassical economics.

Classical economics tended to stress the benefits of trade. Its theory of value was largely displaced by marginalist schools of thought which sees "use value" as deriving from the marginal utility that consumers finds in a good, and "exchange value" (i.e. natural price) as determined by the marginal opportunity- or disutility-cost of the inputs that make up the product. Ironically, considering the attachment of many classical economists to the free market, the largest school of economic thought that still adheres to classical form is the Marxian school.

**Monetary Theory**

British classical economists in the 19th century had a well-developed controversy between the Banking and the Currency school. This parallels recent debates between proponents of the theory of endogeneous money, such as Nicholas Kaldor, and monetarists, such as Milton Friedman. Monetarists and members of the currency school argued that banks can and should control the supply of money. According to their theories, inflation is caused by banks issuing an excessive supply of money. According to proponents of the theory of endogenous money, the supply of money automatically adjusts to the demand, and banks can only control the terms (e.g., the rate of interest) on which loans are made.
Debates on the definition of Classical theory of distribution.

The theory of value is currently a contested subject. One issue is whether classical economics is a forerunner of neoclassical economics or a school of thought that had a distinct theory of value, distribution, and growth.

Sraffians, who emphasize the discontinuity thesis, see classical economics as extending from William Petty's work in the 17th century to the break-up of the Ricardian system around 1830. The period between 1830 and the 1870s would then be dominated by "vulgar political economy", as Karl Marx characterized it. Sraffians argue that: the wages fund theory; Senior's abstinence theory of interest, which puts the return to capital on the same level as returns to land and labor; the explanation of equilibrium prices by well-behaved supply and demand functions; and Say's law, are not necessary or essential elements of the classical theory of value and distribution.

Perhaps Schumpeter's view that John Stuart Mill put forth a half-way house between classical and neoclassical economics is consistent with this view.

Sraffians generally see Marx as having rediscovered and restated the logic of classical economics, albeit for his own purposes. Others, such as Schumpeter, think of Marx as a follower of Ricardo. Even Samuel Hollander has recently explained that there is a textual basis in the classical economists for Marx's reading, although he does argue that it is an extremely narrow set of texts.

The first position is that neoclassical economics is essentially continuous with classical economics. To scholars promoting this view, there is no hard and fast line between classical and neoclassical economics. There may be shifts of emphasis, such as between the long run and the short run and between supply and demand, but the neoclassical concepts are to be found confused or in embryo in classical economics. To these economists, there is only one theory of value and distribution. Alfred Marshall is a well-known promoter of this view. Samuel Hollander is probably its best current proponent.

A second position sees two threads simultaneously being developed in classical economics. In this view, neoclassical economics is a development of certain exoteric (popular) views in Adam Smith. Ricardo was a sport, developing certain esoteric (known by only the select) views in Adam Smith. This view can be found in W. Stanley Jevons, who referred to Ricardo as something like "that able, but wrong-headed man" who put economics on the "wrong track". One can also find this view in Maurice Dobb's Theories of Value and Distribution Since Adam Smith: Ideology and Economic Theory (1973), as well as in Karl Marx's Theories of Surplus Value.

The above does not exhaust the possibilities. John Maynard Keynes thought of classical economics as starting with Ricardo and being ended by the publication of Keynes' General Theory of Employment Interest and Money. The defining criterion of classical economics, on this view, is Say's law.
One difficulty in these debates is that the participants are frequently arguing about whether there is a non-neoclassical theories that should be reconstructed and applied today to describe capitalist economies. Some, such as Terry Peach\textsuperscript{[5]}, see classical economics as of antiquarian interest.

### 1.3 NEOCLASSICAL DISTRIBUTION THEORY

In **neoclassical economics**, the supply and demand of each factor of production interact in factor markets to determine equilibrium output, income, and the income distribution. Factor demand in turn incorporates the marginal-productivity relationship of that factor in the output market. Analysis applies to not only capital and land but the distribution of income in labor markets (Hicks, 1963). In a **perfectly competitive** economy, market equilibrium results in **allocative efficiency** as to the mix of output produced and **distributive efficiency** in the least-cost mix of factors of production. In 1908, the efficiency properties of perfect competition were shown by Enrico Barone to be required as well for efficient resource use in **collectivist planning**.

The **neoclassical growth model** provides an account of how distribution of income between capital and labor are determined in competitive markets at the macroeconomic level over time with technological change and changes in the size of the capital stock and labor force. More recent developments of the distinction between human capital and physical capital and between social capital and personal capital have deepened analysis of distribution.

In fact Neoclassical economics is a term variously used for approaches to economics focusing on the determination of prices, outputs, and income distributions in markets through supply and demand, often as mediated through a hypothesized maximization of income-constrained utility by individuals and of cost-constrained profits of firms employing available information and factors of production, in accordance with **rational choice theory**.\textsuperscript{[1]} Neoclassical economics dominates microeconomics, and together with Keynesian economics forms the neoclassical synthesis, which dominates mainstream economics today.\textsuperscript{[2]} There have been many critiques of neoclassical economics, often incorporated into newer versions of neoclassical theory as human awareness of economic criteria change.

The term was originally introduced by Thorstein Veblen in 1900, in his *Preconceptions of Economic Science*, to distinguish marginalists in the tradition of Alfred Marshall from those in the Austrian School.\textsuperscript{[3][4]} It was later used by John Hicks, George Stigler, and others who presumed that significant disputes amongst marginalist schools had been largely resolved\textsuperscript{[5]} to include the work of Carl Menger, William Stanley Jevons, John Bates Clark and many others.\textsuperscript{[4]} Today it is usually used to refer to mainstream economics, although it has also been used as an umbrella term encompassing a number of mainly defunct schools of thought,\textsuperscript{[6]} notably excluding institutional economics, various historical schools of economics, and Marxian economics, in addition to various other heterodox approaches to economics.
Neoclassical microeconomics of labor markets

Economists see the labor market as similar to any other market in that the forces of supply and demand jointly determine price (in this case the wage rate) and quantity (in this case the number of people employed).

However, the labor market differs from other markets (like the markets for goods or the money market) in several ways. Perhaps the most important of these differences is the function of supply and demand in setting price and quantity. In markets for goods, if the price is high there is a tendency in the long run for more goods to be produced until the demand is satisfied. With labor, overall supply cannot effectively be manufactured because people have a limited amount of time in the day, and people are not manufactured. A rise in overall wages will, in many situations, not result in more supply of labor: it may result in less supply of labor as workers take more time off to spend their increased wages, or it may result in no change in supply. Within the overall labour market, particular segments are thought to be subject to more normal rules of supply and demand as workers are likely to change job types in response to differing wage rates.

Many economists have thought that, in the absence of laws or organizations such as unions or large multinational corporations, labor markets can be close to perfectly competitive in the economic sense — that is, there are many workers and employers both having perfect information about each other and there are no transaction costs. The competitive assumption leads to clear conclusions — workers earn their marginal product of labor.

Other economists focus on deviations from perfectly competitive labor markets. These include job search, training and gaining-of-experience costs to switch between job types, efficiency wage models and oligopsony / monopolistic competition.

Demand for labor and wage determination

Labor demand is a derived demand, in other words the employer's cost of production is the wage, in which the business or firm benefits from an increased output or revenue. The determinants of employing the addition to labour depend on the Marginal Revenue Product (MRP) of the worker. The MRP is calculated by multiplying the price of the end product or service by the Marginal Physical Product of the worker. If the MRP is greater than a firm's Marginal Cost, then the firm will employ the worker. The firm only employs however up to the point where MRP=MC, not lower, in economic theory.

Wage differences exist, particularly in mixed and fully/partly flexible labour markets. For example, the wages of a doctor and a port cleaner, both employed by the NHS, differ greatly. But why? There are many factors concerning this issue. This includes the MRP (see above) of the worker. A doctor's MRP is far greater than that of the port cleaner. In addition, the barriers to becoming a doctor are far greater than that of becoming a port cleaner. For example to become a doctor takes a lot of education and training which is costly, and only those who are socially and intellectually advantaged can succeed in such
a demanding profession. The port cleaner however requires minimal training. The supply of doctors therefore would be much more inelastic than the supply of port cleaners. The demand would also be inelastic as there is a high demand for doctors, so the NHS will pay higher wage rates to attract the profession.

**Neoclassical microeconomic model — Supply**

Households are suppliers of labor. In microeconomics theory, people are assumed rational and seeking to maximize their utility function. In this labor market model, their utility function is determined by the choice between income and leisure. However, they are constrained by the waking hours available to them.

Let \( w \) denote hourly wage.
Let \( k \) denote total waking hours.
Let \( L \) denote working hours.
Let \( n \) denote other incomes or benefits.
Let \( A \) denote leisure hours.

The utility function and **budget constraint** can be expressed as following:

\[
\text{max } U(w L + π, A) \text{ such that } L + A ≤ k.
\]

This can be shown in a that illustrates the trade-off between allocating your time between leisure activities and income generating activities. The linear constraint line indicates that there are only 24 hours in a day, and individuals must choose how much of this time to allocate to leisure activities and how much to working. (If multiple days are being considered the maximum number of hours that could be allocated towards leisure or work is about 16 due to the necessity of sleep) This allocation decision is informed by the curved *indifference curve* labeled IC. The curve indicates the combinations of leisure and work that will give the individual a specific level of utility. The point where the highest indifference curve is just tangent to the constraint line (point A), illustrates the short-run equilibrium for this supplier of labor services.
If the preference for consumption is measured by the value of income obtained, rather than work hours, this diagram can be used to show a variety of interesting effects. This is because the slope of the budget constraint becomes the wage rate. The point of optimization (point A) reflects the equivalency between the wage rate and the marginal rate of substitution, leisure for income (the slope of the indifference curve). Because the marginal rate of substitution, leisure for income, is also the ratio of the marginal utility of leisure ($MU_L$) to the marginal utility of income ($MU_Y$), one can conclude:
If wages increase, this individual's constraint line pivots up from $X,Y_1$ to $X,Y_2$. He/she can now purchase more goods and services. His/her utility will increase from point A on IC$_1$ to point B on IC$_2$. To understand what effect this might have on the decision of how many hours to work, you must look at the income effect and substitution effect.

The wage increase shown in the previous diagram can be decompiled into two separate effects. The pure income effect is shown as the movement from point A to point C in the next diagram. Consumption increases from $Y_A$ to $Y_C$ and — assuming leisure is a normal good — leisure time increases from $X_A$ to $X_C$ (employment time decreases by the same amount; $X_A$ to $X_C$).

But that is only part of the picture. As the wage rate rises, the worker will substitute work hours for leisure hours, that is, will work more hours to take advantage of the higher wage rate, or in other words substitute away from leisure because of its higher opportunity cost. This substitution effect is represented by the shift from point C to point B. The net impact of these two effects is shown by the shift from point A to point B. The relative magnitude of the two effects depends on the circumstances. In some cases the substitution effect is greater than the income effect (in which case more time will be allocated to working), but in other cases the income effect will be greater than the substitution effect (in which case less time is allocated to working). The intuition behind this latter case is that the worker has reached the point where his marginal utility of leisure outweighs his marginal utility of income. To put it in less formal (and less accurate) terms: there is no point in earning more money if you don't have the time to spend it.
Figure 4 the Labor Supply curve

If the substitution effect is greater than the income effect, the labour supply curve (diagram to the left) will slope upwards to the right, as it does at point E for example. This individual will continue to increase his supply of labor services as the wage rate increases up to point F where he is working $H_F$ hours (each period of time). Beyond this point he will start to reduce the amount of labor hours he supplies (for example at point G he has reduced his work hours to $H_G$). Where the supply curve is sloping upwards to the right (positive wage elasticity of labor supply), the substitution effect is greater than the income effect. Where it slopes upwards to the left (negative elasticity), the income effect is greater than the substitution effect. The direction of slope may change more than once for some individuals, and the labor supply curve is likely to be different for different individuals.

Other variables that affect this decision include taxation, welfare, and work environment.

**Neoclassical microeconomic model — Demand**

This article has examined the labour supply curve which illustrates at every wage rate the maximum quantity of hours a worker will be willing to supply to the economy per period of time. Economists also need to know the maximum quantity of hours an employer will demand at every wage rate. To understand the quantity of hours demanded per period of time it is necessary to look at product production. That is, labour demand is a derived demand: it is derived from the output levels in the goods market.

A firm's labour demand is based on its marginal physical product of labour (MPL). This is defined as the additional output (or physical product) that results from an increase of one unit of labour (or from an infinitesimally small increase in labour). If you are not familiar with these concepts, you might want to look at production theory basics before continuing with this article.
In most industries, and over the relevant range of outputs, the marginal physical product of labor is declining. That is, as more and more units of labor are employed, their additional output begins to decline. This is reflected by the slope of the $MPP_L$ curve in the diagram to the right. If the marginal physical product of labor is multiplied by the value of the output that it produces, we obtain the Value of marginal physical product of labor:

$$MPP_L \times P_Q = VMPP_L$$

The value of marginal physical product of labor (VMPP$_L$) is the value of the additional output produced by an additional unit of labor. This is illustrated in the diagram by the VMPP$_L$ curve that is above the MPP$_L$.

In competitive industries, the VMPP$_L$ is in identity with the marginal revenue product of labor (MRP$_L$). This is because in competitive markets price is equal to marginal revenue, and marginal revenue product is defined as the marginal physical product times the marginal revenue from the output (MRP = MPP \times MR).
The marginal revenue product of labor can be used as the demand for labor curve for this firm in the short run. In competitive markets, a firm faces a perfectly elastic supply of labor which corresponds with the wage rate and the marginal resource cost of labor \( W = S_L = MFC_L \). In imperfect markets, the diagram would have to be adjusted because \( MFC_L \) would then be equal to the wage rate divided by marginal costs. Because optimum resource allocation requires that marginal factor costs equal marginal revenue product, this firm would demand \( L \) units of labor as shown in the diagram.

**Neoclassical microeconomic model — Equilibrium**

The demand for labor of this firm can be summed with the demand for labor of all other firms in the economy to obtain the aggregate demand for labor. Likewise, the supply curves of all the individual workers (mentioned above) can be summed to obtain the aggregate supply of labor. These supply and demand curves can be analyzed in the same way as any other industry demand and supply curves to determine equilibrium wage and employment levels. (Morendy Octora)

**1.4 RICARDIAN THEORY OF RENT, LAND AND LABOUR**

"Political Economy, you think, is an enquiry into the nature and causes of wealth -- I think it should rather be called an enquiry into the laws which determine the division of produce of industry amongst the classes that concur in its formation. No law can be laid down respecting quantity, but a tolerably correct one can be laid down respecting proportions. Every day I am more satisfied that the former enquiry is vain and delusive, and the latter the only true object of the science."
David Ricardo followed Smith in the development of the labor theory of value. Ricardo's ideas, tightly reasoned and complex, were much more than a reconsideration of the labor theory. They gave directions to economic theory that, in many ways, it is still following even today. But we will be concerned here with Ricardo's contributions to the discussion on the Labor theory of Value. Ricardo disposed of one important objection to the labor theory and, without quite realizing it, raised two more.

The objection had to do with land and rent. On the one hand, land rent seems to be a cost of production -- shouldn't the "natural price" of an agricultural product depend on the rent of land? On the other hand, labor will be more productive on land that is more fertile. Crops grown on fertile land will cost less labor. Does that mean it has less value?

What Ricardo discovered is that the rent of land would be just enough to offset the differences in labor cost, so that the value of agricultural products would be the same regardless of the fertility of the land where they were produced. That is because rent is based on differential productivity.

Suppose you were a farmer, and you could rent either of two pieces of land, the north field and the south field. With the same labor and other inputs, the north field will produce more output than the south field. Let's say the output of the north field is worth 1100 labor-days and the output of the south field is worth 1000 labor-days, for a difference of 100 labor-days. Then you would be willing to pay up to 100 labor-days' more rent for the north field than for the south field, right? And the landowner, knowing that, wouldn't take less than 100 labor-days of additional rent for the north field, so the rent on the north field will be 100 more than the rent on the south field. That is, the difference in rent is the same as the difference in productivity.

But what will the total rent be, on each parcel of land? Let's suppose a third field, the east field, is so infertile that it isn't worth any rent at all. The east field, and any land so poor it isn't worth any rent, is called "marginal land." Let's suppose the east field can produce 800 labor days' worth of output, and that just pays the cost of cultivation, so there is nothing left for rent. We have seen that the north field produces 300 more labor-days of output than the east field, so it will get 300 more rent. That's 300 more than zero, so -- in other words -- the north field gets 300 man-days of rent in all. And the south field gets 200 labor-days' worth of rent in all.

Ricardo drew two conclusions:

1. It is the labor required for production on marginal land that determines the normal price or value of agricultural products.
2. The surplus of production on more fertile land is absorbed by rent. Landowners don't have to do anything to earn this rent -- they get it automatically as a result of the competition for fertile land.
Thus, Ricardo saved the Labor Theory from what might have been a troubling criticism. In the process he did two other things. First, he discovered a theory of the rent of land. Ricardo's theory of land rent is still regarded by modern economics as the correct theory. Second, Ricardo provided a theory of the distribution of income between landlords and other classes that is also still thought of as correct, but that has some important implications. This theory is not especially favorable to the landlords. According to Ricardo's theory, rents are determined by market processes in which the landlords need not do anything -- just let the farmers compete among themselves to rent the best land. This was no great surprise: everyone knew that landlords could be idle beneficiaries of their wealth -- but Ricardo's theory of rent made that seem, for the first time, an inevitable result of the market process.

Some of Ricardo's other contributions to economic thought were to raise further questions about the labor theory, though. In his discussion of the theory of international trade, in which he (again) discovered principles still central to modern economic theory in the field -- Ricardo stressed that trade would occur only on the basis of differences in the productivity of labor in different countries. That is, each country would export the product in which it had the comparative advantage of relatively lower labor cost. If there were no differences in labor costs, there would be no trade. But, again, this raises the question: if different producers have different labor costs, which labor costs determine exchange-value? Ricardo gave some examples in which the costs in one country or the other prevail, but that didn't settle the question. Later, John Stuart Mill suggested that in many cases, exchange values in international trade would be determined by "supply and demand." This insight would be extended by later economists to become the more modern theory of "exchange value," or rather, market price.

Ricardo also recognized, and to some extent explored, the role of machinery in production, which was growing rapidly at that time. This discussion was to lead to a very tricky problem in the labor theory of value, but no-one would know that until later.

Ricardo left the labor theory a logically sounder and far better entrenched theory than it had been. Nevertheless, it appeared that there were three major exceptions to the labor theory:

1. absolutely scarce goods
2. international trade
3. monopoly

1.5 CONTRIBUTION OF KARL MARX TO DISTRIBUTION THEORY

Marx and the Labor Theory

After Ricardo the Classical Political Economists agreed that the relative value of two commodities in exchange would be equal to the relative quantities of labor-time embodied in the two commodities. Prices might not always correspond to values (old masters, for example) but values always corresponded to labor. Put a little differently (but
not everyone would have put it this way) "labor produces all value." But if labor produces all value, how can there be profits or interest?

Karl Marx was many things -- democratic and socialist revolutionary agitator and leader, journalist, philosopher -- and in his role as an economic theorist, he set out to answer that question. Marx had read Ricardo's ideas, and while Ricardo was no socialist, Marx respected Ricardo's scientific approach. And, as we have seen, Ricardo had found an answer to part of the question. According to Ricardo, landowners would obtain rent without contributing any effort, just because of the workings of the competitive market system and the labor values of products. The landowners were beneficiaries of a surplus-value because they had title to relatively productive land. Marx' idea was that all market payments other than wages -- all profits, interest, and rent -- could be explained in terms of surplus value.

Marx expressed the labor theory of value a little differently, and more precisely, than Ricardo. In Marx's terms, the value of a commodity is the **socially necessary** labor time embodied in it. This phrase, "socially necessary," takes care of some minor confusions in the theory:

Suppose John is a carpenter, but he is very clumsy, so it takes him twice as long as other carpenters to build a house. Does that mean his houses are worth twice as much?

No, since there are other carpenters who can build the house in half the time, half the time is the "socially necessary" labor time. The time that John wastes doesn't go into the value of the house he builds because it is not "socially necessary."

Thanks to technical progress and the extended division of labor, goods today can be produced with much less labor time than would have been required in Adam Smith's time. Does that mean that goods were worth more then?

In a sense, yes. But technical progress and extended division of labor have reduced the labor time socially necessary to produce goods and services, so it is to be expected that their labor-time value would be less. To the point: value depends on the circumstances of time and place, on history and human development. It would be closer to the truth to say that values at other times and places just are not relevant to, or comparable to, values here and now.

This concept of value -- that the value of a commodity is the socially necessary labor time embodied in it -- is basic to Marxist thinking.

**Surplus-Value**

So Marx addressed the question: if value is socially necessary labor time, so that labor produces all value, why does the market award incomes to people who do not work? His key insight was:

- In a competitive capitalist economy, all commodities are priced at their values.
- In a competitive capitalist economy, labor is a commodity.
- Therefore, in a competitive capitalist economy, labor is priced at its value.
In other words: the wage paid for a labor-day would be the labor time socially necessary to produce the labor day. Suppose that it takes just half of a labor day to produce a labor day. Then workers will always be available for half a labor-day of pay, and employers, knowing this, will pay no more than half a labor-day of wages per labor-day. Half a labor day is left to the employers. It is "surplus-value" and is the source of profits, interest, and rent. Employers (and landowners and financiers) don't have to do anything to get it -- it is just "left over" after the competitive wage has been paid.

But what does it mean to talk about "producing" a labor-day? Let's put it this way. To attain a certain "standard of living," workers consume a certain amount of goods and services of various kinds. The "cost of production of labor" is the labor cost of producing those goods and services. Clearly, "subsistence" lets a lower limit to the workers' standard of living. Beyond this, what determines the worker's "standard of living?" We shall pass over the controversy surrounding this point.

Since each day of work produces a labor-day of value (under normal conditions) and costs less than a labor-day of value, there is a fraction of a labor-day left over, the surplus-value. Since labor produces all value, but gets only a part of what it produces, surplus-value is exploitation, in the Marxist conception.

