Practical Guide To Contemporary Economics

Yuri Yevdokimov

Yuri Yevdokimov

Practical Guide To Contemporary Economics

Practical Guide To Contemporary Economics

ISBN 978-87-403-0238-7

Contents

1 The Issues and Methods of Economics

Key concepts discussed in this chapter: economics as a social science, positive economics, normative economics, economic model, economic theory, economic way of thinking, production possibilities frontier (PPF), opportunity cost, productive efficiency, allocative efficiency, Pareto efficiency, specialization, absolute advantage, comparative advantage

1.1 Economics as a science

As a matter of fact, all key economic questions and problems arise because human wants exceed the resources available to satisfy them. Our inability to satisfy all our wants is called scarcity. Faced with scarcity we must make choices. We must choose the available alternatives. Therefore, economics as a science can be defined as follows:

Economics: Social science that studies the choices that individuals, businesses, government and the entire society make as they cope with scarcity

The subject matter of economics is divided into two main components:

- Microeconomics
- Macroeconomics

Microeconomics is the study of the choices that *individual* economic agents make, the interaction of these choices, and the influence that governments exert on these choices. The key word in understanding microeconomics is *individual*.

Macroeconomics is the study of the *aggregate* (total) effects on the national economy and the global economy of the choices that individuals, households, businesses and governments make. The key word in understanding macroeconomics is *aggregate*.

The economic choices that individual economic agents such as individuals, households, businesses and governments make and the interactions of those choices answer the following three major microeconomic questions:

- *What* goods and services should be produce and in what quantities?
- *How* are goods and services produced?
- *For whom* are the various goods and services produced?

Microeconomic theory will help us answer these questions. In answering these questions, economists generally find that individuals want more than is available, and that is why the problem of scarcity arises.

Being social scientists, economists try to discover how the economic world works. In doing so, they distinguish between two types of statements or two types of economic analysis:

- Positive statements or positive economic analysis
- Normative statements or normative economic analysis

Positive statement is a proposition that can be settled by an appeal to facts. It is testable, either true or false and is associated with the statement "what it is" without any policy recommendations. Example of the statement is "Air pollution in large cities is high". Positive analysis is a value-free approach to inquiry.

Normative statement is associated with the proposition "what ought to be". It is not based on facts and usually points to some policy recommendations. Example of the statement is: "We ought to clean up our environment". Normative analysis is based on value judgment. In general, the following words are good indicators of a normative statement: ought to, have to, should, and must.

1.2 Modeling in economics

The task of economic science is first to discover and catalogue positive statements that are consistent with what we observe in the world and that enable us to understand how the economic world works. This task can be broken into three steps:

1. Observing and measuring
2. Model building
3. Testing

Observing and measuring result in economists keeping track of huge amounts of economic data. These data are needed to build economic models.

Economic model: A description of some aspects of the economic world that includes only those features that are needed for the purpose at hand

At large, economic model is an abstraction (simplification) of the real world. It is composed of a number of assumptions and relationships between economic variables from which conclusions and/or predictions are deducted.

Assumptions are apriori statements or what economists call stylized facts. They are accepted without any proof. Assumptions are an important component of an economic model. They help economists create the required environment to make use of mathematical relationships.

In general, relationships take on a form of equations and/or inequalities that involve economic variables, constants and parameters. A set of equations and inequalities defines the structure of a model. Equations in economic models are of 3 types: definitional equations, behavioral equations and conditional equations.

A *definitional equation* is identity which is equality between two alternative expressions that have exactly the same meaning.

A *behavioral equation* specifies the way in which a given variable behaves in response to changes in other variables.

A *conditional equation* states a requirement to be satisfied.

Mathematical way of presenting economic models is not the only one. At large economic models can be presented in various forms such as:

- by words = logical or verbal models
- by tables = statistical models
- by graphs = graphical models
- by mathematical expressions = mathematical models

All of them are useful ways to analyze economic information. For instance, economists make extensive use of graphs because they are very illustrative and help one better absorb specifics of underlying economic processes.

And finally, after a model is constructed it has to be tested because the model might conflict with the existing data. A model that has repeatedly passed the test of corresponding with real-world data is the basis of an economic theory.