Activity 1

1. Discuss the concept of distribution in context of microeconomics.
2. Explain in detail the classical theory of distribution.
3. What do you understand by the neo classical theory of distribution?
4. Distinguish between Ricardian and Marx’s approaches on land, labor and wages.

1.6 SUMMARY

The neoclassical theory of distribution provides an account of how distribution of income between capital and labor are determined in competitive markets at the macroeconomic level over time with technological change and changes in the size of the capital stock and labor force. In the same way the classical approach to distribution was discussed as the first approach of distribution theory which is an idea that free markets can regulate themselves in natural manner according to conditions and situations. Further the Ricardian approach to land, labor and wages was discussed that rent of land would be just enough to offset the differences in labor cost, so that the value of agricultural products would be the same regardless of the fertility of the land where they were produced. That is because rent is based on differential productivity. Finally views of Karl Marx were explained in detail pertaining to distribution factors regarding land and surplus value.
1.7 FURTHER READINGS

UNIT 2

DISTRIBUTION AND RELATED ASPECTS

Objectives

Upon successful completion of this unit, you should be able to:

- Understand the concept of distribution.
- Absorb the theory of marginal productivity in distribution
- Learn the approaches to factor shares, payment and rent.
- Appreciate the concept substitution in context of distribution theory.
- Know the aspects related to technical progress.

Structure

2.1 Introduction
2.2 Marginal productivity theory of distribution
2.3 Factor payment and concept of rent
2.4 Substitution and distribution
2.5 The technical progress
2.6 Summary
2.7 Further readings

2.1 INTRODUCTION

Distribution in economics, is a term applied to two different, but related, processes: (1) the division among the members of society, as individuals, of the national income and wealth; and (2) the apportionment of the value of the output of goods among the factors or agents of production—namely, labor, land, capital, and management. The division or apportionment of this value takes the form of monetary payments, consisting of wages and salaries, rent, interest, and profit. Wages and salaries are paid to workers and managers; rent is paid for the use of land and for certain kinds of physical objects; interest is paid for the use of capital; and profit is realized by the owners of business enterprises as a reward for risk taking.

Recipients of these payments do not receive equal parts of the total. The formulation of the economic laws governing the division of the total of these payments into their various forms and relative portions constitutes the central problem of economic theory in distribution.

Economists have not agreed in formulating these economic laws. Different schools of economists have defined them differently at various times. A large body of authoritative opinion maintains that inequalities in income result, in great part, from the operation of
the law of supply and demand. In this view, for example, an overproduction of cotton will result, through a consequent fall in the price of cotton, in a decrease in the income of cotton growers. It will also tend to result in an increase in the real income, or purchasing power, of the purchasers of cotton, who can buy it more cheaply than would otherwise be possible. Similarly, when capital is abundant and the demand for it is low, interest rates tend to fall. As a result, the relative share of the national income of creditors tends to decrease, while the share of borrowers tends to rise. Variations in the relative share of the national income of workers are also explained in terms of the operation of the law of supply and demand: When labor is plentiful, wage rates tend to fall; when labor is scarce, as in wartime, wages tend to rise. And, finally, inequalities in income among workers are explained by the relative abundance or scarcity of their skills: Skilled workers, less numerous than unskilled workers receive higher wage rates; and workers with rare skills are paid at a higher rate than workers with skills found in abundance.

Economists recognize, however, that the distribution of the national income is influenced by a number of factors in addition to the operation of supply and demand. These factors include the practice by some monopolies and cartels of creating artificial scarcities and fixing prices; collective bargaining by unions and management; and social reform legislation, such as social security and minimum wage and maximum hour laws. Such factors tend to increase the income of one group or another above the level it would reach through the unimpeded operation of the law of supply and demand. Taxation is also an important factor affecting income distribution. Some related concepts to distribution will be discussed in this unit.

2.2 THE MARGINAL PRODUCTIVITY THEORY OF DISTRIBUTION

2.2.1 The Product Exhaustion Theorem

The 1871-4 Marginalist Revolution demolished the Classical Ricardian theory of value. However, the Classical theory of distribution lingered on for a little while. In the 1890s, however, the Neoclassicals finally put forth their own theory -- the "Marginal Productivity" theory of distribution -- that was at the same a generalization and repudiation of the the Classical Ricardian story.

Despite its late appearance, a general marginal productivity theory of factor price determination was already "in the air" by the 1870s. We see hints of it in the work of J.H. von Thünen (1826-50), Mountiford Longfield (1834) and Francis A. Walker (1876). But it was not until the 1890s that the Marginalists realized that the Ricardian "law of rent" applied to all factors and not merely land -- and that a new theory of distribution could be built on that basis. This realization was first forwarded by John Bates Clark (1889, 1891) and John A. Hobson (1891).

Clark and Hobson realized a very simple thing: the Classical had claimed that the supply of labor is endogenous ("men multiply like mice in a barn...", etc.) while the supply of land is fixed and does not vary. But, at least in theory, there is nothing "special" about
the fixity of the supply of land and the variability in the supply of labor. If we hold one factor fixed and vary the other, the rent principle should apply regardless: the quantity of the varying factor should be set where its economic earnings are equal to its marginal product. To see this, suppose that instead of varying the amount of labor applied to a fixed amount of land, why not apply varying amounts of land to a fixed amount of labor? In this case, the horizontal axis in Figure 3 would be $T$, the amount of land, and the curves $MP$ and $AP$ curves would reflect the marginal product and average product of land. Stipulating a fixed payment per unit of land (analogous to the wage), and we would reach the same conclusion: land would be employed up until its payment per unit is equal to its marginal product. Total output would still be by the entire box, with the lightly-shaded box representing payments to land and the remainder (the dark-shaded box) representing payments to the other factors - in this case, labor.

The **Classicals** were not unaware of this possibility. However, they contended that factor shares, computed in this way, would fail to "add up" to total output. To see this, suppose we have a three factor economy, $Y = f(L, K, T)$, where $L$ is labor, $K$ is capital, and $T$ is land. Now, suppose we try calculating the returns on the factor via this procedure. Consequently, varying $L$ and leaving $K$ and $T$ fixed, we end up with the result that payments to labor are $f_1L$ and the residual $Y - f_1L$ will be paid to capital and land, so defining $w$ as the wage, $r$ as the return on capital and $t$ as the rate of payment on land, then:

$$wL = f_1L$$

$$rK + tT = Y - f_1L$$

If instead we vary $K$ and leave $L$ and $T$ fixed, we end up with capital payments being $f_2K$ and the residual $Y - f_2K$ being paid to labor and land.

$$rK = f_2K$$

$$wL + tT = Y - f_2K$$

Finally, if we vary $T$ and leave $K$ and $L$ fixed, then

$$tT = f_1T$$

$$wL + rK = Y - f_1T$$

So far so good. The more enlightened Classical economists would say that, yes, perhaps such a calculation could be made, but that it revealed nothing about the theory of distribution. Now it must be true that (if no entrepreneurial gains are made) $Y = wL + rK + tT$ and this will be the case if *any* of the three cases given above hold. It is a simple accounting principle. Thus, the Classicals were willing to admit, to some extent, that there was nothing "special" about the fixity of land and the flexibility of capital-and-labor.
in principle, that marginal products could be calculated by fixing other things and and varying other things.

But such exercises, the Classicals contended, revealed nothing about distribution because there is no reason to presume that all three cases hold simultaneously. In other words, it is not necessarily true that when calculating factor payments by marginal products, i.e. setting \( w = f_L \), \( r = f_K \) and \( t = f_T \), that the sums of factor payments will add up to total product:

\[
Y = f_L L + f_K K + f_T T
\]

Thus, the Classicals concluded, exercises that try to calculate marginal products of other factors serve no real end as they will not "add up".

But John Bates Clark (1889, 1891) contended that this equality would hold. In other words, he asserted that when each factor is paid its marginal product, the sum of factor incomes will exhaust total output. This proposition became known as the marginal productivity theory of distribution or the product-exhaustion theorem.

Clark himself provided a rather loose verbal "proof" of this contention. Philip H. Wicksteed (1894) was the first to prove it mathematically. However, Wicksteed also revealed that there was a necessary condition for this to hold: namely, the aggregate production function must be linearly homogeneous. Specifically, if the aggregate production function \( Y = f_K, L, T \) is homogeneous of degree one, then if each factor was paid its marginal product, then income shares would indeed "add up". Wicksteed's proof, as A.W. Flux (1894) noted in his review, was a simple application of Euler's Theorem. Namely, by Euler's Theorem, if a function is homogeneous of degree \( r \), then, then:

\[
rY = f_L L + f_K K + f_T T
\]

where \( f_i \) is the first derivative of the function with respect to the \( i \)th argument. Consequently, if, as Wicksteed proposes, the production function is homogeneous of degree one, so \( r = 1 \), then:

\[
Y = f_L L + f_K K + f_T T
\]

But this is precisely the product-exhaustion result we were looking for! So, in sum, the marginal productivity theorem of distribution says that if all factors are paid their marginal products, then the sum of factor incomes will add up to total product.
2.2.2 Early Debates on Marginal Productivity

The marginal productivity theory caused something of a little tornado around the turn-of-the-century, which deserve some attention as they helped clarify what the theory says and what it does not say [accounts of the debates surrounding marginal productivity abound -- those of Joan Robinson (1934), George Stigler (1941: Ch. 12) and John Hicks (1932, 1934) are probably the best. Also worthwhile are the accounts by Henry Schultz (1929), Dennis H. Robertson (1931) and Paul Douglas (1934)].

One of the immediate debates surrounded that of priority. Who "discovered" the marginal productivity theory of distribution? The first verbal exposition of the marginal productivity hypothesis is due to John Bates Clark (1889), which was followed up in Clark (1890, 1899) and, independently, Hobson (1891). Largely unaware of Clark, Philip H. Wicksteed (1894) presented the same theory and proved the product-exhaustion theorem mathematically, although, as noted, it was A.W. Flux (1894) who noted the equivalence between Wicksteed's mathematical statement and Euler's Theorem.

Here a few strange footnotes begin. Enrico Barone had discovered marginal productivity theory independently, but his work was published after he had become aware of Wicksteed's achievement (Barone, 1895, 1896), thus his claim to priority was unluckily lost. Barone convinced Léon Walras to incorporate the marginal productivity theory in the third 1896 edition of his Elements (lesson 36) but then Walras affixed a famous ill-spirited note (App. 3) commending Wicksteed's performance yet claiming that the theorem was already implicit in his early work, and thus that it should be he (Walras), and not Wicksteed, that should be given credit for discovering it. This blatantly opportunistic move outraged even Walras's supporters, and he duly withdrew the note in the fourth edition of his book. During this sorry affair, Knut Wicksell (1900) rose to defend of Wicksteed's claim as discoverer of the theory. But in a surprising twist, it turns out that Wicksell had actually discovered it himself in 1893 -- before Wicksteed -- and had just forgotten about it! To add a bizarre finale, it turns out that a Lausanne mathematician, Hermann Amstein, had basically handed Walras the entire product-exhaustion theorem and its proof in 1877, but Walras had not understood the mathematics and consequently ignored it (cf. Jaffé, 1964).

Despite the fight over priority, the marginal productivity theory of distribution was not immediately embraced by other economists, not even the other Neoclassicals, largely because it was not clear what the theorem said nor what its implications were. As such, it might be useful to clarify a few points of confusion.

The first and most straightforward error (which is sometimes repeated today) is to assume that the marginal productivity theory says that factor prices are determined by marginal products. The tone of the exposition in John Bates Clark (1899) sometimes implies this, and many contemporaries took it at face value. As such, loose critics have gone on to "prove" that the marginal productivity theory is contradictory because it claims that factor prices are determined by marginal products and yet the theory of production tells us quite the opposite, namely that the amount of factors employed (and thus their marginal
products) depend on factor prices. The argument is circular, critics claim, *ergo* the marginal productivity theory is faulty.

The absurdity of this "proof" is evident when one recognizes that the marginal productivity theory does *not* say that marginal products determine factor prices. It has never said that, regardless of whatever Clark let himself say in unguarded moments. Factor prices and factor quantities are determined by the demand and supply of factors, period. The *theory of production*, which makes marginal products dependent on factor prices, gives us only a *factor demand* schedule and *not* the equilibrium position. In other words, *at* equilibrium, factors *are* paid their marginal products because, by *definition*, equilibrium is the equality of demand and supply and, by *derivation* from the theory of the producer, the demand curve is a marginal product schedule. There is thus no "one-way" causality between factor payments and marginal products. At best, as Dennis Robertson (1931) suggests, factor payments are the *measure* of marginal products in equilibrium and consequently, the marginal productivity theory can be regarded as a mere technical *characterization* of that equilibrium. As Alfred Marshall aptly warns us:

"This [marginal productivity] doctrine has sometimes been put forward as a theory of wages. But there is no valid ground for any such pretension. The doctrine that the earnings of a worker tend to be equal to the net product of his work, has by itself no real meaning; since in order to estimate net product, we have to take for granted all the expenses of production of the commodity on which he works, other than his own wages. But though this objection is valid against a claim that it contains a theory of wages; it is not valid against a claim that the doctrine throws into clear light the action of one of the causes that govern wages." (A. Marshall, 1890: p.429-30)

This warning repeated by Gustav Cassel:

"marginal productivity itself is not an objectively ascertained factor in the pricing problem, but is in fact one of the unknowns in the problem...[A factor's] marginal productivity, then, cannot be defined as anything other than [its] price, for this price represents precisely the contribution of the labour in question to the price of the product. The statement that wages are determined by the marginal productivity of labour thus loses all independent meaning." (G. Cassel, 1918: 312-313)

The second misleading element is to take John Bates Clark's argument that marginal productivity is a *natural law* or one that is necessarily "moral" at face value. The following paragraph from his master tome reveals the gist of his claims:

"If each productive function is paid for according to the amount of its product, then each man get what he himself produces. If he works, he gets what he creates by working; if he provides capital, he gets what his capital produces; and if, further, he renders service by coordinating labor and capital, he gets the product that can be separately traced to that function. Only in one of these ways can a man produce anything. If he receives all that he brings into existence through any one of these three functions, he receives all that he creates at all." (J.B. Clark, 1899: p.7)
The tone of Clark’s assertion has led many to think he was referring to the intramarginal worker. This has led critics like George Bernard Shaw to respond in kind:

"[T]hat of giving to every person exactly what he or she has made by his or her labor, seems fair; but when we try to put it into practice we discover, first, that it is quite impossible to find out how much each person has produced." (G.B. Shaw, 1928: p.21)

To see the issues involved, it is best to be clear with an example. Suppose that we have an enterprise which uses one unit of land which can produce ten units of output. Adding a unit of labor, we

Applying successive laborers to a field, we have the following:

<table>
<thead>
<tr>
<th>Qty. of Labor</th>
<th>Total Product</th>
<th>Average Product</th>
<th>Marginal Product</th>
</tr>
</thead>
<tbody>
<tr>
<td>One Laborer</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Two Laborers</td>
<td>18</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>Three Laborers</td>
<td>24</td>
<td>8</td>
<td>6</td>
</tr>
</tbody>
</table>

Let us assume (for the sake of argument, for this is not implied), that the average product represents the actual contribution of the laborer to total output. So, one laborer alone contributes 10 units of output, two laborers contribute 9 each, three laborers contribute 8 each. But, except for the first case, the factors are not paid what they contribute: they are paid the marginal product. Thus, when there are two laborers, each contributing 9, each of them only receives 8 units in wage payments. When there are three, each contributing 8, they each only receive 6 units in wage payments. If laborers are paid their marginal product, we hardly have "moral justice" in this case!

Of course, the perceptive should have noticed immediately that the product exhaustion theorem does not hold in this example as the sum of factor payments is less than the total product. That is because we have not assumed constant returns to scale in this example. Under constant returns to scale, the marginal product will be equal to average product and so, in that case, the payment to a factor in our example will indeed be equal to its contribution and thus Clarkian "moral justice" is achieved. But the lesson should be clear: "moral justice" does not arise merely from paying factors their marginal product; that could be unjust if we do not have constant returns. But if constant returns to scale applies, then paying workers their marginal products may be considered just.

Do we still obtain "moral justice" when we change the assumption about what each laborer contributes to output? Suppose that the contribution of each individual labor in the table above is actually the marginal product, rather than the average product. In other words, suppose the first laborer actually contributes 10, the second laborer contributes 8 and the third laborer contributes 6. By the theory, they all get paid 6. As the first laborer
contributes 10 but only gets paid 6, do we then say that he did not receive what he contributed? Is Clarkian "moral justice" violated?

Not quite. If three laborers, Mr. A, Mr. B and Mr. C, are all of the same type and quality, then there is no meaning to "first", "second" or "third". The "first" laborer contributes 10, but when all three workers are in the field, one cannot determine who exactly was the "first". If A happened to be the first and C the last on the field that morning, then A contributed 10 and C contributed 6. (recall, we are actually assuming that each individual is contributes his marginal product!) If we "reshuffle" the order of entry, so that C enters first and A enters last, then A contributed 6 and C contributed 10. Thus, there is no definite way of identifying who was the "first", "second", etc.

But suppose we can. Suppose that Mr. A is indeed always the "first", in the sense of being the laborer that always contributes 10 and that Mr. C is always the "last" in the sense of being the laborer that always contributes 6, regardless of how we reshuffle the order of worker entry onto the field. But this is equivalent to saying that the three workers are not a "homogeneous" factor class. Each worker forms a distinct "factor class" unto himself. We can identify the marginal product of each of the individual workers (i.e. factor classes) and thus we ought to pay them differently, i.e. Mr. A receiving 10 and Mr. C receiving 6. There is no confusion in this case.

The way in which George Bernard Shaw's remark can make sense, as pointed out by Pareto (1897) and Cassel (1918), is if factors are used in fixed proportions. In this case, it may be impossible to measure the marginal product of a factor type (one can visualize this by attempting to determine the slope of the Leontief isoquant at the corner).

At any rate, we should note that Clark's moral justice argument can be and was seen as an apology for income distribution in capitalist systems -- marginal productivity "is true to the principle on which the right of property rests" (Clark, 1899: p.v). -- and, as a result, attracted the ire of socialists. However, socialist arguments are not, in principle, disabled by this: marginal productivity determines the share of income going to capital, but it is another issue altogether whether capital is privately or publicly owned. One can believe in the marginal productivity theory of distribution and still advocate that all capital ought to be social. Indeed, one of the major features of Soviet planning in the post-Stalin era was precisely the use of the principle of marginal productivity in pricing factors.

The third difficulty lies in the homogeneity assumption. How credible an assumption is this? Wicksteed asserted, rather tentatively, that constant returns are generally true. His somewhat garbled argument was that if we consider every type of every factor as a "unique" and separate factor, then "on this understanding it is of course obvious that a proportional increase in all factors of production will secure a proportional increase in output" (Wicksteed, 1894: p.33). This, of course, is a poor justification for a linear homogeneous production function: if every unit of every factor is unique, the meaning of marginal product becomes confusing (albeit, see the modern restatement of Wicksteed's theory by Makowski and Ostroy, 1992). Thus, he concludes:
"Our law then may be regarded as perfectly general....In this form, it is not a theory of distribution, but an analytical and synthetical law of composition and resolution of industrial factors and products, which would hold equally in Robinson Crusoe's island, an American religious commune, in an Indian village ruled by custom, and in competitive centres of the typical modern industries." (P.H. Wicksteed, 1894: p.42)

Wicksteed was understandably taken to task by his fellow economists for this outlandish statement of universality of application. Vilfredo Pareto (1897, 1901) led the way, chastising Wicksteed and arguing that the assumption of constant returns production was far less applicable than he thought. This was reiterated by Francis Y. Edgeworth (1904, 1911) and Chapman (1906): constant returns to scale evidently ignored the reality of monopoly, indivisibilities of production and thus differing returns to scale which permeate the "real world". Wicksteed's claim was famously derided by Edgeworth:

"This is certainly a remarkable discovery; for the relation between product and factors is to be considered to hold good irrespectively of the play of the market...There is a magnificence in this generalization which recalls the youth of philosophy. Justice is a perfect cube, said the ancient sage; and rational conduct is a homogeneous function, adds the modern savant." (F.Y. Edgeworth, 1904).

Consequently, Phillip Wicksteed, in his later works (e.g. 1910), withdrew this claim of generality.

Léon Walras (1874: Ch. 36) and subsequently Knut Wicksell (1901, 1902) offered a solution to Wicksteed's dilemma: constant returns to scale, they argued, need not be assumed to hold for production functions. However, perfect competition ensures that producers will produce at the point of minimum average cost, i.e. at the point in their production function where there is constant returns to scale. Thus, the possibility of increasing/decreasing returns to scale can be ignored: competition effectively ensures that constant returns will hold in equilibrium. (see our discussion of the theory of the firm).

Note that Wicksteed (1894: p.35) had been aware that monopolistic situations, where entrepreneurs made positive profits, were inconsistent with constant returns and thus that perfect competition was a necessary precondition for his theorem. However, what he had not noticed, and what Walras and Wicksell insisted upon, was that perfect competition implies constant returns to scale in equilibrium.

Léon Walras's argument (Walras, 1874: Ch. 36) was subtle but rigorous. What he sought to demonstrate was the following: firstly, that under perfect competition, firms produce at minimum average cost; secondly, the result that factors of production will be paid their marginal products is implied by the assumption of a cost-minimizing firms. Ergo, Walras concludes, in a perfectly competitive equilibrium, there will be constant returns to scale and thus the marginal productivity theory follows through. Although Walras's argument, as he originally stated it, is a bit confusing (cf. H. Schultz, 1929; J. Hicks, 1934; H. Neisser, 1940), it was made considerably clearer and more accessible by Knut Wicksell.
Wicksell's (1901, 1902) argument ran something as follows. Consider the following scenario: suppose there are no entrepreneurs and that either capitalists or laborers can own and run the enterprise. The question is this: would capitalists enjoy being the owners of the enterprise - in which case they hire labor, pay them their marginal products and then take the residual as their own income - or would they prefer to be employees - in which case, they are hired by labor, receive their marginal products as income, while the laborer-owners take the residual. Under constant returns to scale, it makes no difference: as the marginal productivity theory shows, \( f_K = Y - f_L \), so that the residual the capitalists would get as owners \( (Y - f_L) \) is identical to the payment they would receive as employees \( (f_K) \). However, suppose now that we have diminishing returns to scale so that, by Euler's Theorem:

\[
y > \sum_{i=1}^{m} f_i \cdot x_i
\]

thus factor payments fail to exhaust output (incidentally, this was pointed out by Knut Wicksell (1901: p.128) using a Cobb-Douglas production function). In this case, \( f_K < Y - f_L \), thus the residual income the capitalist gets as an owner is greater than what he would receive as an employee. Conversely, under increasing returns to scale, Euler's Theorem implies that:

\[
y < \sum_{i=1}^{m} f_i \cdot x_i
\]

so if factors are paid their marginal products, then total factor payments will exceed output. Thus, \( f_K > Y - f_L \), i.e. the capitalist would earn more as an employee than he would earn as an employer. The same story can be applied to laborer's decisions on whether to own or be employed in an enterprise. Thus, Wicksell informs us, the general rule is this: under decreasing returns to scale, all factors will prefer to be owners; under increasing returns, all factors will prefer to be employees; and under constant returns, the factors do not care whether they are owners or employees.

Suppose, Wicksell suggests, that we do indeed have decreasing returns to scale and perfect competition. All factors would thus desire to become owners and consequently they would all try to set up their own enterprises and hire each other. This would lead to a bidding up of factor prices that would eat into the profits made by the residual earner. In other words, the extra amount these aspiring entrepreneurs would make as employers would be dissipated by their competition to employ factors. Conversely, suppose we are under increasing returns to scale so that all factors prefer to be employees rather than employers; the bidding process would work in reverse, and factor prices would fall. Thus, Wicksell concludes, the only point consistent with stable competition would be where factors are indifferent between being employees or employers, i.e. the constant returns to scale case. John Bates Clark had recognized effectively the same problem and offered the same solution:

"It is clear that one group cannot keep its profit-making position in the system, if entrepreneurs who are making no such net gains are at liberty to enter it. May not all entrepreneurs be making the same rate of profits, and making them at the same time?"
Clearly not: for it would be an universal invitation to capitalists to become *entrepreneurs* and, as such, to bid against each other for labor and capital till the profit should everywhere vanish, by being made over to laborers and capitalists in the shape of additional wages and interest." (J.B. Clark, 1899: p.290-1).

Notice that these arguments reiterate Walras's old idea (cf. Walras, 1874: p.225-6) that under perfect competition, the entrepreneur makes "no profit" -- which, in this context, means that the owner receives no more in residual income than he would receive as an employed factor of production. However, it was this very definition of "profit" that irked unsympathetic commentators such as Edgeworth (1904).

A fourth objection to the marginal productivity theory was set forth by John A. Hobson (1910, 1911) and Albert Aftalion (1911). They argued that when one re-defines the concept of marginal product in terms of *loss*, the definition which Carl Menger (1871) had used, then if all factors are paid their marginal product, it will not "add up". To see this, consider the example offered even earlier by Friedrich von Wieser (1889: p.83): suppose three units of a factor are employed in the "best" enterprise which, jointly, produces 10 units of output. The alternative use of each of these factors by itself yields 3 units each. Consequently, *removing* a factor means that the remaining two units produce 3 each and thus a total of 6. Consequently, the removal of the third factor has reduced output from 10 to 6. Thus, the *marginal product* of the factor, computed in the Menger-Hobson loss form, is 4. Yet if each factor was indeed paid 4, then, added up, the total payments would be 12, which more than exhausts total product available (10).

As Alfred Marshall (1890: p.339n.1) and, more satisfactorily, F.Y. Edgeworth (1904) and John Bates Clark (1901) pointed out, Hobson's argument relies on the fact that he was using *large* units to compute marginal products. Marginal units are infinitesimal - i.e. \( f_i = \frac{\partial f}{\partial x_i} \) is the marginal product of the ith factor - and if \( f_i \) is defined, then the marginal productivity theorem holds true. Note that Hobson's argument can hold true in the non-differentiable case. Suppose we have a Leontief production function:

\[ Y = \min \{ \frac{K}{v}, \frac{L}{u} \} \]

where \( v \) is the required capital-output ratio and \( v \) the required labor-output ratio. By Hobson's definition, removing a marginal unit of capital will reduce output by \( 1/u \), thus this is the marginal product of capital. Similarly, \( 1/v \) would be Hobson's marginal productivity of labor. Consequently, if factors are paid their marginal products, total factor income is \( K/v + L/v \), which is certainly greater than \( Y \).

But is the marginal productivity theorem always disabled in the non-differentiable case? No. How does the marginal productivity theory of distribution work when the production function is not differentiable, e.g. of the activity analysis sort? This question leads us the fixed coefficients equilibrium models of Léon Walras (1874), Friedrich von Wieser (1889), Vilfredo Pareto (1906), Gustav Cassel (1918) and Abraham Wald (1936), thus we must refer to our review of the *Walras-Cassel model* for the complete answer.
Nonetheless, a few brief remarks may be worthwhile making here (cf. Schultz, 1929; Hicks, 1932, 1934). We must differentiate between two central questions when facing non-smooth technology: firstly, that of the *determinacy* of factor prices and quantities; secondly, of whether one somehow conceive that factor prices will be equal to the "marginal products" of the relevant factors in some manner. The determinacy question is swiftly answered: in the Walras-Cassel model, fixed coefficient production models will yield us determinate factor prices if the following conditions hold: (1) prices of all processes are set equal to their cost of production ("perfect competition"); (2) if there are m factors of production, then there are at least m output processes that employ *all* the m factors.

Condition (1) is familiar: knowing output prices, we can immediately determine factor prices in Walras-Cassel models (even though these may be negative, etc.). However, it does not work unless condition (2) is also imposed. What this means, effectively, is that indeterminacy can arise if the number of processes using the factors is *less* than the number of factors. Thus, the Hobsonesque instance we proposed earlier, a traditional Leontief production function, fails the determinacy condition (2) because it employs two factor (K and L) and only one output process. In terms of Walras-Cassel diagrams, this is equivalent to having only one price-cost equality determining two factor prices: clearly, factor prices would be indeterminate in this case.

So, if factor prices are determinate, are they equal to marginal products? This will be true, if we define the *marginal product* of a factor as the increase in output that arise from the marginal *release* of the relevant factor supply constraint. In other words, the marginal productivity of a factor is merely the shadow value of the factor supply constraint. As the Walras-Cassel-Wald model indicates, this shadow value will be *precisely* the factor prices determined in the primal exercise. Thus, factor prices *are* equal to marginal products.

In sum, the marginal productivity theory of distribution holds good in activity-analysis type of general equilibrium models *in spite* of the fact that we cannot differentiate the production function to derive the marginal product as a derivative. It comes in as a shadow value.

A final point we wish to make regards Alfred Marshall's unique formulation of marginal productivity theory (cf. Marshall, 1890: p.426-30). For Marshall, the marginal product of a factor should be computed *only* once *all other factors* have been adjusted to their optimal values -- what he called the *net marginal product*. In other words, Marshall recognized the marginal product concept can be a bit misleading; the marginal product of labor is the change in output that results from an increased employment of a unit of labor. But what if the resulting factor proportions are *not* optimal for the entrepreneur? What if, as a result of employing a unit of labor, he decides to let go or add on an extra unit of capital as well?

Marshall's proposed the concept of net marginal product to solve the problem (cf. Hicks, 1932: p.12-15; Machlup, 1937). The net marginal product of labor can be defined as the
increase in output that arises from the employment of an extra unit of labor after we allow all other factors to be adjusted to their new optimal (i.e. profit-maximizing) levels. Mathematically, we can conceive of this by employing the profit function, $\pi (p, w, r)$. Let us define the following:

$$\pi (p, w, r) - wL = \max [pY - wL - rK] - wL$$

as net profits, or profits minus labor costs, evaluated at the optimal, profit-maximizing position. Consequently, Marshall's net marginal product of labor can be defined as $\partial \pi / \partial L$, i.e. the product of an extra unit of labor minus the value of inputs cooperating with labor when the levels of these inputs are chosen optimally. Notice that in a position of maximum net profits (as in equilibrium), we have from the first order conditions that $\partial \pi / \partial L = w$, i.e. the net marginal product is equal to the wage. It is in terms of net marginal products that Marshall goes on to formulate his marginal productivity theory.

### 2.3 FACTOR PAYMENTS AND THE CONCEPT OF RENT

The first thing to remember is this: in Neoclassical theory, factor prices and quantities employed are determined simultaneously by the supply and demand for factors. Period. If any of the ensuing discussion seems confusing, one can regain one's bearing by reminding one's self of this. This is the thread out of the labyrinth which follows.

The second thing to remember is this: in all that follows, there are no produced factors of production, i.e. there is no capital. More precisely, for the rest of this discussion, the word "capital" is used in the same sense as "land", i.e. capital is assumed to be an endowed factor of production (which effectively contradicts the definition of capital! -- but more on that later). For an analysis of the theory of distribution with capital properly speaking, turn to our discussion of capital theory.

The reason for these initial warning is that the Neoclassical theory of distribution -- what has become known as the "Marginal Productivity Theory of Distribution -- has been a subject of much debate and confusion since it was formulated in the 1890s. We shall survey this debate here.

Before proceeding, we ought to be clear about a few terms. By "distribution" we mean the relative income received by the owners of factors of production. If L units of labor are employed in the economy, each unit being paid a wage w, then the income of laborers (the owners of labor) is wL. If K units of (fixed, endowed) capital are employed and paid a return r, then the income of capitalists (the owners of capital) is rK. If we denote by Y the economy-wide level of output, then the income share of labor can be expressed as wL/Y and the income share of capital is rK/Y. Consequently, the relative income shares of the capital and labor can be expressed as a ratio wL/rK.

The distribution of income is about how total output in the economy Y, is divided up among people. Edgeworth called it "the species of exchange by which produce is divided between the parties who have contributed to its production " (Edgeworth, 1904). The
laborers get $wL$, the capitalists get $rK$ and, possibly, there might be some residual amount. This residual amount, the amount of income/output produced which is not paid back to the owners of capital and labor for factor services, is $R = Y - rK - wL$. The residual is usually paid out to a the class of people known as entrepreneurs.

It is important not to confuse this "residual" with the "surplus". The "surplus" is defined as the amount of output that is not paid out to factors in reward for "factor services." So, if we define $r$ and $w$ as the rate of return and wage in "reward" for factor services, the surplus is defined as $S = Y - rK - wL$. This seems mathematically similar to the entrepreneurial residual, but it is, in fact, quite different. Explicit in the definition of the surplus is the assumption that $r$ and $w$ are what is called "economic earnings" alone. In contrast, the $r$ and $w$ in the definition of the entrepreneurial residual include both economic earnings and "rental earnings". So, if we define $r_e$ and $w_e$ as the economic earnings of capital and labor and $r_r$ and $w_r$ as their "rental" earnings, then the surplus is:

$$S = r_eK + w_eL$$

while the entrepreneurial residual is:

$$R = (r_e + r_r)K + (w_e + w_r)L$$

So, implicitly, while the residual accrues to the entrepreneur alone, the "surplus" includes amounts that accrue to labor and capital in the form of rental earnings.

This may all seem a bit obscure and so we need to define things a bit better. Just how do we distinguish payments for factor services from payments derived from the surplus, i.e. between economic earnings and rental earnings? The difference differs in meaning between Classical and Neoclassical economists. However, in general, we can define them as follows:

*Economic earnings* are that portion of factor payments by a producer which is necessary and sufficient to employ the particular factor, i.e. to obtain "command" of the factor services.

*Rental earnings* are any payments received by the factor above their economic earnings and as a result of their being in fixed supply.

Turning to economic earnings, what does "necessary and sufficient to employ a factor" mean? For the Classical Ricardian school, the economic earnings of a factor are merely the payments necessary to maintain the factor "intact". Thus, for laborers, economic earnings are wages required to keep the laborer alive and well, i.e. "subsistence" wages. (for capital, the story is more complicated; see our discussion of Classical capital theory).

Although some Neoclassicals have agreed to this Classical definition, most have taken on the Austrian definition of economic earnings in terms of *opportunity costs*. If a producer
wishes to secure the employment of a particular factor, it has to pay that factor at least what it might receive in alternative employments. This is the opportunity cost of the factor. So, if a factor is paid $7 an hour by a particular producer and could find alternative employment only for $5 an hour, then the factor's opportunity cost (and thus its economic earnings) are $5 and its surplus earnings are $2.

We can fix our ideas better by examining the factor market equilibrium for a particular factor. An example for the labor market is shown in Figure 1. The $L^d$ curve is the economy-wide demand for labor by firms, $L^s$ is the economy-wide supply of labor by households. The demand for labor is downward sloping with respect to the wage for reasons that have already been extensively analyzed in the Neoclassical theory of production: specifically, as the wage increases (holding all other factor prices constant), firms will choose techniques of production that substitute away from labor and towards other factors. We know, from profit-maximization, that they will choose to employ labor until the marginal value product is equal to the wage. Thus, heuristically, the labor demand curve $L^d$ can be seen as the economy-wide marginal value product curve (if we can define such a thing as an economy-wide MVP!). The labor supply curve is upward sloping because of labor-leisure choice issues: the greater the wage, the greater the opportunity cost of leisure, and thus the more households substitute away from leisure and towards labor.

![Figure 1 - Factor Market Equilibrium](image)

Factor market equilibrium is established where the economy-wide demand for labor $L^d$ is equal to the economy-wide supply of labor ($L^s$). In Figure 1, this will be at $w^*$, where $L^d = L^s$. Notice that at a lower wage, e.g. $w_1$, there is an excess demand for labor as $L^d = L_3 > L_1 = L^s$. At a higher wage, e.g. $w_3$, we have excess supply of labor as $L^d = L_1 < L_3 = L^s$. 
The wage $w^*$ in Figure 1 is the equilibrium wage. Equilibrium quantity is $L^*$, thus economy-wide labor earnings are, in equilibrium, $w^*L^*$, the area of the box formed by $0L^*ew^*$ in Figure 1. Economic earnings and rental earnings are noted in Figure 1 by the areas $E$ (for economic earnings) and $R$ (for rental earnings).

The reasoning for labelling the $R$ and $E$ areas in this manner can be readily understood. When we are at the factor market equilibrium ($w^*$, $L^*$), every worker is individually paid the equilibrium wage, $w^*$. However, it may be that some of these workers might have been willing to work at a lower wage. They nonetheless receive $w^*$. For instance, notice in Figure 1 that at wage $w_1$, the amount of labor supplied is $L_1$. Thus, the "economic earnings" of the first $L_1$ workers, the payment that would have been sufficient to command their labor, is not more than $w_1$. However, in equilibrium, these same set of workers, $L_1$, are paid the equilibrium wage $w^*$, which is considerably higher than $w_1$. This principle applies to all the "intramarginal" workers supplied between zero and $L^*$. Thus, the area below the labor supply curve reflects economic earnings, while the area above the labor supply curve reflects, as we shall see, rental earnings.

We spoke earlier that economic earnings are payments required to command labor, which, we noted, in the Austrian sense, are conceived as opportunity cost payments. Opportunity cost is captured by the shape of the labor supply curve. In this simplified scenario, we can conceive of the "alternative" employment of the laborers to be "leisure". Thus, the greater the "rewards" of leisure, the lower will be the labor supply at any wage rate (i.e. $L^*$ shifts left) and thus the greater the wage required to command any laborer and thus the higher the economic earnings of any employed laborer.

Note the implications of the two extreme scenarios, both depicted in Figure 2. Suppose leisure is so disliked that, in fact, workers do not consider it a gainful alternative to employment. In this case, the labor supply curve is vertical as shown by $L^s$ in Figure 2. In other words, any wage rate will call forth the entire labor force. Now, equilibrium will still be where the labor demand and vertical labor supply curve meet at $e$, thus we still have a strictly positive equilibrium wage, $w^* > 0$ and strictly positive labor earnings (the area of the box, $w^*L^*$). However, notice that now that all earnings of labor, $w^*L^*$, are rental earnings and economic earnings are nil. Conversely, suppose labor supply is supplied with infinite elasticity, i.e. we have a horizontal labor supply curve such as $L^s'$ in Figure 2. In this case, an infinite amount of labor is supplied when the wage is greater than $w^*$ and no labor is supplied when the wage is below $w^*$. Labor earnings are still defined at equilibrium $e$ as the area of the box $w^*L^*$. However, note that now equilibrium labor income $w^*L^*$ will be entirely composed of economic earnings and no rental earnings are received.
Figure 2 - Pure Rental versus Pure Economic Earnings

The reasoning for these two extreme cases is readily apparent. When the labor supply is inelastic (vertical $L^s$), i.e. when there is no alternative employment, then any earnings made by labor must necessarily arise because firms are fighting over a limited supply of them. In other words, firms are bidding up their wages "artificially" above what is necessary to get them to work. The supply of workers will be $L^*$ regardless of what the wage offered is. As such, workers are experiencing a windfall gain in this case: they would be willing to work for less, much less (indeed, near zero), but competition among firms has bid their wages up regardless.

In contrast, when the labor supply is perfectly elastic (horizontal $L^s'$), then there are no "intramarginal" workers. In other words, at least $w^*$ is necessary to call forth any labor and, furthermore, $w^*$ calls forth an infinite amount of labor. Labor supply is not finite at $w^*$. This implies that, as long as firms pay at least $w^*$, they do not need to "fight" each other over a limited supply of workers. As there is no "bidding war" ensuing from limited labor supply, firms will have no incentive to pay workers above their minimal opportunity cost wage, $w^*$.

These examples permit us to better define the meaning of rental earnings of a factor as that portion of earnings that arise purely out of the fact that the factor is in fixed supply. This concept of rent, or differential rent or Ricardian rent as it has been variously called, was introduced simultaneously but independently by T.R. Malthus (1815), Robert Torrens (1815), Edward West (1815) and David Ricardo (1815), and became one of cornerstones of the Classical Ricardian theory of distribution.

Classical theory generally did not assume that factors were fixed in supply; in other words, they assumed that capital and labor could be "produced". In terms of Figure 2,
they believed the labor supply curve was horizontal so that all payments to labor were economic earnings. However, following David *Ricardo* (1815, 1817) they recognized that *land* was fixed in supply and thus that *land* made rental earnings.

Figure 3 illustrates the Classical Ricardian theory of rent. Here we are assuming only two factors of production: labor (L) and land (T), where the labor is completely variable but land is in fixed supply. The production function is thus $Y = f (L, T_0)$, where $T_0$ is the fixed total supply of land. In contrast labor is supplied with infinite elasticity (a typical Classical assumption). This is captured in Figure 3 by the infinitely-elastic labor supply curve, $L^*$, at the subsistence wage rate, $w$.

![Ricardian (Differential) Rent](image)

**Fig. 3 - Ricardian (Differential) Rent**

The horizontal axis in Figure 3 measures different amounts of labor being applied to the fixed amount of land. The curves $AP_L$ and $MP_L$ are the "economically-relevant" portions of the average product and marginal product of labor curves (i.e. the portions where marginal product of labor is diminishing and below the average product curve, what Ricardo called the portion above the *extensive margin*; see our discussion on marginal products).

The Classical Ricardian story now proceeds as follows. For a given amount of land, the more labor we apply to it, the smaller the marginal and average products. This the Classicals conceived as a natural truth with regard to agriculture *alone*. The basic idea was that land was in fixed supply and of differing quality. The most fertile lands were always used first and the less fertile ones only used later. Thus, the more the scale of production increases (i.e. the more dollops of labor are applied), the increasingly *worse*
land would be taken into cultivation and thus the lower the productivity of labor on the marginal piece of land.

Now, the Ricardians argued that at least enough output must be generated to pay for the factors of production. The wage paid to labor is \( w \) and this reflects economic earnings entirely, i.e. must cover "subsistence". In contrast, land, although a factor of production, does not need to be paid for. One can justify this in Classical terms by saying it does not need to be maintained intact; in neoclassical terms, this implies there are no alternative uses for land and thus no opportunity costs to be compensated.

However, and this is the gist of the Classical Ricardian story, although land makes no economic earnings, we see that because land is fixed in supply, it takes receives rental earnings. In fact, as we shall see, in Figure 3, all surpluses in production resolve itself into rental payments for land. To see this, note that we can increase the scale of production, and thus take in more land and apply more labor, up until the marginal product of labor is equal to the subsistence wage. This will be at \( L^* \) in Figure 3. Thus, total wage payments \( wL^* \) are the area of the light-shaded box in Figure 3.

Now, at \( L^* \), average product is \( y = Y/L^* \). Thus, total output is \( Y = yL^* \), i.e. the area of the box \( 0L^*ay \) (alternatively, we could have represented total output as the sum of the light-shaded box and the triangle \( ewb \) formed under the \( MP_L \) curve; the areas are equivalent). Consequently, the surplus produced, defined as \( Y - wL \), is the darkly shaded box in Figure 3. This is the amount of output that is produced over and above payments to factors. This remainder, the early Classical contended, accrues to landowners, thus it is referred to as rent.

2.4 SUBSTITUTION AND DISTRIBUTION

One of the interesting results from empirical analysis of income distributions is that relative income shares do not tend to vary much over time, i.e. the relative share of capital in income \( rK/Y \) and the relative share of labor in income \( wL/Y \) have been rather constant. This has been noticed at least as far back as John Maynard Keynes (1939) and has been so repeatedly confirmed by empirical studies that Nicholas Kaldor (1961) considered it to be one of the "stylized facts" about the economy.

Now, constancy of relative income shares does not contradict the marginal productivity theory of distribution. But neither is it implied by it. This has led some economists, notably Nicholas Kaldor (1955), to call for the abandonment of the marginal productivity theory of distribution as it does not "explain" this stylized fact. Others, such as Robert Solow (1958), have disputed its empirical validity.