Economic theory: A generalization that summarizes what we understand about economic choices that people make and the economic performance of industries and nations

Theories are then usually used to address normative aspects of economic analysis.

So, economics is a social science that studies the allocation of scarce resources to satisfy unlimited wants. This involves analyzing the production, distribution, trade and consumption of goods and services. Economics is said to be positive when it attempts to explain the consequences of different choices given a set of assumptions or a set of observations, and normative when it prescribes that a certain action should be taken.

1.3 Economic way of thinking

The economic way of thinking assumes that a typical response to an economic problem of scarcity is *rational behavior.* This way of thinking is somewhat different from the natural sciences'. Five core ideas summarize it, and these ideas form the basis of all microeconomics models. They are:

1. People make rational choices by comparing costs and benefits
2. Cost is what you must give up to get something
3. Benefit is what you gain when you get something and is measured by what you are willing to give up to get it
4. Rational choice is made on a margin
5. People respond to incentives

In general, rational choice is a choice that uses the available resources most effectively to satisfy the wants of an economic agent making the choice.

Cost in economics is viewed in terms of *opportunity cost*. Opportunity cost is the cost of something you must give up to get what you want. The concept arises because of scarcity: If you use resource in some specific way then you actually forgo opportunity to use the scarce resource in any other way. Formal definition of opportunity cost is

Opportunity cost: The value of the most valuable alternative that was not chosen

The benefit of something is the gain or pleasure that it brings. Economists measure benefits by what a person is willing to give up getting it.

A choice on the margin is a choice that is made by comparing all the relevant alternatives systematically and incrementally. Therefore, instead of just costs and benefits we have to take into consideration marginal costs and marginal benefits:

Marginal cost: The cost that arises from a one-unit increase in an activity. The marginal cost of something is what you must give up to get one more unit of it

Marginal benefit: The benefit that arises from a one-unit increase in an activity. The marginal benefit of something is measured by what you are willing to give up getting one more unit of it

People make rational choices and use our scarce resources in the way that makes them as well off as possible when they take those actions for which marginal benefits exceed or equals marginal costs.

In making their choices, people respond to incentives. An incentive is an inducement to take a particular action. The inducement can be a reward in the form of an increase in benefit or a decrease in cost. On the other hand, an incentive can be a punishment with the opposite result. In general, a change in marginal benefit or a change in marginal cost brings a change in the incentives that people face and eventually leads them to change their actions.

Therefore, in order to make a rational choice, we must determine the costs and benefits of the alternatives.

1.4 Production possibilities frontier (PPF)

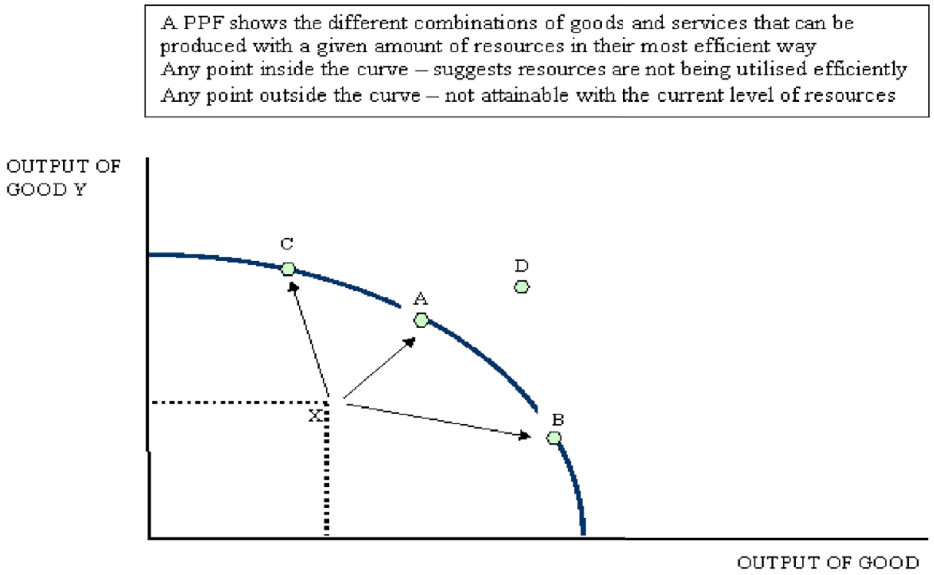
Formal definition of the production possibilities frontier is:

Production Possibilities Frontier: The boundary between the combinations of goods and services that can be produced and the combinations that cannot be produced, given the available factors of production and state of technology

First of all, it is necessary to define what economists mean by *factors of production* also called inputs of production or resources. At large it is possible to define the following factors of production (inputs of production, resources):

1) Labor – a set of human efforts
2) Physical capital – a set of capital goods (assets) such as equipment, machinery, tools, buildings, structures, etc.
3) Land and natural resources such as, for example, oil, coal, natural gas, etc.
4) Entrepreneurship – a set of managerial skills

In order to illustrate the limits of production, we focus our attention on two goods only. The following graph illustrates the concept of the PPP: In that graph, X-axis measures quantity of good X in some units while Y-axis measures quantity of good Y.



It turns out that the PPF is a valuable tool for illustrating the effects of scarcity and its consequences. It puts three features of production possibilities in focus:

- Attainable versus unattainable combinations: Combinations of two goods represented by points X, A, B and C or in general any point inside or on the PPF are attainable while combination associated with point D is unattainable given the economy's resources
- Efficient versus inefficient combinations: Combinations A, B and C or any point on the PPF are efficient combinations while combinations inside the PPF like point X are inefficient
- Trade-offs: Movement from p. C to point A shows the trade-off between goods X and Y since an increase in X corresponds to a decrease in Y and vice versa

Full employment of all economy's resources occurs when all the available factors of production are being used up which is represented by points A, B, C on the above graph. Therefore, combinations inside the PPF are associated with unemployment of some or all resources.

1.5 PPF and opportunity costs

The concept of opportunity cost is often expressed by economists in the form of the following expression: "There is no such a thing as free lunch" which means that there is always cost involved. However, if the economy produces inside the PPF (point X), then it is possible to increase production of one good without giving up the other: Consider movement from point X to point A on the above graph. When production takes place on the PPF, we face a trade-off; however, we do not face a trade-off if we produce inside the PPF.

Trade-off or movement along the PPP is always associated with opportunity costs. PPF helps us define the opportunity costs associated with production of both goods numerically. Technically opportunity cost is a ratio – the change in the quantity of one good divided by the change in the quantity of the other good or in our case we can express opportunity cost of good X as follows:

$$OC_X = \frac{\Delta Y}{\Delta X}$$

It means that opportunity cost of good X is equal to the ΔY quantity of good Y given up per extra unit of good X gained. It is associated with movement along the PPF like from p. A to p. B. Graphically it is slope of the PPF.

In turn, opportunity cost of good Y is inverse of the opportunity cost of good X or

$$OC_Y = \frac{1}{OC_X}$$

and it is associated with movement along the PPF like from p. B to p. A. In general, when we move along the PPF in any direction extra unit of one good requires more and more of the other good being sacrificed which is reflected in the bowed out shape of the PPF.

As intermediate conclusion:

- negative slope of the PPF reflects trade-off between two goods
- slope of the PPF at any point shows opportunity cost of the good on the horizontal axis
- opportunity cost of one good is the inverse of the opportunity cost of the other
- opportunity cost of producing a good increases with an increase in the quantity of the good which results in a concave (bowed outward) PPF

1.6 Economic efficiency

In economics, *efficiency* occurs when we produce the quantities of goods and services that people value the most. Resource use is efficient when we cannot produce more of a good or service without giving up some of another good or service that people value more highly. These two statements characterize economic efficiency as *allocative* efficiency and *productive* efficiency:

Productive efficiency: A situation in which an economy produces the maximum output with given technology and resources; it cannot produce more of one good or service without producing less of some other good or service

It means that under productive efficiency, production takes place on the PPF.

Allocative efficiency: The most highly valued combination of goods and services on the PPF

Given the above definition we can state that all combinations on the PPF achieve productive efficiency. Each of these combinations, however, is associated with specific distribution. Only one of them achieves allocative efficiency or only one combination is the most highly valued by people – the consumers. In order to find this combination, we need to know the value of each available combination. We can express it in terms of marginal benefits people receive from consumption.