But there are other more plausible hypotheses to explain the constancy of distribution in terms of the marginal productivity theory. One is that the capital-labor ratio \( k = K/L \) has itself been constant over time. This is possible, but not necessarily plausible. After all, at least since the Industrial Revolution in Western countries, we have seen a rather large increase in the amount of capital per laborer over time. Alternatively, Robert Solow
(1957) offered an explanation of constancy in terms of technical progress: although the capital-labor ratio has indeed increased, the reason that the shares have stayed constant can be possibly explained that the capital-effective labor ratio has been constant (where by effective labor, we mean technically-augmented labor). As it is the marginal product of effective labor which determines the wage, then the constancy of distribution can be explained by a constant capital-effective labor ratio.

A more fruitful explanation was proposed by Martin Bronfenbrenner (1960): namely, instead of speculating whether capital-labor ratios are constant or not, why not examine the conditions under which relative income shares might be constant when the capital-labor ratio does change. Bronfenbrenner identified the keystone to be the elasticity of substitution between capital and labor. To see this, consider our canonical production function again, $Y = f(K, L)$. In order to obtain the marginal productivity theory of distribution, we will have to assume it is constant returns to scale. Consequently, we can express it in intensive form as:

$$y = \phi(k)$$

where $y = Y/L$ and $k = K/L$. The intensive production function is depicted in Figure 4.

Figure 4 - Intensive Production Function

So what exactly are the income shares of capital and labor in Figure 4? At a given capital-labor ratio, the slope of the intensive production function is the marginal productivity of capital, i.e. $\phi_k = f_K$. We also know that the marginal product of labor is thus the remainder, $f_L = y - \phi_k k$ (see our earlier discussion).

Consider now the following procedure (originally due to Knut Wicksell (1893) and Joan Robinson (1956)). At a given capital-labor ratio, $k^*$, the tangent line will have slope $f_k$. 
which is the marginal product of capital. Now, this tangent line crosses the vertical axis at point \( w \) in Figure 4. We claim now three things: (1) \( w \) is the real wage; (2) the portion on the vertical axis \( 0w \) are payments to labor per capita \( (w = f_1L/L = f_1) \); (3) the remaining portion on the vertical axis \( wy^* \) are payments to capital per capita \( (wy^* = f_KK/L) \). To see this, note that the following relationships hold in the diagram:

\[
f_K = \phi_k = y^*/sk^* = w/s0 = wy^*/k^*
\]

Consequently, \( f_K/K/L = f_k K^* = wy^* \), i.e. per capita capital income is precisely the portion \( wy^* \) on the vertical axis in Figure 4. As per capita labor income is \( y^* - f_Kk \), then consequently the rest of the vertical axis (portion \( 0w \)) is per capita labor income, \( f_1 \), thus \( w \) represents the real wage.

Relative shares, \( f_KK/Y \) and \( f_1L/Y \) are obtained by dividing the per capita shares by the output-labor ratio, \( y^* \). Specifically, \( f_1/y = f_1L/Y \) and \( f_kK/y = f_KK/Y \). Thus, the relative income shares are obtained by calculating the relative sizes of the portions \( w \) and \( wy^* \) on the vertical axis in Figure 4.

What happens then when the capital-labor ratio changes? If \( k \) increases, \( y \) increases and so will \( w \), the intercept of the tangent line on the vertical axis. But whether and how the relative shares of capital and labor change can be read diagramatically: if \( w \) rises more than \( y^* \) rises, then the relative share of labor has risen relative to the share of capital.

As Martin Bronfenbrenner (1960) has shown, the evolution of relative income shares depends in good part on the elasticity of substitution \( \sigma \). To see this formally, note that this question reduces to finding out whether \( f_KK/Y \) is positive or negative. This can be rewritten as \( f_kk/y \), thus differentiating with respect to \( k \):

\[
\hat{\sigma} (f_k k/y)/k = (yf_{kk}k + yf_k - f_k^2k)/y^2
\]

As \( y^2 \geq 0 \) and \( yf_{kk}k < 0 \), then in order to prove that the relative income share of capital rises with \( k \), we must prove that:

\[
yf_k - f_k^2k > -[yf_{kk}k]
\]

or simply:

\[
-f_k(y-f_kk)/yf_{kk}k > 1
\]

We write this in this manner because consultation with an earlier section ought to reveal that this is exactly the expression for the elasticity of substitution, \( \sigma \), in the constant returns to scale case in intensive form. Thus, for \( f_KK/Y \) to rise (fall) with \( k \), then it must be that \( \sigma > 1 \) (\( \sigma < 1 \)). To see this intuitively, recall that \( \sigma \) is defined as:

\[
\sigma = \hat{\sigma} \ln (L/K)/\hat{\sigma} \ln (f_k/f_1)
\]
We can see how $\sigma$ captures the essence of this problem by trying to plot how relative income shares, $f_{K}/f_{L}$, change in response to a rise in the capital-labor ratio, $k = K/L$. A little bit of math will show that:

$$1/\sigma - 1 = -\partial \ln (f_{K}/f_{L})/\partial \ln (K/L)$$

Various possible cases are depicted in Figure 5. Suppose $\sigma = 1$, then the elasticity of substitution claims that $f_{K}/f_{L}$ and $L/K$ rise proportionally, or, in other words, $f_{K}/f_{L}$ falls in exactly the proportion that $K/L$ rises, thus $f_{K}/f_{L}$ is constant as $K/L$ changes. This is shown by the horizontal line $\sigma = 1$ in Figure 5. Suppose now that $\sigma > 1$, then $f_{K}/f_{L}$ falls less than the amount $K/L$ rises, thus $f_{K}/f_{L}$ rises when $K/L$ rises. This is captured by the upward slope of the curve $\sigma > 1$ in Figure 5. Finally, if $\sigma < 1$, then $f_{K}/f_{L}$ falls more than the amount $K/L$ rises, thus $f_{K}/f_{L}$ falls as $K/L$ rises - thus the downward slope of the $\sigma < 1$ curve in Figure 5.

![Figure 5 - Relative Income Shares and Elasticities of Substitution](image)

The three curves $\sigma < 1$, $\sigma = 1$ and $\sigma > 1$ depicted in Figure 5 assume that we have a constant elasticity of substitution production function, thus $\sigma$ is the same value throughout for any $K/L$. More generally, however, it is common to assume a relationship with varying elasticity of substitution, depicted by the hill-shaped locus $\sigma^{*}$ in Figure 5: up to the capital-labor ratio $k^{*}$, we have $\sigma > 1$, at $k^{*}$, $\sigma = 1$ and after $k^{*}$, $\sigma < 1$. In this case, the relative shares of capital and labor vary differently with $K/L$ depending on whether we are below or above the critical value $k^{*}$. Thus, Bronfenbrenner's answer to Kaldor's "stylized fact" on constant relative income shares is that the elasticity of substitution tends not to stray from $\sigma = 1$, thus relative income shares have remained constant.

A second interesting hypothesis was laid out by David Ricardo (1817: Ch. 31), in his famous chapter "On Machinery" added to the third (1821) edition of the *Principles*. In
effect, Ricardo had argued that the introduction of machinery would have an adverse effect on labor income. In our terms, the question is what is the impact of a rise in K on the absolute shares - and not relative shares - of income, i.e. the impact of K on rK and wL. (the issue of the impact of specific types of technical progress on relative income shares is discussed elsewhere).

We shall concentrate mainly on the impact of a rise in K on rK. The specific question can be posed this way: does the amount of income going to capital increase if more capital is employed? The question can thus be written: when do we have it that \( \frac{\partial (f_{K}K)}{\partial K} > 0? \) Note that this is not self-evident: K increasing increases the numerator, but, as we know, substitution implies that \( f_{K} \) would fall, thus \( f_{K}K \) may rise or fall accordingly. By definition note that:

\[
\partial (f_{K}K)/\partial K = f_{K} + Kf_{KK}
\]

thus if this is to be positive, then it must be that:

\[-f_{KK}(K/f_{K}) < 1\]

as we are assuming that \( f_{K} > 0 \) and \( f_{KK} < 0 \). But notice that this is the inverse of the elasticity of the marginal product (i.e. capital demand) curve, i.e. \( -f_{KK}(K/f_{K}) = \varepsilon^{-1} \) where \( \varepsilon = -(\partial K/\partial f_{K})(f_{K}/K) \). Thus, if the demand for capital is elastic (i.e. \( \varepsilon > 1 \)), then the absolute income going to capital will rise with K; if demand for capital is inelastic (i.e. \( \varepsilon < 1 \)), then the absolute income going to capital will fall with K. This result, of course, is quite intuitive: elastic and thus relatively flat demand curve for capital means that a rise in K will reduce \( f_{K} \) very slightly, thus the income going to capital is bound to increase quite a bit; in contrast, an inelastic (and thus quite steep) demand curve for capital means that even a small rise in K will reduce \( f_{K} \) by a lot, thus capital income \( f_{K}K \) is bound to fall.

It is easy to show that we can reduce this to a statement in terms of the elasticity of substitution. Recall that as \( f(\cdot) \) is homogeneous of degree one, then marginal products \( f_{K}(\cdot) \) are homogeneous of degree zero, i.e. \( f_{KK}K + f_{KL}L = 0 \). Thus, \( f_{KK}K = -Lf_{KL} \), so substituting into our earlier expression, we have:

\[
Lf_{KL}/f_{K} < 1
\]

as the necessary condition for a rise in K to increase \( f_{K}K \). Now it is common to assume (and this is not implied) that \( f_{KL} > 0 \), thus we can rewrite this condition as:

\[
f_{K}/f_{KL} > L
\]

or, multiplying by \( f_{L}/Y \):

\[
f_{L}f_{K}/f_{KL}Y > f_{L}L/Y
\]
But again, by our earlier section we should recognize the term on the left as alternative form of the elasticity of substitution, $\sigma$, for the constant returns to scale case. Thus, in order for a rise in $K$ to increase $f_{K}$ we need:

$$\sigma > f_{L}L/Y$$

i.e. the elasticity of substitution must be greater than the relative share of labor. Now, $f_{L}L/Y < 1$, so this is not enough to tell us much. But one thing should be recognized: it is entirely possible that $f_{L}L/Y < \sigma < 1$. This means that it is possible that, when capital increases, that we simultaneously have it that the *relative* share of capital falls (as $\sigma < 1$) while the *absolute* share of capital rises ($f_{L}L/Y < \sigma$).

Now we turn to the Ricardo question: does the amount of labor income rise or fall when machinery ($K$) is increased? This is read directly: $\partial f_{L}/\partial K = f_{KL}L$. If $f_{KL} > 0$, as has been assumed so far, then the absolute share of labor rises. Note the peculiarity in this result: whether labor income rises or falls when capital increases does *not* depend at all on what happens to the absolute share of capital, nor indeed on labor's own relative share. The amount of income received by a factor *always* rises when the quantity of the other factor increases. However, we must remind ourselves that this is rests on the $f_{KL} > 0$ assumption and is *not* implied by CRS or anything else.

### 2.5 THE TECHNICAL PROGRESS

By technical progress one understands an improvement of the technical initial position of a national economy or the whole of all technical innovations of a culture. Either a same output (output) with a smaller employment at work or means of production (inputs) can be provided by technical progress or a higher quantity with the same employment of means of production and work. Apart from the quantitative improvement of the input output relationship there are also qualitative improvements like new products (technology history). Technical progress has among other things cultural and social effects.

**Historical considerations**

In earlier times of mankind history the speed of the technical progress was very slowly, even if it likewise came in larger time intervals to large circulations, for instance the Neolithic revolution.

Historically it quite gave times with technical backward step apart from times with technical progress also. As classical example the fall of the antique culture with the following Middle Ages is considered. However the historical scientists in this question argue, to what extent for example in certain ranges (spreading of the technical progress continued to go also during the middle Ages.)
In the most recent time the question often arose whether technical progress creates jobs or on the contrary a cause for unemployment was. This question dipped already 1821 with David Ricardo and later in the discussion around automation and rationalization.

However some acceptance is made. A sinking active volume is used for the production of the rising goods and service offer, because the purchasing power and/or the purchase desires rises more slowly than the productivity, which is expressed in a rising rate of saving. In addition a free market or capitalistic society is presupposed, at which the employees have to do a certain work time that sinking active volume is not thus converted in shorter hours. On this assumption a pessimism problem develops. Under these conditions it looks in such a way, as if technical progress must lead to the fact that the number of jobs must decrease.

Indeed it can happen on the assumption from all day busy employees that manpower are replaced by machines in the course of technical progress. The classical economist Ricardo admitted that technical progress can destroy jobs. Sometimes similar views are subordinated to Karl Marx (law of the tendentious case of the profit rate), since he expected always far rising employment at means of production a for each job in production as a condition for technical progress. If a machine replaces ten workers in the pin factory, then would have ten workers will dismiss - if not the demand for pins to reconciliation rises accordingly or is worked accordingly more briefly. Thus around 1900 in Germany in the year approx. 3000 h, today is one worked on the average approx. 1400 h.

Such vague considerations tried philosophers to represent such as Karl Popper or the growth theory of the economic science more systematically (see lower sections).

**Manifestations**

Technical progress can take place evolutionary or in a revolutionary manner.

The three main manifestations of the technical progress are:

1. Automation
2. Rationalization and those
3. Synergies, positive scale effects

It concerns with technical progress however not only around the increase of the productivity that about a certain number of humans can manufacture ever more cars, but also qualitative changes, innovations, innovations with the produced products for the consumption of humans.
Evaluation

Whether the technical progress of the society is useful or not and/or whether this is positively or taken up negatively, depends on how the policy carries ensuring briskly that the won economic or technical clearance comes to the entire society.

If one relies only on the pure market forces, it can happen that workers are set free, of progress nothing to then abbekommen. In the public opinion such an unfavorable picture over automation can develop because of unemployment threatening thereby, although thereby work can be settled perhaps even better and more rapidly, one thinks e.g. of ticket automats in the comparison to the old ticket salesman.

In the nature the criticism is not really directed to technical progress against this, when many more against the increasing poverty of humans, after which it job-destruction-unites states that the job destruction would be the result of the technical progress. Against the fact history shows that technical progress is connected with shorter hours and rising standard of living (of 2000 h/Jahr around 1960 on 1340 h/Jahr around 2004). The insufficient of the work time destroys jobs and lowers production - see Okun law. The robot is an example of the technical progress and thus not debt at unemployment. But nevertheless of many enterprises or enterprises states it would be debt at the job destruction.

The technical progress is endangered also today by the fact that the clearing-up of the hardly takes place. To their arguments can be worked against, if by saving of human work and the replacement of work by machines etc. - rationalization investments - the unearned incomes of the entrepreneurs or owners of fortune are not only in accordance with-honoured, but over the income distribution also all different.

It is advisable to ensure that the saved work time e.g. benefits also against the interests of the enterprises over shorter hours with full wage adjustment again the employees. Also the profits could be taxed to the existence safety device of the broad population. This should prevent the envy debate. Otherwise the danger exists the fact that the machine storming he mentality celebrates merry and also further over subsidies artificially work remains, which became technically actually redundant. The labor-saving effect of the technical progress comes into conflict with Biblical ark types la "who does not work, is not also to eat", which are however still a substantial ideology component of our society.

This ideology is one of the barriers, which can brake the technical progress. It was lasting for thousands of years primarily humans, the mainly produced and technical progress with saving of human work by machines, tools or robots was rare or took place in very slow processes. From Roman realm for example is history well-known, according to which an inventor of shatterproof glass, after it had told unklugerweise to the emperor that except it nobody knew the secret of the production, was executed, because the emperor was afraid to large disturbances in the economic life by this new invention. "Gold would be only worth shit", as it is called drastic in the guest meal of the Trimalchio.
As children of the are today occasionally criticized: 1. over entitled concerns going out fears over everything that concerns nuclear power, in favor of an expansion of human occupation in the alternative Jobs must create the password generally ““we”” detached by the actual original purpose by work to create i.e. more to place a better supply and the infrastructure ready for it.

Also the experienced complaint in addition, it is to be saved badly humans, only for growth, is not conclusive. Because growth does not mean more products and services only for humans and less or for something else than humans (times apart from the armament). Altogether by the technical progress ever more is taken humans given than. Nobody produces simply more, only for the own stocks.

The technical progress actually does not work simply negatively, negatively is only the delay of the policy to also give the winning the apparent losers of the progress. The economy does not have to do the task this. For their requests is appropriate only in the profit.

Briefly: Who and the money wants to work, is not only debt, even if he gets on a long-term basis only work and not money. The industry is not carrier of the connection. The nature from work did not turn around what must be done, as expiration in a production process and production processes for more work, which needs for its part none.

**Activity 2**

1. Discuss in detail the marginal productivity theory of distribution.
2. What do you understand by economic earnings and rental earnings?
3. Explain the approach to factor payment and rent.
4. Give a brief note on substitution and distribution.
5. Evaluate the progress of technical progress in context of a developing nation like India.

**2.6 SUMMARY**

In economics, the systematic attempt to account for the sharing of the national income among the owners of the factors of production—land, labor, and capital. Traditionally, economists have studied how the costs of these factors and the size of their return—rent, wages, and profits—are fixed. Followed by this, the marginal productivity theory of distribution was discussed. Further, concepts related to factor payment and rents were explained. Substitution and distribution approaches were given in detail. Finally the technical progress in context of overall economy was explained in detail.
2.7 FURTHER READINGS

- Milton Friedman and Simon Kuznets (1945). Income from Independent Professional Practice
Objectives

After completing this unit, you should be able to:

- Understand the approach of welfare economics
- Become aware of the two mainstream approaches to welfare economics
- Understand the concepts of income distribution and social welfare maximization
- Simplify the 7 equations model to social welfare function.
- Determine the Pareto’s welfare economics
- Know the relevance of welfare economics in other subjects

Structure

3.1 Introduction
3.2 Two approaches
3.3 Income distribution
3.4 Simplified 7 equation model
3.5 Efficiency between production and consumption
3.6 Social welfare maximization
3.7 Pareto’s welfare economics
3.8 Fundamental Theorems
3.9 Welfare economics in relation to other subjects
3.10 Summary
3.11 Further readings

3.1 INTRODUCTION

Welfare economics is a branch of economics that uses microeconomic techniques to simultaneously determine allocative efficiency within an economy and the income distribution associated with it. It analyzes social welfare, however measured, in terms of economic activities of the individuals that comprise the theoretical society considered. As such, individuals, with associated economic activities, are the basic units for aggregating to social welfare, whether of a group, a community, or a society, and there is no "social welfare" apart from the "welfare" associated with its individual units.

Welfare economics typically takes individual preferences as given and stipulates a welfare improvement in Pareto efficiency terms from social state A to social state B if at least one person prefers B and no one else opposes it. There is no requirement of a unique
quantitative measure of the welfare improvement implied by this. Another aspect of welfare treats income/goods distribution, including equality, as a further dimension of welfare.

Social welfare refers to the overall welfare of society. With sufficiently strong assumptions, it can be specified as the summation of the welfare of all the individuals in the society. Welfare may be measured either cardinally in terms of "utils" or dollars, or measured ordinally in terms of Pareto efficiency. The cardinal method in "utils" is seldom used in pure theory today because of aggregation problems that make the meaning of the method doubtful, except on widely challenged underlying assumptions. In applied welfare economics, such as in cost-benefit analysis, money-value estimates are often used, particularly where income-distribution effects are factored into the analysis or seem unlikely to undercut the analysis.

Since the early 1980s economists have been interested in a number of new approaches and issues in welfare economics. The capabilities approach to welfare argues that what people are free to do or be should also be included in welfare assessments and the approach has been particularly influential in development policy circles where the emphasis on multi-dimensionality and freedom has shaped the evolution of the Human Development Index.

Economists have also been interested in using life satisfaction to measure what Kahneman and colleagues call experienced utility.

What follows, for the most part, therefore refers to a particular approach to welfare economics, possibly best referred to as 'neo-classical' or 'traditional' welfare economics.

Other classifying terms or problems in welfare economics include externalities, equity, justice, inequality, and altruism.

### 3.2 TWO APPROACHES

There are two mainstream approaches to welfare economics: the early Neoclassical approach and the New welfare economics approach.

The early Neoclassical approach was developed by Edgeworth, Sidgwick, Marshall, and Pigou. It assumes that:

- Utility is cardinal, that is, scale-measurable by observation or judgment.
- Preference is exogenously given and stable.
- Additional consumption provides smaller and smaller increases in utility (diminishing marginal utility).
- All individuals have interpersonally comparable utility functions (an assumption that Edgeworth avoided in his Mathematical 'Psychics).
With these assumptions, it is possible to construct a social welfare function simply by summing all the individual utility functions.

The New Welfare Economics approach is based on the work of Pareto, Hicks, and Kaldor. It explicitly recognizes the differences between the efficiency part of the discipline and the distribution part and treats them differently. Questions of efficiency are assessed with criteria such as Pareto efficiency and the Kaldor-Hicks compensation tests, while questions of income distribution are covered in social welfare function specification. Further, efficiency dispenses with cardinal measures of utility: ordinal utility, which merely ranks commodity bundles, such as represented by an indifference-curve map is adequate for this analysis.

**EFFICIENCY**

Situations are considered to have distributive efficiency when goods are distributed to the people who can gain the most utility from them.

Many economists use Pareto efficiency as their efficiency goal. According to this measure of social welfare, a situation is optimal only if no individuals can be made better off without making someone else worse off.

This ideal state of affairs can only come about if four criteria are met:

- The marginal rates of substitution in consumption are identical for all consumers. This occurs when no consumer can be made better off without making others worse off.
- The marginal rate of transformation in production is identical for all products. This occurs when it is impossible to increase the production of any good without reducing the production of other goods.
- The marginal resource cost is equal to the marginal revenue product for all production processes. This takes place when marginal physical product of a factor must be the same for all firms producing a good.
- The marginal rates of substitution in consumption are equal to the marginal rates of transformation in production, such as where production processes must match consumer wants.

There are a number of conditions that, most economists agree, may lead to inefficiency. They include:

- Imperfect market structures, such as a monopoly, monopsony, oligopoly, oligopsony, and monopolistic competition.
- Factor allocation inefficiencies in production theory basics.
- Market failures and externalities; there is also social cost.
- Imperfect Price discrimination and price skimming.
- Asymmetric information, principal-agent problems.
- Long run declining average costs in a natural monopoly.
- Certain types of taxes and tariffs.

To determine whether an activity is moving the economy towards Pareto efficiency, two compensation tests have been developed. Any change usually makes some people better off while making others worse off, so these tests ask what would happen if the winners were to compensate the losers. Using the *Kaldor criterion*, an activity will contribute to Pareto optimality if the maximum amount the gainers are prepared to pay is greater than the minimum amount that the losers are prepared to accept. Under the *Hicks criterion*, an activity will contribute to Pareto optimality if the maximum amount the losers are prepared to offer to the gainers in order to prevent the change is less than the minimum amount the gainers are prepared to accept as a bribe to forgo the change. The Hicks compensation test is from the losers' point of view, while the Kaldor compensation test is from the gainers' point of view. If both conditions are satisfied, both gainers and losers will agree that the proposed activity will move the economy toward Pareto optimality. This is referred to as *Kaldor-Hicks efficiency* or the Scitovsky criterion.

### 3.3 INCOME DISTRIBUTION

There are many combinations of consumer utility, production mixes, and factor input combinations consistent with efficiency. In fact, there are an infinity of consumer and production equilibria that yield Pareto optimal results. There are as many optima as there are points on the aggregate production possibilities frontier. Hence, Pareto efficiency is a necessary, but not a sufficient condition for social welfare. Each Pareto optimum corresponds to a different income distribution in the economy. Some may involve great inequalities of income. So how do we decide which Pareto optimum is most desirable? This decision is made, either tacitly or overtly, when we specify the *social welfare function*. This function embodies value judgements about interpersonal utility. The social welfare function is a way of mathematically stating the relative importance of the individuals that comprise society.

A utilitarian welfare function (also called a *Benthamite* welfare function) sums the utility of each individual in order to obtain society's overall welfare. All people are treated the same, regardless of their initial level of utility. One extra unit of utility for a starving person is not seen to be of any greater value than an extra unit of utility for a millionaire. At the other extreme is the Max-Min, or Rawlsian *John Rawls* utility function (Stiglitz, 2000, p102). According to the Max-Min criterion, welfare is maximized when the utility of those society members that have the least is the greatest. No economic activity will increase social welfare unless it improves the position of the society member that is the worst off. Most economists specify social welfare functions that are intermediate between these two extremes.

The social welfare function is typically translated into social *indifference curves* so that they can be used in the same graphic space as the other functions that they interact with. A utilitarian social indifference curve is linear and downward sloping to the right. The Max-Min social indifference curve takes the shape of two straight lines joined so as they
form a 90 degree angle. A social indifference curve drawn from an intermediate social welfare function is a curve that slopes downward to the right

![Figure 1](image)

The intermediate form of social indifference curve can be interpreted as showing that as inequality increases, a larger improvement in the utility of relatively rich individuals is needed to compensate for the loss in utility of relatively poor individuals.

A crude social welfare function can be constructed by measuring the subjective dollar value of goods and services distributed to participants in the economy.

### 3.4 A SIMPLIFIED SEVEN-EQUATION MODEL

The basic welfare economics problem is to find the theoretical maximum of a social welfare function, subject to various constraints such as the state of technology in production, available natural resources, national infrastructure, and behavioural constraints such as consumer utility maximization and producer profit maximization. In the simplest possible economy this can be done by simultaneously solving seven equations. This simple economy would have only two consumers (consumer 1 and consumer 2), only two products (product X and product Y), and only two factors of production going into these products (labour (L) and capital (K)). The model can be stated as:

maximize social welfare: \( W = f(U_1, U_2) \) subject to the following set of constraints:
- \( K = K^x + K^y \) (The amount of capital used in the production of goods X and Y)
- \( L = L^x + L^y \) (The amount of labour used in the production of goods X and Y)
- \( X = X(K^x L^x) \) (The production function for product X)
- \( Y = Y(K^y L^y) \) (The production function for product Y)
- \( U_1 = U_1(X^1 Y^1) \) (The preferences of consumer 1)
\[ U^2 = U^2(X^2 Y^2) \] (The preferences of consumer 2)

The solution to this problem yields a Pareto optimum. In a more realistic example of millions of consumers, millions of products, and several factors of production, the math gets more complicated.

Also, finding a solution to an abstract function does not directly yield a policy recommendation! In other words, solving an equation does not solve social problems. However, a model like the one above can be viewed as an argument that solving a social problem (like achieving a Pareto-optimal distribution of wealth) is at least theoretically possible.

### 3.5 EFFICIENCY BETWEEN PRODUCTION AND CONSUMPTION

The relation between production and consumption in a simple seven equation model (2x2x2 model) can be shown graphically. In the diagram below, the aggregate production possibility frontier, labeled PQ shows all the points of efficiency in the production of goods X and Y. If the economy produces the mix of good X and Y shown at point A, then the marginal rate of transformation (MRT), X for Y, is equal to 2.

![Figure 2](image)

Point A defines the boundaries of an Edgeworth box diagram of consumption. That is, the same mix of products that are produced at point A, can be consumed by the two consumers in this simple economy. The consumers' relative preferences are shown by the indifference curves inside the Edgeworth box. At point B the marginal rate of substitution (MRS) is equal to 2, while at point C the marginal rate of substitution is equal to 3. Only
at point B is consumption in balance with production (MRS=MRT). The curve 0BCA (often called the contract curve) inside the Edgeworth box defines the locus of points of efficiency in consumption (MRS1=MRS ²). As we move along the curve, we are changing the mix of goods X and Y that individuals 1 and 2 choose to consume. The utility data associated with each point on this curve can be used to create utility functions.

3.6 SOCIAL WELFARE MAXIMIZATION

Utility functions can be derived from the points on a contract curve. Numerous utility functions can be derived, one for each point on the production possibility frontier (PQ in the diagram above). A social utility frontier (also called a grand utility frontier) can be obtained from the outer envelope of all these utility functions. Each point on a social utility frontier represents an efficient allocation of an economy's resources; that is, it is a Pareto optimum in factor allocation, in production, in consumption, and in the interaction of production and consumption (supply and demand). In the diagram below, the curve MN is a social utility frontier. Point D corresponds with point B from the earlier diagram. Point D is on the social utility frontier because the marginal rate of substitution at point B is equal to the marginal rate of transformation at point A. Point E corresponds with point C in the previous diagram, and lies inside the social utility frontier (indicating inefficiency) because the MRS at point C is not equal to the MRT at point A.

![Figure 3](image)

Figure 3

Although all the points on the grand social utility frontier are Pareto efficient, only one point identifies where social welfare is maximized. This is point Z where the social utility frontier MN is tangent to the highest possible social indifference curve labelled SI.
3.7 PARETIAN WELFARE ECONOMICS

Paretian welfare economics rests on the assumed value judgment that, if a particular change in the economy leaves at least one individual better off and no individual worse off, social welfare may be said to have increased. (One individual being better off than other individuals and not leaving other individuals worse off is possible in societies, where political power is not related to economic power.) In this sense, an individualistic approach to social welfare is defined, with concern extending to all individuals in society, and with an explicit rejection of any ‘organic’ concept of the State.

CRITICISMS

Some, such as economists in the tradition of the Austrian School, doubt whether a cardinal utility function, or cardinal social welfare function, is of any value. The reason given is that it is difficult to aggregate the utilities of various people that have differing marginal utility of money, such as the wealthy and the poor.

Also, the economists of the Austrian School question the relevance of pareto optimal allocation considering situations where the framework of means and ends is not perfectly known, since neoclassical theory always assumes that the ends-means framework is perfectly defined.

Some even question the value of ordinal utility functions. They have proposed other means of measuring well-being as an alternative to price indices, “willingness to pay” functions, and other price oriented measures. These price based measures are seen as promoting consumerism and productivism by many. It should be noted that it is possible to do welfare economics without the use of prices, however this is not always done.

Value assumptions explicit in the social welfare function used and implicit in the efficiency criterion chosen make welfare economics a highly normative and subjective field. This can make it controversial.

3.8 FUNDAMENTAL THEOREMS

There are two fundamental theorems of welfare economics. The first states that any competitive equilibrium or Walrasian equilibrium leads to a Pareto efficient allocation of resources. The second states the converse, that any efficient allocation can be sustainable by a competitive equilibrium. Despite the apparent symmetry of the two theorems, in fact the first theorem is much more general than the second, requiring far weaker assumptions.

The first theorem is often taken to be an analytical confirmation of Adam Smith’s "invisible hand" hypothesis, namely that competitive markets tend toward the efficient allocation of resources. The theorem supports a case for non-intervention in ideal conditions: let the markets do the work and the outcome will be Pareto efficient. However, Pareto efficiency is not necessarily the same thing as desirability or even more
general definitions of "efficiency"; it merely indicates that no one can be made better off without someone being made worse off. There can be many possible Pareto efficient allocations of resources and not all of them may be equally desirable by society.

These ideal conditions, however, collectively known as Perfect Competition, do not exist in the real world. The Greenwald-Stiglitz Theorem, for example, states that in the presence of either imperfect information, or incomplete markets, markets are not Pareto efficient. Thus, in most real world economies, the degree of these variations from ideal conditions must factor into policy choices.[1]

The second theorem states that out of the infinity of all possible Pareto efficient outcomes one can achieve any particular one by enacting a lump-sum wealth redistribution and then letting the market take over. This appears to make the case that intervention has a legitimate place in policy – redistributions can allow us to select from among all efficient outcomes for one that has other desired features, such as distributional equity. However, it is unclear how any real-world government might enact such redistributions. Lump-sum transfers are difficult to enforce and virtually never used, and proportional taxes may have large distortionary effects on the economy since taxes change the relative remunerations of the factors of production, distorting the structure of production. Additionally, the government would need to have perfect knowledge of consumers' preferences and firms' production functions (which are in fact unknowable[2]) in order to choose the transfers correctly. In addition, this remedy cannot be expected to work if large numbers of people do not understand the economy, and how to make effective use of any transfers they receive.

### 3.8.1 THE FIRST FUNDAMENTAL THEOREM OF WELFARE ECONOMICS

The first fundamental theorem of welfare economics is often misunderstood, especially by technical economists. Briefly, the theorem says that a market outcome is efficient (Pareto-optimal). The theorem, as proven with great mathematical beauty by Arrow and Debreu, requires a number of reasonably strong assumptions such as very large numbers of buyers and sellers who have perfect rationality and perfect information. Since the conditions required for the theorem's proof are unlikely to hold in the real world it's common for people to reverse the theorem to suggest that markets cannot be efficient.

The First Fundamental Theorem of Welfare Economics is proof, in view of its long list of prerequisites, that market outcome can be improved by well-designed interventions.

Now what is wrong with this is very simple. The First Theorem gives sufficient conditions for a market to be efficient it does not give necessary conditions. Thus, as a matter of logic, the fact that the theorem's conditions are not satisfied does not prove that market outcomes can be improved, even by "well-designed" interventions. As an empirical matter, the difference between the sufficient and necessary conditions turns out to be quite large. We know from Vernon Smith's work, for example, that markets can be competitive with only a handful of traders; nor do the traders have to be perfectly rational. In fact, markets can be very efficient with zero-intelligence traders.
Perhaps even more importantly technical economists seem to think that the First Theorem is the ultimate expression of "the invisible hand" or what makes markets good but in fact the First Theorem is but a pinched and limited expression of the virtues of markets. The First Theorem, for example, says nothing about innovation, experimentation, or the discovery process. Nor does the First Theorem say anything about markets and political philosophy. You will not learn from the First Theorem that markets are not simply a "mechanism," markets are peaceful exchange.

To be clear, I am correcting a misuse of the First Theorem. I am not asserting that markets are always perfectly efficient. But really what kind of standard is perfect efficiency anyway? The internal combustion engine isn't even close to being perfectly efficient but my car gets me to work every day, is fun to drive and gives me the freedom of the open road.

3.8.2 SECOND FUNDAMENTAL THEOREM OF WELFARE ECONOMICS

While every equilibrium is efficient, it is clearly not true that every efficient allocation of resources will be an equilibrium. However, the Second Theorem states that every efficient allocation can be supported by some set of prices. In other words all that is required to reach a particular outcome is a redistribution of initial endowments of the agents after which the market can be left alone to do its work. This suggests that the issues of efficiency and equity can be separated and need not involve a trade off.

However, the conditions for the Second Theorem are stronger than those for the First, as now we need consumers' preferences to be convex (convexity roughly corresponds to the idea of diminishing rates of marginal substitution, or to preferences where "averages are better than extrema").

PROOF OF THE SECOND FUNDAMENTAL THEOREM

The second fundamental theorem of welfare economics states that, under the assumptions that every production set $Y_j$ is convex and every preference relation $\succeq_i$ is convex and locally nonsatiated, any desired Pareto-efficient allocation can be supported as a price quasi-equilibrium with transfers. Further assumptions are needed to prove this statement for price equilibriums with transfers. We will proceed in two steps: first we prove that any Pareto-efficient allocation can be supported as a price quasi-equilibrium with transfers, then we give conditions under which a price quasi-equilibrium is also a price equilibrium.

Let us define a price quasi-equilibrium with transfers as an allocation $(x^*, y^*)$, a price vector $p$, and a vector of wealth levels $w$ (achieved by lump-sum transfers) with $\sum_i w_i = p \cdot \omega + \sum_j p \cdot y_j^*$ (where $\omega$ is the aggregate endowment of goods and $y_j^*$ is the production of firm $j$) such that:

i. $p \cdot y_j \leq p \cdot y^*_j$ for all $y_j \in Y_j$ (firms maximize profit by producing $y^*_j$)
ii. For all $i$, if $x_i \succeq_i x_i^*$ then $p \cdot x_i \geq w_i$(if $x_i$ is strictly preferred to $x_i^*$ then it cannot cost less than $x_i^*$)

iii. $\sum_i x_i^* = \omega + \sum_j y_j^*$ (budget constraint satisfied)

The only difference between this definition and the standard definition of a price equilibrium with transfers is in statement (ii). The inequality is weak here ($p \cdot x_i \geq w_i$) making it a price quasi-equilibrium. Later we will strengthen this to make a price equilibrium.

Define $V_i$ to be the set of all consumption bundles strictly preferred to $x_i^*$ by consumer $i$, and let $V$ be the sum of all $V_i$. $V_i$ is convex due to the convexity of the preference relation $\succeq_i$. $V$ is convex because every $V_i$ is convex. Similarly $Y + \{\omega\}$, the union of all production sets $Y_i$ plus the aggregate endowment, is convex because every $Y_i$ is convex.

We also know that the intersection of $V$ and $Y + \{\omega\}$ must be empty, because if it were not it would imply there existed a bundle that is strictly preferred to $(x^*, y^*)$ by everyone and is also affordable. This is ruled out by the Pareto-optimality of $(x^*, y^*)$.

These two convex, non-intersecting sets allow us to apply the separating hyperplane theorem. This theorem states that there exists a price vector $p \neq 0$ and a number $r$ such that $p \cdot z \geq r$ for every $z \in V$ and $p \cdot z \leq r$ for every $z \in Y + \{\omega\}$. In other words, there exists a price vector that defines a hyperplane that perfectly separates the two convex sets.

Next we argue that if $x_i \succeq_i x_i^*$ for all $i$ then $p \cdot (\sum_i x_i) \geq r$. This is due to local nonsatiation: there must be a bundle $x_i'$ arbitrarily close to $x_i$ that is strictly preferred to $x_i^*$ and hence part of $V_i$, so $p \cdot (\sum_i x_i') \geq r$. Taking the limit as $x_i' \rightarrow x_i$ does not change the weak inequality, so $p \cdot (\sum_i x_i) \geq r$ as well. In other words, $x_i$ is in the closure of $V$.

Using this relation we see that for $x_i^*$ itself $p \cdot (\sum_i x_i^*) \geq r$. We also know that $\sum_i x_i^* \in Y + \{\omega\}$, so $p \cdot (\sum_i x_i^*) \leq r$ as well. Combining these we find that $p \cdot (\sum_i x_i^*) = r$. We can use this equation to show that $(x^*, y^*, p)$ fits the definition of a price quasi-equilibrium with transfers.

Because $p \cdot (\sum_i x_i^*) = r$ and $\sum_i x_i^* = \omega + \sum_j y_j^*$ we know that for any firm $j$:

$$p \cdot (\omega + y_j + \sum_h y_h^*) \leq r = p \cdot (\omega + y_j^* + \sum_h y_h^*) \text{ for } h \neq j$$

which implies $p \cdot y_j \leq p \cdot y_j^*$. Similarly we know:
\[ p \cdot (x_i + \sum_k x_k^*) \geq r = p \cdot (x_i^* + \sum_k x_k^*) \text{for } k \neq i \]

which implies \( p \cdot x_i \geq p \cdot x_i^* \). These two statements, along with the feasibility of the allocation at the Pareto optimum, satisfy the three conditions for a price quasi-equilibrium with transfers supported by wealth levels \( w_i = p \cdot x_i^* \) for all \( i \).

We now turn to conditions under which a price quasi-equilibrium is also a price equilibrium, in other words, conditions under which the statement "if \( x_i \succ_i x_i^* \) then \( p \cdot x_i \geq w_i \)" implies "if \( x_i \succ_i x_i^* \) then \( p \cdot x_i > w_i \)." For this to be true we need now to assume that the consumption set \( X_i \) is convex and the preference relation \( \succeq_i \) is continuous. Then, if there exists a consumption vector \( x_i' \) such that \( x_i' \in X_i \) and \( p \cdot x_i^* < w_i \), a price quasi-equilibrium is a price equilibrium.

To see why, assume to the contrary \( x_i \succ_i x_i^* \) and \( p \cdot x_i = w_i \), and \( x_i \) exists. Then by the convexity of \( X_i \) we have a bundle \( x_i'' = \alpha x_i + (1 - \alpha) x_i^* \in X_i \) with \( p \cdot x_i'' < w_i \). By the continuity of \( \succeq_i \) for \( \alpha \) close to 1 we have \( \alpha x_i + (1 - \alpha) x_i^* \succ_i x_i^* \). This is a contradiction, because this bundle is preferred to \( x_i^* \) and costs less than \( w_i \).

Hence, for price quasi-equilibria to be price equilibria it is sufficient that the consumption set be convex, the preference relation to be continuous, and for there always to exist a "cheaper" consumption bundle \( x_i' \). One way to ensure the existence of such a bundle is to require wealth levels \( w_i \) to be strictly positive for all consumers \( i \).

**Existence**

Even though every equilibrium is efficient, neither of the above two theorems say anything about the equilibrium existing in the first place. To guarantee that an equilibrium exists we once again need consumer preferences to be convex (although with enough consumers this assumption can be relaxed both for existence and the Second Welfare Theorem). Similarly, but less plausibly, feasible production sets must be convex, excluding the possibility of economies of scale.

Proofs of the existence of equilibrium generally rely on fixed point theorems such as Brouwer fixed point theorem or its generalization, the Kakutani fixed point theorem. In fact, one can quickly pass from a general theorem on the existence of equilibrium to Brouwer’s fixed point theorem. For this reason many mathematical economists consider proving existence a deeper result than proving the two Fundamental Theorems.
3.9 WELFARE ECONOMICS IN RELATION TO OTHER SUBJECTS

Welfare economics uses many of the same techniques as microeconomics and can be seen as intermediate or advanced microeconomic theory. Its results are applicable to macroeconomic issues so welfare economics is somewhat of a bridge between the two branches of economics.

Cost-benefit analysis is a specific application of welfare economics techniques, but excludes the income distribution aspects.

Political science also looks into the issue of social welfare (political science), but in a less quantitative manner.

Human development theory explores these issues also, and considers them fundamental to the development process itself.

Activity 3

1. Explain the assumptions in first fundamental theorem of welfare economics.
2. How the second fundamental theorem of welfare economics is important in allocation of resources?
3. Discuss the existence of both the theorems with suitable examples.

3.10 SUMMARY

Welfare economics is a branch of economics that uses microeconomics techniques to simultaneously determine allocative efficiency within an economy and the income Distribution associated with it. Followed by these mainstream concepts like income distribution, efficiency to production and consumption, social welfare maximization and Pareto’s welfare economics were explained. There are two fundamental theorems of Welfare economics. The first states that any Competitive equilibrium or Walrasian equilibrium leads to Pareto efficiency allocation of resources. The second states the converse, that any efficient allocation can be sustainable by a competitive equilibrium. Despite the apparent symmetry of the two theorems, in fact the first theorem is much more general than the second, requiring far weaker assumptions. After this discussion finally relevance of welfare economics in other subjects was given.

3.11 FURTHER READINGS


UNIT 4

BASIC CONCEPTS OF WELFARE ECONOMICS

Objectives

After reading this unit, you should be able to:

• Understand the approach to Pareto’s optimal conditions
• Appreciate the relevance of value judgment
• Know the approaches to social welfare function by various experts
• Identify the concepts compensation principle.
• Have deep understanding of market imperfections, non existence, incompleteness and market failure
• Discuss the ways to obtain optimum welfare through social choice theory

Structure

4.1 Introduction
4.2 Pareto optimal conditions
4.3 Value judgment
4.4 Social welfare function
4.5 Compensation principle
4.6 Market imperfections, incomplete markets and market failure
4.7 Optimum welfare
4.8 Summary
4.9 Further readings

4.1 INTRODUCTION

Welfare economics is a branch of economics that studies efficiency and the overall well-being of society based on alternative allocations of scarce resources. Welfare economics extends the microeconomic analysis of indifference curves to society as a whole. It is concerned with broad efficiency questions and criteria (Pareto efficiency and Kaldor-Hicks efficiency) as well as more specific efficiency issues (market failures, externalities, public goods). Some of the basic concepts of welfare economics will be discussed in this unit.

4.2 PARETO OPTIMAL CONDITIONS

Pareto efficiency, or Pareto optimality, is an important concept in economics with broad applications in game theory, engineering and the social sciences. The term is named after Vilfredo Pareto, an Italian economist who used the concept in his studies of economic
efficiency and income distribution. Informally, Pareto efficient situations are those in which any change to make any person better off would make someone else worse off.

Given a set of alternative allocations of, say, goods or income for a set of individuals, a change from one allocation to another that can make at least one individual better off without making any other individual worse off is called a Pareto improvement. An allocation is defined as Pareto efficient or Pareto optimal when no further Pareto improvements can be made. This is often called a strong Pareto optimum (SPO).

A weak Pareto optimum (WPO) satisfies a less stringent requirement, in which a new allocation is only considered to be a Pareto improvement if it is strictly preferred by all individuals (i.e., all must gain with the new allocation). In other words, when an allocation is WPO there are no possible alternative allocations where every individual would gain. An SPO is a WPO: a WPO is an allocation where every change causes some individual to NOT IMPROVE, whereas with an SPO every change causes some individual to DO WORSE (and hence not improve).

Formally, a (strong/weak) Pareto optimum is a maximal element for the partial order relation of Pareto improvement/strict Pareto improvement: it is an allocation such that no other allocation is "better" in the sense of the order relation.

A common criticism of a state of Pareto efficiency is that it does not necessarily result in a socially desirable distribution of resources, as it makes no statement about equality or the overall well-being of a society; notably, allocating all resources to one person and none to anyone else is Pareto efficient if preference relations are monotone increasing.

An economic system that is Pareto inefficient implies that a certain change in allocation of goods (for example) may result in some individuals being made "better off" with no individual being made worse off, and therefore can be made more Pareto efficient through a Pareto improvement. Here 'better off' is often interpreted as "put in a preferred position." It is commonly accepted that outcomes that are not Pareto efficient are to be avoided, and therefore Pareto efficiency is an important criterion for evaluating economic systems and public policies.

If economic allocation in any system (in the real world or in a model) is not Pareto efficient, there is theoretical potential for a Pareto improvement — an increase in Pareto efficiency: through reallocation, improvements to at least one participant's well-being can be made without reducing any other participant's well-being.

In the real world ensuring that nobody is disadvantaged by a change aimed at improving economic efficiency may require compensation of one or more parties. For instance, if a change in economic policy dictates that a legally protected monopoly ceases to exist and that market subsequently becomes competitive and more efficient, the monopolist will be made worse off. However, the loss to the monopolist will be more than offset by the gain in efficiency. This means the monopolist can be compensated for its loss while still
leaving an efficiency gain to be realized by others in the economy. Thus, the requirement of nobody being made worse off for a gain to others is met.

In real-world practice, the compensation principle often appealed to is hypothetical. That is, for the alleged Pareto improvement (say from public regulation of the monopolist or removal of tariffs) some losers are not (fully) compensated. The change thus results in distribution effects in addition to any Pareto improvement that might have taken place. The theory of hypothetical compensation is part of Kaldor-Hicks efficiency, also called Potential Pareto Criterion. (Ng, 1983).

Under certain idealized conditions, it can be shown that a system of free markets will lead to a Pareto efficient outcome. This is called the first welfare theorem. It was first demonstrated mathematically by economists Kenneth Arrow and Gerard Debreu. However, the result does not rigorously establish welfare results for real economies because of the restrictive assumptions necessary for the proof (markets exist for all possible goods, all markets are in full equilibrium, markets are perfectly competitive, transaction costs are negligible, there must be no externalities, and market participants must have perfect information). Moreover, it has since been demonstrated mathematically that, in the absence of perfect competition or complete markets, outcomes will generically be Pareto inefficient.

**PARETO FRONTIER**

Given a set of choices and a way of valuing them, the Pareto frontier or Pareto set is the set of choices that are Pareto efficient. The Pareto frontier is particularly useful in engineering: by restricting attention to the set of choices that are Pareto-efficient, a designer can make tradeoffs within this set, rather than considering the full range of every parameter.

The Pareto frontier is defined formally as follows.

Consider a design space with \( n \) real parameters, and for each design-space point there are \( m \) different criteria by which to judge that point. Let \( f : \mathbb{R}^n \to \mathbb{R}^m \) be the function which assigns, to each design-space point \( x \), a criteria-space point \( f(x) \). This represents the way of valuing the designs. Now, it may be that some designs are infeasible; so let \( X \) be a set of feasible designs in \( \mathbb{R}^n \), which must be a compact set. Then the set which represents the feasible criterion points is \( f(X) \), the image of the set \( X \) under the action of \( f \). Call this image \( Y \).

Now construct the Pareto frontier as a subset of \( Y \), the feasible criterion points. It can be assumed that the preferable values of each criterion parameter are the lesser ones, thus minimizing each dimension of the criterion vector. Then compare criterion vectors as follows: One criterion vector \( x \) strictly dominates (or "is preferred to") a vector \( y \) if each parameter of \( x \) is no greater than the corresponding parameter of \( y \) and at least one parameter is strictly less: that is, \( x_i \leq y_i \) for each \( i \) and \( x_i < y_i \) for some \( i \). This is
written as \( \mathbf{X} \succ \mathbf{Y} \) to mean that \( \mathbf{x} \) strictly dominates \( \mathbf{y} \). Then the Pareto frontier is the set of points from \( Y \) that are not strictly dominated by another point in \( Y \).

Formally, this defines a partial order on \( Y \), namely the (opposite of the) product order on \( \mathbb{R}^m \) (more precisely, the induced order on \( Y \) as a subset of \( \mathbb{R}^m \)), and the Pareto frontier is the set of maximal elements with respect to this order.

Algorithms for computing the Pareto frontier of a finite set of alternatives have been studied in computer science. There, this task is known as the maximum vector problem or as skyline query.

**RELATIONSHIP TO MARGINAL RATE OF SUBSTITUTION**

An important fact about the Pareto frontier in economics is that at a Pareto efficient allocation, the marginal rate of substitution is the same for all consumers. A formal statement can be derived by considering a system with \( m \) consumers and \( n \) goods, and a utility function of each consumer as \( z_i = f^i(x^i) \) where \( x^i = (x_1^i, x_2^i, \ldots, x_n^i) \) is the vector of goods, both for all \( i \). The supply constraint is written \( \sum_{i=1}^{m} x_j^i = b_j^0 \) for \( j = 1, \ldots, n \). To optimize this problem, the Lagrangian is used:

\[
L(x, \lambda, \Gamma) = f^1(x^1) + \sum_{i=2}^{m} \lambda_i (z_i^0 - f^i(x^i)) + \sum_{j=1}^{n} \Gamma_j (b_j^0 - \sum_{i=1}^{m} x_j^i)
\]

and \( \lambda \) and \( \Gamma \) are multipliers.

Taking the partial derivative of the Lagrangian with respect to one good, \( i \), and then taking the partial derivative of the Lagrangian with respect to another good, \( j \), gives the following system of equations:

\[
\frac{\partial L}{\partial x_j^i} = f_{x_j^i}^1 - \Gamma_j^0 = 0 \quad \text{for } j=1,\ldots,n.
\]

\[
\frac{\partial L}{\partial x_j^i} = -\lambda_i^0 f_{x_j^i}^1 - \Gamma_j^0 = 0 \quad \text{for } i = 2,\ldots,m \text{ and } j=1,\ldots,m,
\]

where \( f_{x_i}^i \) is the marginal utility on \( f^i \) of \( x \) (the partial derivative of \( f^i \) with respect to \( x \)).

\[
\frac{f_i^j}{x_j^i} = f_{x_j^k}^i \quad \text{for } i,k=1,\ldots,m \text{ and } j,s=1,\ldots,n\ldots.
\]
CRITICISM

Pareto efficiency does not consider the equity of resource allocations. It may be that one economic agent owns all of the world's resources; it would be impossible to make anyone else better off without taking something away from this agent. Thus this situation is described as "Pareto optimal", even though it may be seen as inequitable.

More generally, it can be misleading, in that "not Pareto optimal" implies "can be improved" (making someone better off without hurting anyone), but "Pareto optimal" does not imply "cannot be improved" by some measure—it only implies that someone must receive less. Thus if an allocation is not Pareto optimal, it means that one can improve it, but does not mean that one should categorically reject it for any arbitrary Pareto optimal solution, as many of those Pareto optimal solutions will not be Pareto improvements.

More importantly, not all Pareto optimal states are necessarily equally desirable from the standpoint of society in general. For instance, a one-time transfer of wealth from the very wealthy to the very poor may not be a Pareto improvement but may nevertheless result in a new Pareto optimal state that might be seen as more socially desirable than the previous one.

4.3 VALUE JUDGMENT

A value judgment is a judgment of the rightness or wrongness of something, or of the usefulness of something, based on a personal view. As a generalization, a value judgment can refer to a judgment based upon a particular set of values or on a particular value system. A related meaning of value judgment is an expedient evaluation based upon limited information at hand, an evaluation undertaken because a decision must be made on short notice.

The term value judgment can be used both in a positive sense, signifying that a judgment must be made taking a value system into account, or in a disparaging sense, signifying a judgment made by personal whim rather than rational, objective thought.

In its positive sense, recommendation to make a value judgment is an admonition to consider carefully, to avoid whim and impetuousness, and search for consonance with one’s deeper convictions.

In its disparaging sense the term value judgment implies a conclusion is insular, one-sided, and not objective — contrasting with judgments based upon deliberation, balance and rationality.

Value judgment also can refer to a tentative judgment based on a considered appraisal of the information at hand, taken to be incomplete and evolving, for example, a value judgment on whether to launch a military attack, or as to procedure in a medical emergency. In this case the quality of judgment suffers because the information
available is incomplete as a result of exigency, rather than as a result of cultural or personal limitations.

Most commonly the term *value judgment* refers to an individual's opinion. Of course, the individual's opinion is formed to a degree by their belief system, and the culture to which they belong. So a natural extension of the term *value judgment* is to include declarations seen one way from one value system, but which may be seen differently from another. Conceptually this extension of definition is related both to the anthropological axiom "cultural relativism" (that is, that cultural meaning derives from a context) and to the term "moral relativism" (that is, that moral and ethical propositions are not universal truths, but stem from cultural context). In the pejorative sense, a value judgment formed within a specific value system may be parochial, and may be subject to dispute in a wider audience.

**Nonjudgmental**

*Nonjudgmental* is a descriptor that conveys the opposite meaning to the pejorative sense of value judgment: it expresses avoidance of personal opinion and reflex "knee-jerk" reactions.

**Judgment call**

*Judgment call* is a term describing decision made between alternatives that are not clearly right or wrong, and so must be made on a personal basis.

**Value-neutral**

*Value-neutral* is a related adjective suggesting independence from a value system. For example, the classification of an object sometimes depends upon context: Is it a tool or a weapon, an artifact or an ancestor? The object itself might be considered value-neutral being neither good nor bad, neither useful nor useless, neither significant nor trite, until placed in some social context.

**VALUE JUDGMENTS AND THEIR CONTEXT**

Some argue that true objectivity is impossible, that even the most rigorous rational analysis is founded on the set of values accepted in the course of analysis. See [Free On-Line Dictionary of Philosophy](https://www.fordham.edu/halsall/mod/philosophy/). Consequently, all conclusions are necessarily value judgments (and therefore maybe suspect). Of course, putting all conclusions in one category does nothing to distinguish between them, and is therefore a useless descriptor except as a rhetorical device intended to discredit a position claiming higher authority.

As an example of a more nuanced view, scientific "truths" are considered objective, but are held tentatively, with the understanding that more careful evidence and/or wider experience might change matters. Further, a scientific view (in the sense of a conclusion based upon a value system) is a *value judgment* based upon rigorous evaluation and wide
consensus. With this example in mind, characterizing a view as a *value judgment* is vague without description of the context surrounding it.

However, as noted in the first segment of this article, in common usage the term *value judgment* has a much simpler meaning with context simply implied, not specified.

### 4.4 SOCIAL WELFARE FUNCTION

In *economics*, a social welfare function is a *real-valued function* that ranks conceivable social states (alternative complete descriptions of the society) from lowest to highest. Inputs of the function include any variables considered to affect welfare of the society (Sen, 1970, p. 33). In using welfare measures of persons in the society as inputs, the social welfare function is individualistic in form. One use of a social welfare function is to represent prospective patterns of collective choice as to alternative social states.

The social welfare function is analogous to an *indifference-curve* map for an individual, except that the social welfare function is a mapping of individual preferences or judgments of everyone in the society as to collective choices, which apply to all, whatever individual preferences are. One point of a social welfare function is to determine how close the analogy is to an ordinal *utility function* for an individual with at least minimal restrictions suggested by welfare economics. Kenneth Arrow proved a more basic point for a set of seemingly reasonable conditions.

#### 4.4.1 BERGSON-SAMUELSON SOCIAL WELFARE FUNCTION

In a 1938 article Abram Bergson introduced the *social welfare function*. The object was "to state in precise form the value judgments required for the derivation of the conditions of maximum economic welfare" set out by earlier writers, including Marshall and Pigou, Pareto and Barone, and Lerner. The function was real-valued and differentiable. It was specified to describe the society as a whole. Arguments of the function included the quantities of different commodities produced and consumed and of *resources* used in producing different commodities, including labor.

Necessary general conditions are that at the maximum value of the function:

- The marginal "dollar's worth" of welfare is equal for each individual and for each commodity.
- The marginal "dissatisfaction" of each "dollar's worth" of labor is equal for each commodity produced of each labor supplier.
- The marginal "dollar" cost of each unit of resources is equal to the marginal value productivity for each commodity.

Auxiliary specifications enable comparison of different social states by each member of society in preference satisfaction. These help define *Pareto efficiency*, which holds if all alternatives have been exhausted to put at least one person in a *more preferred position* with no one put in a *less preferred position*. Bergson described an "economic welfare
increase" (later called a Pareto improvement) as at least one individual moving to a more preferred position with everyone else indifferent. The social welfare function could then be specified in a substantively individualistic sense to derive Pareto efficiency (optimality). Paul Samuelson (2004, p. 26) notes that Bergson's function "could derive Pareto optimality conditions as necessary but not sufficient for defining interpersonal normative equity." Still, Pareto efficiency could also characterize one dimension of a particular social welfare function with distribution of commodities among individuals characterizing another dimension. As Bergson noted, a welfare improvement from the social welfare function could come from the "position of some individuals" improving at the expense of others. That social welfare function could then be described as characterizing an equity dimension.

Samuelson (1947, p. 221) himself stressed the flexibility of the social welfare function to characterize any one ethical belief, Pareto-bound or not, consistent with:

- a complete and transitive ranking (an ethically "better", "worse", or "indifferent" ranking) of all social alternatives and
- one set out of an infinity of welfare indices and cardinal indicators to characterize the belief.

He also presented a lucid verbal and mathematical exposition of the social welfare function (1947, pp. 219-49) with minimal use of Lagrangean multipliers and without the difficult notation of differentials used by Bergson throughout. As Samuelson (1983, p. xxii) notes, Bergson clarified how production and consumption efficiency conditions are distinct from the interpersonal ethical values of the social welfare function.

Samuelson further sharpened that distinction by specifying the Welfare function and the Possibility function (1947, pp. 243-49). Each has as arguments the set of utility functions for everyone in the society. Each can (and commonly does) incorporate Pareto efficiency. The Possibility function also depends on technology and resource restraints. It is written in implicit form, reflecting the feasible locus of utility combinations imposed by the restraints and allowed by Pareto efficiency. At a given point on the Possibility function, if the utility of all but one person is determined, the remaining person's utility is determined. The Welfare function ranks different hypothetical sets of utility for everyone in the society from ethically lowest on up (with ties permitted), that is, it makes interpersonal comparisons of utility. Welfare maximization then consists of maximizing the Welfare function subject to the Possibility function as a constraint. The same welfare maximization conditions emerge as in Bergson's analysis.

### 4.4.2 ARROW SOCIAL WELFARE FUNCTION (CONSTITUTION)

Kenneth Arrow (1963) generalizes the analysis. Along earlier lines, his version of a social welfare function, also called a 'constitution', maps a set of individual orderings (ordinal utility functions) for everyone in the society to a social ordering, a rule for ranking alternative social states (say passing an enforceable law or not, ceteris paribus). Arrow finds that nothing of behavioral significance is lost by dropping the requirement of
social orderings that are real-valued (and thus cardinal) in favor of orderings, which are merely complete and transitive, such as a standard indifference-curve map. The earlier analysis mapped any set of individual orderings to one social ordering, whatever it was. This social ordering selected the top-ranked feasible alternative from the economic environment as to resource constraints. Arrow proposed to examine mapping different sets of individual orderings to possibly different social orderings. Here the social ordering would depend on the set of individual orderings, rather than being imposed (invariant to them). Stunningly (relative to a course of theory from Adam Smith and Jeremy Bentham on), Arrow proved the General Possibility Theorem that it is impossible to have a social welfare function that satisfies a certain set of "apparently reasonable" conditions.

4.4.3 CARDINAL SOCIAL WELFARE FUNCTIONS

In the above contexts, a social welfare function provides a kind of social preference based on only individual utility functions, whereas in others it includes cardinal measures of social welfare not aggregated from individual utility functions. Examples of such measures are life expectancy and per capita income for the society. The rest of this article adopts the latter definition.

The form of the social welfare function is intended to express a statement of objectives of a society. For example, take this example of a social welfare function:

\[ W = Y_1 + Y_2 + \ldots + Y_n \]

where \( W \) is social welfare and \( Y_i \) is the income of individual \( i \) among \( n \) in the society. In this case, maximising the social welfare function means maximising the total income of the people in the society, without regard to how incomes are distributed in society. Alternatively, consider the Max-Min utility function (based on the philosophical work of John Rawls):

\[ W = \min(Y_1, Y_2, \ldots, Y_n) \]

Here, the social welfare of society is taken to be related to the income of the poorest person in the society, and maximising welfare would mean maximising the income of the poorest person without regard for the incomes of the others.

These two social welfare functions express very different views about how a society would need to be organised in order to maximise welfare, with the first emphasizing total incomes and the second emphasising the needs of the poorest. The max-min welfare function can be seen as reflecting an extreme form of uncertainty aversion on the part of society as a whole, since it is concerned only with the worst conditions that a member of society could face.

Amartya Sen proposed a welfare function in 1973:

\[ W_{\text{Gini}} = \frac{\text{Income}}{\text{Income}} \cdot (1 - G^2) \]
The average per capita income of a measured group (e.g. nation) is multiplied with \( (1 - G) \) where \( G \) is the Gini index, a relative inequality measure. James E. Foster (1996) proposed to use one of Atkinson's Indexes, which is an entropy measure. Due to the relation between Atkinsons entropy measure and the Theil index, Foster's welfare function also can be computed directly using the Theil-L Index.

\[
W_{\text{Theil-L}} = \ln(\text{Income}) \cdot e^{-T_L}
\]

The value yielded by this function has a concrete meaning. There are several possible incomes which could be earned by a person, who randomly is selected from a population with an unequal distribution of incomes. This welfare function marks the income, which a randomly selected person is most likely to have. Similar to the median, this income will be smaller than the average per capita income.

\[
W_{\text{Theil-T}}^{-1} = \ln(\text{Income}) \cdot e^{T_T}
\]

Here the Theil-T index is applied. The inverse value yielded by this function has a concrete meaning as well. There are several possible incomes to which an Euro may belong, which is randomly picked from the sum of all unequally distributed incomes. This welfare function marks the income, which a randomly selected Euro most likely belongs to. The inverse value of that function will be larger than the average per capita income.

### 4.5 COMPENSATION PRINCIPLE

In welfare economics, the compensation principle refers to a decision rule used to select between pairs of alternative feasible social states. One of these states is the hypothetical point of departure ("the original state"). According to the compensation principle, if the prospective gainer could compensate (any) prospective losers and leave no one worse off, the other state is to be selected (Chipman, 1987, p. 524). An example of a compensation principle is the Pareto criterion in which a change in states entails that such compensation is not merely feasible but required. Two variants are:

- the Pareto principle, which requires any change such that all gain.
- the (strong) Pareto criterion, which requires any change such that at least one gains and no one loses from the change.

In non-hypothetical contexts such that the compensation occurs (say in the marketplace), invoking the compensation principle is unnecessary to effect the change. But its use is more controversial and complex with some losers (where full compensation is feasible but not made) and in selecting among more than two feasible social states. In its specifics, it is also more controversial where the range of the decision rule itself is at issue.

Uses for the compensation principle include:
comparisons between the welfare properties of perfect competition and imperfect competition
the Pareto principle in social choice theory
cost-benefit analysis.

4.6 MARKET IMPERFECTIONS, INCOMPLETE MARKETS AND MARKET FAILURE

The Theory of Incomplete Markets is an extension of the general equilibrium approach to intertemporal economies with uncertainty, where the set of available contracts which can be used to transfer wealth across time is limited relative to the possible probabilistic states that an economy might find itself in. Unlike in the standard Arrow-Debreu model where all trade is planned at beginning of time. These agents can do this for their descendents at the beginning of time in the Classical model (i.e. with complete markets) because agents are assumed to have costless contractual enforcement and perfect calculations along with perfect knowledge of the likelihood of all possible future states (across an unlimited range of contracts). In an economy with incomplete markets agents trade in sequential spot markets.

There are at least two results that significantly depart from those well-known results in complete markets.

1. Generic Existence of Equilibrium
2. Non Pareto-Optimality of Allocations

The First Welfare Theorem concerning the Pareto Optimality of general equilibrium no longer holds.

In economics, a market failure exists when the production or use of goods and services by the market is not efficient. That is, there exists another outcome where market participants' total gains from the new outcome outweigh their losses (even if some participants lose under the new arrangement). Market failures can be viewed as scenarios where individuals' pursuit of pure self-interest leads to results that are not efficient – that can be improved upon from the societal point-of-view. The first known use of the term by economists was in 1958, but the concept has been traced back to the Victorian philosopher Henry Sidgwick. Market failures are often associated with non-competitive markets, externalities or public goods. The existence of a market failure is often used as a justification for government intervention in a particular market. Economists, especially microeconomists, are often concerned with the causes of market failure, and possible means to correct such a failure when it occurs. Such analysis plays an important role in many types of public policy decisions and studies. However, some types of government policy interventions, such as taxes, subsidies, bailouts, wage and price controls, and regulations, including attempts to correct market failure, may also lead to an inefficient allocation of resources, (sometimes
called government failures). Thus, there is often a choice between imperfect outcomes, i.e. imperfect market outcomes with or without government interventions.

CAUSES

According to mainstream economic analysis, a market failure (relative to Pareto efficiency) can occur for three main reasons.

- First, agents in a market can gain market power, allowing them to block other mutually beneficial gains from trades from occurring. This can lead to inefficiency due to imperfect competition, which can take many different forms, such as monopolies, monopsonies, cartels, or monopolistic competition, if the agent does not implement perfect price discrimination. In a monopoly, the market equilibrium will no longer be Pareto optimal. The monopoly will use its market power to restrict output below the quantity at which the MSB is equal to the MSC of the last unit produced, so as to keep prices and profits high.

- Second, the actions of agents can have externalities, which are innate to the methods of production, or other conditions important to the market. For example, when a firm is producing steel, it absorbs labor, capital and other inputs, it must pay for these in the appropriate markets, and these costs will be reflected in the market price for steel. If the firm also pollutes the atmosphere when it makes steel, however, and if it is not forced to pay for the use of this resource, then this cost will be borne not by the firm but by society. Hence, the market price for steel will fail to incorporate the full opportunity cost to society of producing. In this case, the market equilibrium in the steel industry will not be optimal. More steel will be produced than would occur were the firm to have to pay for all of its costs of production. Consequently, the MSC of the last unit produced will exceed its MSB. Finally, some markets can fail due to the nature of certain goods, or the nature of their exchange. For instance, goods can display the attributes of public goods or common-pool resources, while markets may have significant transaction costs, agency problems, or informational asymmetry. In general, all of these situations can produce inefficiency, and a resulting market failure.

More fundamentally, the underlying cause of market failure is often a problem of property rights. As Hugh Gravelle and Ray Rees put it,

*A market is an institution in which individuals or firms exchange not just commodities, but the rights to use them in particular ways for particular amounts of time. [...] Markets are institutions which organize the exchange of control of commodities, where the nature of the control is defined by the property rights attached to the commodities.*

As a result, agents’ control over the uses of their commodities can be imperfect, because the system of rights which defines that control is incomplete. Typically, this falls into two generalized rights – excludability and transferability. Excludability deals with the ability of agents to control who uses their commodity, and for how long – and the related costs associated with doing so. Transferability reflects the right of agents to transfer the rights
of use from one agent to another, for instance by selling or leasing a commodity, and the costs associated with doing so. If a given system of rights does not fully guarantee these at minimal (or no) cost, then the resulting distribution can be inefficient. Considerations such as these form an important part of the work of institutional economics. Nonetheless, views still differ on whether something displaying these attributes is meaningful without the information provided by the market price system.

There are many examples cited by economists as examples of market failure. For instance, traffic congestion is considered an example, since driving can impose hidden costs on other drivers and society, whereas use of public transportation or other ways of avoiding driving would be more beneficial to society as a whole. Other common examples of market failure include environmental problems such as pollution or overexploitation of natural resources.

**INTERPRETATIONS AND POLICY**

The above causes represent the mainstream view of what market failures mean and of their importance in the economy. This analysis follows the lead of the neoclassical school, and relies on the notion of Pareto efficiency – and specifically considers market failures absent considerations of the "public interest", or equity, citing definitional concerns. This form of analysis has also been adopted by the Keynesian or new Keynesian schools in modern macroeconomics, applying it to Walrasian models of general equilibrium in order to deal with failures to attain full employment, or the non-adjustment of prices and wages.

Many social democrats and "New Deal liberals", have adopted this analysis for public policy, so they view market failures as a very common problem of any unregulated market system and therefore argue for state intervention in the economy in order to ensure both efficiency and social justice (usually interpreted in terms of limiting avoidable inequalities in wealth and income). Both the democratic accountability of these regulations and the technocratic expertise of the economists play an important role here in shaping the kind and degree of intervention. Neoliberals follow a similar line, often focusing on "market-oriented solutions" to market failure: for example, they propose going beyond the common idea of having the government charge a fee for the right to pollute (internalizing the external cost, creating a disincentive to pollute) to allow polluters to sell the pollution permits.

**EXAMPLES OF POTENTIAL MARKET FAILURE**

There are plenty of reasons why the normal operation of market forces may not lead to economic efficiency.

**Public Goods**

Public Goods not provided by the free market because of their two main characteristics
- **Non-excludability** where it is not possible to provide a good or service to one person without it thereby being available for others to enjoy

- **Non-rivalry** where the consumption of a good or service by one person will not prevent others from enjoying it

**Examples**: Street lighting / Lighthouse Protection, Police services, Air defence systems, Roads / motorways, Terrestrial television, Flood defence systems, Public parks & beaches

Because of their nature the private sector is unlikely to be willing and able to provide public goods. The government therefore provides them for collective consumption and finances them through general taxation.

**Merit Goods**

**Merit Goods** are those goods and services that the government feels that people left to themselves will under-consume and which therefore ought to be subsidised or provided free at the point of use.

Both the public and private sector of the economy can provide merit goods & services. Consumption of merit goods is thought to generate positive externality effects where the social benefit from consumption exceeds the private benefit.

**Examples**: Health services, Education, Work Training, Public Libraries, Citizen's Advice, Inoculations

**Monopoly**

Few modern markets meet the stringent conditions required for a perfectly competitive market. The existence of monopoly power is often thought to create the potential for market failure and a need for intervention to correct for some of the welfare consequences of monopoly power.

The classical economic case against monopoly is that

- Price is higher and output is lower under monopoly than in a competitive market
- This causes a net economic welfare loss of both consumer and producer surplus
- Price > marginal cost - leading to allocative inefficiency and a pareto sub-optimal equilibrium.
- Rent seeking behaviour by the monopolist might add to the standard costs of monopoly. This includes high (possibly excessive) amounts of spending on persuasive advertising and marketing.
- Libenstein's X-inefficiency may also result if the monopolist allows cost efficiency to drop. An upward drift in costs because of a lack of effective
competition in the market-place can lead to consumers facing higher prices and a reduction in their real standard of living.

**Externalities**

Any exam question on market failure must make some reference to externalities. What are the potential market failures arising from externalities?

*The social optimum output or level of consumption diverges from the private optimum.*

Main problem is the absence of clearly defined property rights for those agents operating in the market. When property rights are not clearly defined, market failure is likely because producers & consumers may not be held to account.

Don't forget that positive externalities can also justify intervention if goods are under-consumed (social benefit > private benefit).

**Inequality**

Market failure can also be caused by the existence of inequality throughout the economy. Wide differences in income and wealth between different groups within our economy leads to a wide gap in living standards between affluent households and those experiencing poverty. Society may come to the view that too much inequality is unacceptable or undesirable.

Note here that value judgements come into play whenever we discuss the distribution of income and wealth in society. The government may decide to intervene to reduce inequality through changes to the tax and benefits system and also specific policies such as the national minimum wage.

**GOVERNMENT INTERVENTION AND MARKET FAILURE**

Government intervention may seek to correct for the distortions created by market failure and to improve the efficiency in the way that markets operate.

- Pollution taxes to correct for externalities
- Taxation of monopoly profits (the Windfall Tax)
- Regulation of oligopolies/cartel behaviour
- Direct provision of public goods (defence)
- Policies to introduce competition into markets (de-regulation)
- Price controls for the recently privatised utilities

**4.7 OPTIMUM WELFARE**

In view of most of the economists it is the right of all individuals to get optimum benefit from the resources they are being provides by their nations. Still there is a huge
imbalance of individual expectations of welfare and welfare provisions. The problem can be solved by way of social choice theory.

Social choice theory studies how measures of individual interests, values, or welfare in theory could be aggregated to reach a collective decision. A non-theoretical example of a collective decision is passing a set of laws under a constitution. Social choice theory dates from Condorcet’s formulation of the voting paradox. Kenneth Arrow’s 1951 book *Social Choice and Individual Values* and Arrow’s impossibility theorem in it established the theory in its modern form.

Social choice theory blends elements of welfare economics and voting theory and generalizes them. It is methodologically individualistic, that is, "bottom-up," in aggregating from individuals to society. A characteristic method proceeds from formulating some set of apparently reasonable axioms of social choice to construct a social welfare function (or constitution) and derive the implications of those axioms. Many earlier results indicate the logical incompatibility of different axioms, revealing an aggregation problem and suggesting reformulation or theoretical triage in dropping some axiom(s).

A related field is public choice theory. Public Choice theory is often used to explain how political decision-making results in outcomes that conflict with the preferences of the general public. For example, many special interest and pork barrel projects are not the desire of the overall democracy. However, it makes sense for politicians to support these projects. It may make them feel powerful and important. It can also benefit them financially by opening the door to future wealth as lobbyists. The project may be of interest to the politician's local constituency, increasing district votes or campaign contributions. The politician pays little or no cost to gain these benefits, as he is spending public money. Special-interest lobbyists are also behaving rationally. They can gain government favors worth millions or billions for relatively small investments. They face a risk of losing out to their competitors if they don't seek these favors. The taxpayer is also behaving rationally. The cost of defeating any one government give-away is very high, while the benefits to the individual taxpayer are very small. Each citizen pays only a few pennies or a few dollars for any given government favor, while the costs of ending that favor would be many times higher. Everyone involved has rational incentives to do exactly what they're doing, even though the desire of the general constituency is opposite. (It is notable that the political system considered here is very much that of the United States, with "pork" a main aim of individual legislators; in countries such as Britain with strong party systems the issues would differ somewhat.)

**Activity 4**

1. What do you understand by Pareto efficiency?
2. Discuss the concept of market failure giving suitable examples for potential market failure.
3. Give a brief note on Bergson-Samuelson social welfare function.
4. write short notes on the following:
   - value judgment
   - social welfare function
   - compensation principle

4.8 SUMMARY

The unit was started discussing Pareto’s optimal conditions along with criticism. Followed by this their was the explanation of term value judgment as a judgment of the rightness or wrongness of something, or of the usefulness of something, based on a personal view. Further the social welfare function and compensation principle were discussed in detail to give readers the clear understanding of concepts. Market failure was discussed as the market condition when the production or use of goods and services by the market is not efficient. Similarly imperfections of markets were explained giving suitable examples. Finally, why and how the inability to obtain optimum welfare can be avoided was discussed in brief.

4.9 FURTHER READINGS

UNIT 5

THE SECOND BEST THEORY AND IMPOSSIBILITY THEOREM

Objectives

After reading this unit, you should be able to:

- Understand the theory of the second best.
- Explain the solution to the theory of second best
- Use the approach of Arrow’s impossibility theorem in decision making
- Identify the concept of voting paradox.
- Know the assumptions in Arrow Debreu model

Structure

5.1 Introduction
5.2 Theory of the second best
5.3 Theory of the second best solution
5.4 Arrow’s impossibility theorem
5.5 The voting paradox
5.6 Arrow Debreu model
5.7 Summary
5.8 Further readings

5.1 INTRODUCTION

The Theory of the Second Best concerns what happens when one or more optimality conditions cannot be satisfied in an economic model. Canadian economist Richard Lipsey and Australian-American economist Kelvin Lancaster showed in a 1956 paper that if one optimality condition in an economic model cannot be satisfied, it is possible that the next-best solution involves changing other variables away from the ones that are usually assumed to be optimal.

This means that in an economy with some unavoidable market failure in one sector, there can actually be a decrease in efficiency due to a move toward greater market perfection in another sector. In theory, at least, it may be better to let two market imperfections cancel each other out rather than making an effort to fix either one. Thus, it may be optimal for the government to intervene in a way that is contrary to laissez faire policy. This suggests that economists need to study the details of the situation before jumping to the theory-based conclusion that an improvement in market perfection in one area implies a global improvement in efficiency.
Even though the theory of the second best was developed for the Walrasian general equilibrium system, it also applies in microeconomic (partial equilibrium) cases. For example, consider a mining monopoly that's also a polluter: mining leads to tailings being dumped in the river and deadly dust in the workers’ lungs. Suppose in addition that there is nothing at all that we can do about the pollution. But the government is able to break up the monopoly.

The problem here is that increasing competition in this market is likely to increase production (since competitors have such a hard time restricting production compared to a monopoly). Because pollution is highly associated with production, pollution will most likely increase. This may actually make the world worse off than before.

On the other hand Arrow’s impossibility theorem is based on the different notion. Proof that something cannot be done or cannot be had. The most famous such result in politics, due to K. J. Arrow, proves that if a choice or ordering system (such as an electoral procedure) produces results that are transitive and consistent (see economic man), satisfies ‘universal domain’ (that is, works for all possible combinations of individual preference), satisfies the weak Pareto condition, and is independent of irrelevant alternatives, then it is dictatorial. ‘Dictatorial’ here has a technical meaning, namely, that the preferences of one individual may determine the social choice, irrespective of the preferences of any other individuals in the society. A non-technical interpretation of Arrow's theorem is as follows. In a society, group choices, or group rankings, often have to be made between courses of action or candidates for a post. We would like a good procedure to satisfy some criteria of fairness as well as of logicality. Arrow's startling proof shows that a set of extremely weak such criteria is inconsistent. We would like a good procedure to satisfy not only these but much more besides. But that is logically impossible.

Impossibility theorems save time. For instance, much work by electoral reformers amounts to trying to evade Arrow's theorem. As we know it cannot be done, this removes the need to scrutinize many such schemes in detail. This is not to say that all electoral systems are equally bad, however; there remains an important job for electoral reformers within the limit set by Arrow's and other impossibility theorems.

5.2 THEORY OF THE SECOND BEST

The Theory of the Second Best concerns what happens when one or more optimality conditions cannot be satisfied in an economic model. Canadian economist Richard Lipsey and Australian-American economist Kelvin Lancaster showed in a 1956 paper that if one optimality condition in an economic model cannot be satisfied, it is possible that the next-best solution involves changing other variables away from the ones that are usually assumed to be optimal.

This means that in an economy with some unavoidable market failure in one sector, there can actually be a decrease in efficiency due to a move toward greater market perfection in another sector. In theory, at least, it may be better to let two market imperfections cancel
each other out rather than making an effort to fix either one. Thus, it may be optimal for the government to intervene in a way that is contrary to laissez faire policy. This suggests that economists need to study the details of the situation before jumping to the theory-based conclusion that an improvement in market perfection in one area implies a global improvement in efficiency.

Even though the theory of the second best was developed for the Walrasian general equilibrium system, it also applies in microeconomic (partial equilibrium) cases. For example, consider a mining monopoly that's also a polluter: mining leads to tailings being dumped in the river and deadly dust in the workers’ lungs. Suppose in addition that there is nothing at all that we can do about the pollution. But the government is able to break up the monopoly.

The problem here is that increasing competition in this market is likely to increase production (since competitors have such a hard time restricting production compared to a monopoly). Because pollution is highly associated with production, pollution will most likely increase. This may actually make the world worse off than before.

UNDERLYING ASSUMPTIONS OF THE THEORY OF SECOND-BEST

The theory of the second-best provides the theoretical underpinning to explain many of the reasons that trade policy can be shown to be welfare enhancing for an economy. In most (if not all) of the cases in which trade policy is shown to improve national welfare, the economy begins at an equilibrium that can be characterized as second best. Second best equilibria arise whenever the market has distortions or imperfections present. In these cases it is relatively straightforward to conceive of a trade policy which corrects the distortion or imperfection sufficiently to outweigh the detrimental effects of the policy itself. In other words, whenever there are market imperfections or distortions present it is always theoretically or conceptually possible to design a trade policy that would improve national welfare. As such the theory of the second best provides a rationale for many different types of protection in an economy.

The main criticism suggested by the theory is that rarely is trade policy the first best policy choice to correct a market imperfection or distortion. Instead trade policy is second best. The first best policy, generally, would be a purely domestic policy targeted directly at the market imperfection or distortion.

On the following pages we use the theory of the second best to explain many of the justifications commonly given for protection or for government intervention with some form of trade policy. In each case we also discuss the likely first best policies.
5.3 THE THEORY OF A SECOND-BEST SOLUTION

The theory of a second-best solution concerns the events that happen when a condition for an optimal outcome isn't met. In that case a second-best solution should be sought. But the second-best solution isn't always the one where every other condition is met except the one missing to make the solution optimal. Thus, in order to get the second-best solution where one or more necessary conditions haven't been met, it isn't necessary, it is in fact a bad idea to try to keep the other, already met conditions. In other words, one should allow the market deficiencies to cancel themselves out. E.g. in a perfect competitive state the optimum is found if the price and border cost is equal in all market sectors. Should the price in a sector grow above the border costs, the second-best solution will, for example, require taxes to make the prices grow elsewhere, because that way the consumers’ border decisions about the allocation of their budget to various products stay almost unchanged. After being brought up in the 1965 work by a Canadian, Richard Lipsey (born in 1928) and an Australian, Kelvin Lancaster (1924-1999), the theory was also used, except in economy, in the legislative sciences.

The primary focus of the theory is on what happens when the optimum conditions are not satisfied in an economic model. Lipsey and Lancaster's results have important implications for the understanding of, not only, trade policies but many other government policies as well.

In this section we will provide an overview of the main results and indicate some of the implications for trade policy analysis. We will then consider various applications of the theory to international trade policy issues.

First of all, one must note that economic models consist of exercises in which a set of assumptions are used to deduce a series of logical conclusions. The solution of a model is referred to as equilibrium. Equilibrium is typically described by explaining the conditions or relationships that must be satisfied in order for the equilibrium to be realized. These are called the equilibrium conditions. In economic models these conditions arise out of the maximizing behavior of producers and consumers. Thus the solution is also called an optimum.

For example, in a standard perfectly competitive model, the equilibrium conditions include, 1) output price equal to marginal cost for each firm in an industry, 2) the ratio of prices between any two goods is equal to each consumer's marginal rate of substitution between the two goods, 3) the long-run profit of each firm is equal to zero, and 4) supply of all goods is equal to demand for all goods. In a general equilibrium model, with many consumers, firms, industries and markets there will be numerous equilibrium conditions that must be satisfied simultaneously.

Lipsey and Lancaster's analysis asks the following simple question: What happens to the other optimal equilibrium conditions when one of the conditions cannot be satisfied for some reason? For example, what happens if one of the markets does not clear, i.e. supply
does not equal demand in that one market? Would it still be appropriate for the firms to set price equal to marginal cost? Should consumers continue to set each price ratio equal to their marginal rate of substitution? Or, would it be better if firms and consumers deviate from these conditions? Lipsey and Lancaster show that, generally, when one optimal equilibrium condition is not satisfied, for whatever reason, all of the other equilibrium conditions will change. Thus if one market does not clear, it would no longer be optimal for firms to set price equal to marginal cost or for consumers to set the price ratio equal to the marginal rate of substitution.

**FIRST-BEST VS. SECOND-BEST EQUILIBRIA**

Consider a small perfectly competitive open economy that has no market imperfections or distortions, no externalities in production or consumption, no public goods. An economy in which all resources are privately owned, where the participants maximize their own well-being, firms maximize profit and consumers maximize utility always in the presence of perfect information. An economy in which markets always clear, in which there are no adjustment costs or unemployment of resources.

The optimal government policy in this case is laissez-faire. With respect to trade policy the optimal policy is free trade. Any type of tax or subsidy implemented by the government under these circumstances can only reduce economic efficiency and national welfare. Thus with a laissez-faire policy the resulting equilibrium would be called first-best.. It is useful to think of this market condition as economic nirvana since there is no conceivable way of increasing economic efficiency at a first-best equilibrium.

Of course, the real world is unlikely to be so perfectly characterized. Instead markets will likely have numerous distortions and imperfections. Some production and consumption activities have externality effects. Some goods have public good characteristics. Some markets have a small number of firms, each of which has some control over the price that prevails and makes positive economic profit. Governments invariably set taxes on consumption, profit, property and assets, etc. Finally, information is rarely perfectly and costlessly available.

Now imagine again a small open perfectly competitive economy with no market imperfections or distortions. Suppose we introduce one distortion or imperfection into such an economy. The resulting equilibrium will now be less efficient from a national perspective than when the distortion was not present. In other words the introduction of one distortion would reduce the optimal level of national welfare.

In terms of Lipsey and Lancaster's analysis, the introduction of the distortion into the system would severe one or more of the equilibrium conditions that must be satisfied to obtain economic nirvana. For example, suppose the imperfection that is introduced is the presence of a monopolistic firm in an industry. In this case the firm's profit maximizing equilibrium condition would be to set its price greater than marginal cost rather than equal to marginal cost as would be done by a profit maximizing perfectly competitive firm. Since the economic optimum obtained in these circumstances would be less
efficient than in economic nirvana, we would call this equilibrium a second-best equilibrium. Second-best equilibria arise whenever all of the equilibrium conditions satisfying economic nirvana cannot occur simultaneously. In general, second-best equilibria arise whenever there are market imperfections or distortions present.

**WELFARE IMPROVING POLICIES IN A SECOND-BEST WORLD**

An economic rationale for government intervention in the private market arises whenever there are uncorrected market imperfections or distortions. In these circumstances the economy is characterized by a second-best rather than a first-best equilibrium. In the best of cases the government policy can correct the distortions completely and the economy would revert back to the state under economic nirvana. If the distortion is not corrected completely then at least the new equilibrium conditions, altered by the presence of the distortion, can all be satisfied. In either case an appropriate government policy can act to correct, or reduce the detrimental effects of the market imperfection or distortion, raise economic efficiency and improve national welfare.

It is for this reason that many types of trade policies can be shown to improve national welfare. Trade policies, chosen appropriate to the market circumstances, act to correct the imperfections or distortions. This remains true even though the trade policies themselves would act to reduce economic efficiency if applied starting from a state of economic nirvana. What happens is that the policy corrects the distortion or imperfection and thus raises national welfare by more than the loss in welfare arising from the application of the policy.

Many different types of policies can be applied even for the same distortion or imperfection. Governments can apply taxes, subsidies or quantitative restrictions. It can apply these to production, to consumption, or to factor usage. Sometimes it even applies two or more of these policies simultaneously in the same market. Some policies, like tariffs or export taxes, are designed to directly affect the flow of goods and services between countries. These are called trade policies. Other policies, like production subsidies or consumption taxes, are directed at a particular activity that occurs within the country but is not targeted directly at trade flows. These can be referred to as domestic policies.

One prominent area of trade policy research focuses on identifying the optimal policy to be used in a particular second-best equilibrium situation. Invariably this research has considered multiple policy options in any one situation and has attempted to rank order the potential policies in terms of their efficiency enhancing capabilities. As with the ranking of equilibria described above, the ranking of policy options is also typically characterized using the first-best and second-best labels.

Thus, the ideal or optimal policy choice in the presence of a particular market distortion or imperfection is referred to as a first-best policy. The first-best policy will raise national welfare, or enhance aggregate economic efficiency, to the greatest extent possible in a particular situation.
Many other policies can often be applied, some of which would be welfare-improving. If any such policy raises welfare to a lesser degree than a first-best policy, then it would be called a second-best policy. If there are many policy options which are inferior to the first-best policy, then it is common to refer to them all as second-best policies. Only if one can definitively rank three or more policy options would one ever refer to a third-best or fourth-best policy. Since these rankings are often difficult, third-best et al., policies are not commonly denoted.

**TRADE POLICIES IN A SECOND-BEST WORLD**

In a 1971 paper titled "A General Theory of Domestic Distortions and Welfare", Jagdish Bhagwati provided a framework for understanding the welfare implications of trade policies in the presence of market distortions. This framework applied the theory of the second-best to much of the welfare analysis that had been done in international trade theory up until that point. Bhagwati demonstrated the result that trade policies can improve national welfare if they occur in the presence of a market distortion and if they act to correct the detrimental effects caused by the distortion. However, Bhagwati also showed that in almost all circumstances a trade policy will be a second-best rather than a first-best policy choice. The first-best policy would likely be a purely domestic policy that is targeted directly at the distortion in the market. One exception to this rule occurs when a country is "large" in international markets and thus can affect international prices with its domestic policies. In this case, as was shown with optimal tariffs, quotas, VERs and export taxes, trade policy is the first-best policy.

Since Bhagwati’s paper, international trade policy analysis has advanced to include market imperfections such as monopolies, duopolies and oligopolies. In many of these cases it has been shown that appropriately chosen trade policies can improve national welfare. The reason trade policies can improve welfare, of course, is that the presence of the market imperfection means that the economy begins at a second-best equilibrium. The trade policy, if properly targeted, can reduce the negative aggregate effects caused by the imperfection and thus raise national welfare.

### 5.4 ARROW'S IMPOSSIBILITY THEOREM

In social choice theory, Arrow’s impossibility theorem, or Arrow’s paradox, demonstrates that no voting system can convert the ranked preferences of individuals into a community-wide ranking while also meeting a certain set of reasonable criteria with three or more discrete options to choose from. These criteria are called unrestricted domain, non-imposition, non-dictatorship, Pareto efficiency, and independence of irrelevant alternatives. The theorem is often cited in discussions of election theory as it is further interpreted by the Gibbard–Satterthwaite theorem.

The theorem is named after economist Kenneth Arrow, who demonstrated the theorem in his Ph.D. thesis and popularized it in his 1951 book *Social Choice and Individual Values*. The original paper was titled "A Difficulty in the Concept of Social Welfare". Arrow was a co-recipient of the 1972 Nobel Prize in Economics.
STATEMENT OF THE THEOREM

The need to aggregate preferences occurs in many different disciplines: in welfare economics, where one attempts to find an economic outcome which would be acceptable and stable; in decision theory, where a person has to make a rational choice based on several criteria; and most naturally in voting systems, which are mechanisms for extracting a decision from a multitude of voters' preferences.

The framework for Arrow's theorem assumes that we need to extract a preference order on a given set of options (outcomes). Each individual in the society (or equivalently, each decision criterion) gives a particular order of preferences on the set of outcomes. We are searching for a preferential voting system, called a social welfare function, which transforms the set of preferences into a single global societal preference order. The theorem considers the following properties, assumed to be reasonable requirements of a fair voting method:

Non-dictatorship
   The social welfare function should account for the wishes of multiple voters. It cannot simply mimic the preferences of a single voter.

Unrestricted domain
   (or universality) The social welfare function should account for all preferences among all voters to yield a unique and complete ranking of societal choices. Thus:

   - The voting mechanism must account for all individual preferences.
   - It must do so in a manner that results in a complete ranking of preferences for society.
   - It must deterministically provide the same ranking each time voters' preferences are presented the same way.

Independence of irrelevant alternatives (IIA)
   The social welfare function should provide the same ranking of preferences among a subset of options as it would for a complete set of options. Changes in individuals' rankings of irrelevant alternatives (ones outside the subset) should have no impact on the societal ranking of the relevant subset.

Positive association of social and individual values
   (or monotonicity) If any individual modifies his or her preference order by promoting a certain option, then the societal preference order should respond only by promoting that same option or not changing, never by placing it lower than before. An individual should not be able to hurt an option by ranking it higher.

Non-imposition
   (or citizen sovereignty) Every possible societal preference order should be achievable by some set of individual preference orders. This means that the social welfare function is surjective: It has an unrestricted target space.
Arrow's theorem says that if the decision-making body has at least two members and at least three options to decide among, then it is impossible to design a social welfare function that satisfies all these conditions at once.

A later (1963) version of Arrow's theorem can be obtained by replacing the monotonicity and non-imposition criteria with:

- **Pareto efficiency**: if every individual prefers a certain option to another, then so must the resulting societal preference order. This, again, is a demand that the social welfare function will be minimally sensitive to the preference profile.

The later version of this theorem is stronger—has weaker conditions—since monotonicity, non-imposition, and independence of irrelevant alternatives together imply Pareto efficiency, whereas Pareto efficiency, non-imposition, and independence of irrelevant alternatives together do not imply monotonicity.

**FORMAL STATEMENT OF THE THEOREM**

Let $A$ be a set of outcomes, $N$ a number of voters or decision criteria. We shall denote the set of all full linear orderings of $A$ by $L(A)$ (this set is equivalent to the set $S_{|A|}$ of permutations on the elements of $A$).

A (strict) social welfare function is a function $F : L(A)^N \to L(A)$ which aggregates voters' preferences into a single preference order on $A$. The $N$-tuple $(R_1, \ldots, R_N)$ of voter's preferences is called a *preference profile*. In its strongest and most simple form, Arrow's impossibility theorem states that whenever the set $A$ of possible alternatives has more than 2 elements, then the following three conditions become incompatible:

**unanimity, or Pareto efficiency**

If alternative $a$ is ranked above $b$ for all orderings $R_1, \ldots, R_N$, then $a$ is ranked higher than $b$ by $F(R_1, R_2, \ldots, R_N)$. (Note that unanimity implies non-imposition).

**non-dictatorship**

There is no individual $i$ whose preferences always prevail. That is, there is no $i \in \{1, \ldots, N\}$ such that $\forall (R_1, \ldots, R_N) \in L(A)^N$, $F(R_1, R_2, \ldots, R_N) = R_i$.

**independence of irrelevant alternatives**

For two preference profiles $(R_1, \ldots, R_N)$ and $(S_1, \ldots, S_N)$ such that for all individuals $i$, alternatives $a$ and $b$ have the same order in $R_i$ as in $S_i$, alternatives $a$ and $b$ have the same order in $F(R_1, R_2, \ldots, R_N)$ as in $F(S_1, S_2, \ldots, S_N)$.
INFORMAL PROOF

Based on the proof by John Geanakoplos of Cowles Foundation, Yale University.\[2\]

We wish to prove that any social choice system respecting unrestricted domain, unanimity, and independence of irrelevant alternatives (IIA) is a dictatorship.

PART ONE: THERE IS A "PIVOTAL" VOTER FOR B

Say there are three choices for society, call them A, B, and C. Suppose first that everyone prefers option B the least. That is, everyone prefers every other option to B. By unanimity, society must prefer every option to B. Specifically, society prefers A and C to B. Call this situation Profile 1.

On the other hand, if everyone preferred B to everything else, then society would have to prefer B to everything else by unanimity. So it is clear that, if we take Profile 1 and, running through the members in the society in some arbitrary but specific order, move B from the bottom of each person's preference list to the top, there must be some point at which B moves off the bottom of society's preferences as well, since we know it eventually ends up at the top.

We now want to show that, during this process, at the point when the pivotal voter n moves B off the bottom of his preferences to the top, the society's B moves to the top of its preferences as well, not to an intermediate point.

To prove this, consider what would happen if it were not true. Then, after n has moved B to the top (i.e., when voters \{1, \cdots, n - 1\} have B at the bottom and voters \{n, \cdots, \} have B at the top) society would have some option it prefers to B, say A, and one less preferable than B, say C. (If otherwise, just swap A's and C's names).

Now if each person moves his preference for C above A, then society would prefer C to A by unanimity. By the fact that A is already preferred to B, C would now be preferred to B as well in the social preference ranking. But moving C above A should not change anything about how B and C compare, by independence of irrelevant alternatives. That is, since B is either at the very top or bottom of each person's preferences, moving C or A around does not change how either compares with B. We have reached an absurd conclusion.

Therefore, when the voters \{1, \cdots, n\} have moved B from the bottom of their preferences to the top, society moves B from the bottom all the way to the top, not some intermediate point.
PART TWO: VOTER N IS A DICTATOR FOR A–C

In the second part of the proof, we show how voter \( n \) can be a dictator over society's decision between A and C. Call the case with all voters up to (not including) \( n \) having B at the top of their preferences and the rest (including \( n \)) with B at the bottom Profile 2. Call the case with all voters up through (and including) \( n \) having B at the top and the rest having B at the bottom Profile 3.

Now suppose everyone up to \( n \) ranks B at the bottom, \( n \) ranks B below A but above C, and everyone else ranks B at the top. As far as the A–B decision is concerned, this organization is just as in Profile 2, which we proved puts B below A (in Profile 2, B is actually at the bottom of the social ordering). C's new position is irrelevant to the B–A ordering for society because of IIA. Likewise, \( n \)'s new ordering has a relationship between B and C that is just as in Profile 3, which we proved has B above C (B is actually at the top). Hence we know society puts A above B above C. And if person \( n \) flipped A and C, society would have to flip its preferences by the same argument. Hence person \( n \) gets to be a dictator over society's decision between A and C.

Since B is irrelevant (IIA) to the decision between A and C, the fact that we assumed particular profiles that put B in particular places does not matter. This was just a way of finding out, by example, who the dictator over A and C was. But all we need to know is that he exists.

PART THREE: THERE CAN BE AT MOST ONE DICTATOR

Finally, we want to show that the dictator can also dictate over the A–B pair and over the C–B pair. Consider that we have proven that there are dictators over the A–B, B–C, and A–C pairs, but they are not necessarily the same dictator. However, if you take the two dictators who can dictate over A–B and B–C, for example, they together can determine the A–C outcome, contradicting the idea that there is some third dictator who can dictate over the A–C pair. Hence the existence of these dictators is enough to prove that they are the same person, otherwise they would be able to overrule one another, a contradiction.

INTERPRETATIONS OF THE THEOREM

Arrow's theorem is a mathematical result, but it is often expressed in a non-mathematical way with a statement such as "No voting method is fair", "Every ranked voting method is flawed", or "The only voting method that isn't flawed is a dictatorship". These statements are simplifications of Arrow's result which are not universally considered to be true. What Arrow's theorem does state is that a voting mechanism cannot comply with all of the conditions given above simultaneously for all possible preference orders.

Arrow did use the term "fair" to refer to his criteria. Indeed, Pareto efficiency, as well as the demand for non-imposition, seems trivial. Various theorists have suggested
weakening the IIA criterion as a way out of the paradox. Proponents of ranked voting methods contend that the IIA is an unreasonably strong criterion, which actually does not hold in most real-life situations. Indeed, the IIA criterion is the one breached in most useful voting systems.

Advocates of this position point out that failure of the standard IIA criterion is trivially implied by the possibility of cyclic preferences. If voters cast ballots as follows:

- 7 votes for A > B > C
- 6 votes for B > C > A
- 5 votes for C > A > B

then the net preference of the group is that A wins over B, B wins over C, and C wins over A: these yield rock-paper-scissors preferences for any pairwise comparison. In this circumstance, any system that picks a unique winner, and satisfies the very basic majoritarian rule that a candidate who receives a majority of all first-choice votes must win the election, will fail the IIA criterion. Without loss of generality, consider that if a system currently picks A, and B drops out of the race (as e.g. in a two-round system), the remaining votes will be:

- 7 votes for A > C
- 11 votes for C > A

Thus, C will win, even though the change (B dropping out) concerned an "irrelevant" alternative candidate who did not win in the original circumstance.

So, what Arrow's theorem really shows is that voting is a non-trivial game, and that game theory should be used to predict the outcome of most voting mechanisms. This could be seen as a discouraging result, because a game need not have efficient equilibria, e.g., a ballot could result in an alternative nobody really wanted in the first place, yet everybody voted for.

Note, however, that not all voting systems require (or even allow), as input, a strict ordering of all candidates. These systems may then trivially fail the universality criterion. Some systems may satisfy a version of Arrow's theorem with some reformulation of universality and independence of irrelevant alternatives; Warren Smith claims that range voting is such a system.

**OTHER POSSIBILITIES**

The preceding discussion assumes that the "correct" way to deal with Arrow's paradox is to eliminate (or weaken) one of the criteria. The IIA criterion is the most natural candidate. Yet there are other "ways out".

Duncan Black has shown that if there is only one agenda by which the preferences are judged, then all of Arrow's axioms are met by the majority rule. Formally, this means that
if we properly restrict the domain of the social welfare function, then all is well. Black's restriction, the "single-peaked preference" principle, states that there is some predetermined linear ordering $P$ of the alternative set. Every voter has some special place he likes best along that line, and his dislike for an alternative grows larger as the alternative goes further away from that spot. For example, if voters were voting on where to set the volume for music, it would be reasonable to assume that each voter had their own ideal volume preference and that as the volume got progressively too loud or too quiet they would be increasingly dissatisfied. In such a restricted case, Arrow's theorem does not apply: in particular, it lacks an unrestricted domain.

Indeed, many different social welfare functions can meet Arrow's conditions under such restrictions of the domain. It has been proved, however, that under any such restriction, if there exists any social welfare function that adheres to Arrow's criteria, then the majority rule will adhere to Arrow's criteria. Under single-peaked preferences, then, the majority rule is in some respects the most natural voting mechanism.

Another common way "around" the paradox is limiting the alternative set to two alternatives. Thus, whenever more than two alternatives should be put to the test, it seems very tempting to use a mechanism that pairs them and votes by pairs. As tempting as this mechanism seems at first glance, it is generally far from meeting even the Pareto principle, not to mention IIA. The specific order by which the pairs are decided strongly influences the outcome. This is not necessarily a bad feature of the mechanism. Many sports use the tournament mechanism—essentially a pairing mechanism—to choose a winner. This gives considerable opportunity for weaker teams to win, thus adding interest and tension throughout the tournament. In effect, the mechanism by which the choices are limited to two candidates is best considered as a part of the balloting system, and hence Arrow's theorem applies.

There has developed an entire literature following from Arrow's original work which finds other impossibilities as well as some possibility results. For example, if we weaken the requirement that the social choice rule must create a social preference ordering which satisfies transitivity and instead only require acyclicity (if $a$ is preferred to $b$, and $b$ is preferred to $c$, then it is not the case that $c$ is preferred to $a$) there do exist social choice rules which satisfy Arrow's requirements.

Economist and Nobel prize winner Amartya Sen has suggested at least two other alternatives. He has offered both relaxation of transitivity and removal of the Pareto principle. He has shown the existence of voting mechanisms which comply with all of Arrow's criteria, but supply only semi-transitive results.

Also, he has demonstrated another interesting impossibility result, known as the "impossibility of the Paretian Liberal". Sen went on to argue that this demonstrates the futility of demanding Pareto optimality in relation to voting mechanisms.

Advocates of approval voting consider unrestricted domain to be the best criterion to weaken. In approval voting, voters can only vote 'for' or 'against' each candidate,
Advocates of range voting also consider unrestricted domain to be the best criterion to violate—but instead of limiting voter options like approval voting, range voting increases the number of voter options beyond what Arrow's Theorem allows.

**SCALAR RANKINGS FROM A VECTOR OF ATTRIBUTES AND THE IIA PROPERTY**

The IIA property might not be satisfied in human decision-making of realistic complexity because the scalar preference ranking is effectively derived from the weighting—not usually explicit—of a vector of attributes (one book dealing with the Arrow theorem invites the reader to consider the related problem of creating a scalar measure for the track and field decathlon event—e.g. how does one make scoring 600 points in the discus event "commensurable" with scoring 600 points in the 1500 m race) and this scalar ranking can depend sensitively on the weighting of different attributes, with the tacit weighting itself affected by the context and contrast created by apparently "irrelevant" choices. Edward MacNeal discusses this sensitivity problem with respect to the ranking of "most livable city" in the chapter "Surveys" of his book *MathSemantics: making numbers talk sense* (1994).

### 5.5 THE VOTING PARADOX

The voting paradox (also known as Condorcet's paradox or the paradox of voting) is a situation noted by the Marquis de Condorcet in the late 18th century, in which collective preferences can be cyclic (i.e. not transitive), even if the preferences of individual voters are not. This is paradoxical, because it means that majority wishes can be in conflict with each other. When this occurs, it is because the conflicting majorities are each made up of different groups of individuals. For example, suppose we have three candidates, A, B, and C, and that there are three voters with preferences as follows (candidates being listed in decreasing order of preference):

Voter 1: A B C  
Voter 2: B C A  
Voter 3: C A B

If C is chosen as the winner, it can be argued that B should win instead, since two voters (1 and 2) prefer B to C and only one voter (3) prefers C to B. However, by the same argument A is preferred to B, and C is preferred to A, by a margin of two to one on each occasion. The requirement of majority rule then provides no clear winner.

Also, if an election were held with the above three voters as the only participants, nobody would win under majority rule, as it would result in a three way tie with each candidate getting one vote. However, Condorcet's paradox illustrates that the person who can reduce alternatives can essentially guide the election. For example, if Voter 1 and Voter 2
choose their preferred candidates (A and B respectively), and if Voter 3 was willing to drop his vote for C, then Voter 3 can choose between either A or B - and become the agenda-setter.

When a Condorcet method is used to determine an election, a voting paradox among the ballots can mean that the election has no Condorcet winner. The several variants of the Condorcet method differ on how they resolve such ambiguities when they arise to determine a winner. Note that there is no fair and deterministic resolution to this trivial example because each candidate is in an exactly symmetrical situation.

The phrase "Voter's Paradox" is sometimes used for the paradox of voting, the rational choice theory prediction that voter turnout should be 0.

### 5.6 ARROW-DEBREU MODEL

The Arrow-Debreu model, also referred to as the Arrow-Debreu-McKenzie model suggests that, should the assumptions made about the conditions under which it works hold (i.e. convexity, perfect competition and demand independence), then there will be a set of prices such that aggregate supplies will equal aggregate demands for every commodity in the economy.

The model (ADM model) is the central model in the General (Economic) Equilibrium Theory and often used as a general reference for other microeconomic models. It is named after Kenneth Arrow, Gerard Debreu and Lionel W. McKenzie.

Compared to earlier models, the Arrow-Debreu model radically generalized the notion of a commodity, differentiating commodities by time and place of delivery. So, for example, 'apples in New York in September' and 'apples in Chicago in June' are regarded as distinct commodities. The Arrow-Debreu model applies to economies with maximally complete markets, in which there exists a market for every time period and forward prices for every commodity at all time periods and in all places.

The ADM model is one of the most general models of competitive economy and is a crucial part of general equilibrium theory, as it can be used to prove the existence of general equilibrium (or Walrasian equilibrium) of an economy. Once we can prove the existence of such an equilibrium, it is possible to show that it is unique.

### APPLICATIONS TO FINANCE THEORY

The Arrow-Debreu model specifies the conditions of perfectly competitive markets.

In financial economics the term Arrow-Debreu is most commonly used with reference to an Arrow-Debreu security. A canonical Arrow-Debreu security is a security that pays one unit of numeraire if a particular state of the world is reached and zero otherwise (a so called "state price"). As such, any derivatives contract whose settlement value is a
function on an underlying whose value is uncertain at contract date can be decomposed as linear combination of Arrow-Debreu securities.

The concept of Arrow-Debreu security is a good pedagogical tool to understand pricing and hedging issues in derivatives analysis. Its practical use in financial engineering, however, has turned out to be very limited, especially in the multi-period or continuous markets.

The Black Scholes analysis and its extensions, despite their strongly formulated and somewhat questionable assumptions, have proven more successful in practice and have contributed directly to the exponential growth in the size of the global derivatives industry over the past 30 years.

Activity 5

1. Discuss the underlying assumptions in the theory of second best. What do you understand by theory of second best solution?
2. Explain Arrow’s impossibility theorem and interpretations of the theorem.
3. Write a brief note on the voter’s paradox.
4. What is an Arrow Debreu model? Discuss its relevance in context of different functional areas.

5.7 SUMMARY

Theory of second best is the concept introduced by a Canadian economist Richard Lipsey and Australian-American economist Kelvin Lancaster and considered as an economics concept that if two or more requirements for achieving a most desirable economic situation cannot be satisfied, a concerted attempt to satisfy the requirements that can be met is not necessarily the second best option and may not be beneficial.

Arrow's impossibility theorem is also known as Arrow's theorem or Arrow's possibility theorem, after the economist Kenneth Arrow. The theorem addresses the question of whether there is a reliable way of determining the aggregate interest of society from the interests of individuals. Such a method would, in principle, provide policy makers with a mechanism for social choice. The theorem states that no social choice mechanism exists that satisfies certain minimal reasonable conditions

5.8 FURTHER READINGS

- Arrow, K.J., "A Difficulty in the Concept of Social Welfare", Journal of Political Economy