In general, the more we have of any good or service, the smaller is our marginal benefit from extra unit of it which is known as *the principle of diminishing (decreasing) marginal benefit*. Mathematically the principle of diminishing marginal benefit implies that marginal benefit is a decreasing function of the quantity of a good or service consumed.

As previously discussed, the bowed out shape of the PPF is due to the fact that marginal opportunity cost or just marginal cost is an increasing function of quantity. In order to achieve allocative efficiency, we must compare the marginal benefit of a good or service *MB* with its marginal cost *MC*. The point when $MC = MB$ is the point of allocative efficiency. In the above discussed case of two goods, this condition with respect to one good (usually good X) coupled with the PPF for two goods X and Y produces optimal combination of the two goods that is allocatively efficient.

It should be pointed out that allocative efficiency is a broader and deeper concept than productive efficiency since it is associated with the so-called Pareto efficiency:

Pareto efficiency: An allocation is Pareto efficient if there is no other allocation in which some other individual is better off and no individual is worse off

1.7 Specialization, absolute and comparative advantage

Producers can produce several goods or they can concentrate on producing one good and then exchange some of their own good for those produced by others. The latter is called specialization:

Specialization: Concentrating on the production of only one good

The same can be said about nations and regions. According to economic theory, people (nations, regions) have to produce the good in which they have *comparative advantage* defined as follows:

Comparative advantage: The ability of a person to perform an activity or produce a good or service at a lower opportunity cost than someone else

This concept is usually compared to the concept of *absolute advantage*:

Absolute advantage: When one person is more productive than another person in several or even all activities

It turns out that it is possible to have absolute advantage in producing all goods and services while it is impossible to have comparative advantage in all goods and services. Moreover, it appears to be that it is beneficial for a producer (nation, region) to find its comparative advantage, specialize on production of a good or service according to it and then exchange it for other goods and services.

The concept of comparative advantage is the most powerful in the context of international trade. Usually trading countries have different absolute advantages in producing goods. For example (Lee, 1999), suppose that there are only two goods, cars and computers, and one productive resource (input of production) which is some composite of land, labor, and capital. Assume also that producing 100 cars requires 2 units of the productive resource (PR) in Country 1 and 4 units in Country 2, and producing 1,000 computers requires 3 units of PR in Country 1 and 4 in Country 2. This information is summarized in the following table:

	Country 1	Country 2
100 cars	2	4
1,000 computers	3	4

As seen from the table, Country 1 has an absolute advantage in producing both cars and computers. It may seem that Country 1 can realize no gain by trading with Country 2. Why not produce both cars and computers here, in Country 1? The answer to this question is: Because it *costs* more to produce computers in Country 1 than in Country 2.

All costs are opportunity costs. The cost of producing computers is the cars that *could* have been produced. Using the three units of PR required to produce 1,000 computers in Country 1 requires sacrificing the production of 150 cars. Using the four units of PR required to produce 1,000 computers in Country 2 requires sacrificing only 100 cars. So even though Country 1 has an absolute advantage in producing computers, Country 2 has a comparative advantage.

Compared to what has to be sacrificed, Country 2 produces computers for only two-thirds as much as it costs in Country 1. However, Country 1 has a comparative advantage over Country 2 in the production of cars. Producing 100 cars here costs 666 computers, while producing 100 cars in Country 2 costs 1,000 computers. Clearly Country 1 benefits from specializing in cars, which it produces more cheaply than Country 2, and trading with Country 2 for some of the computers it produces more cheaply.

If, for example, Country 1 produced both cars and computers it might devote 70 units of PR to car production and 30 units to computer production, yielding 3,500 cars and 10,000 computers. If Country 2 produced both products, it might devote 56 units of PR to car production and 24 to computer production, yielding 1,400 cars and 6,000 computers. On the other hand, by specializing in their comparative advantages, Country 1 can produce 5,000 cars and Country 2 can produce 20,000 computers, or a total of 100 additional cars and 4,000 additional computers. Country 1 could trade 1,450 cars to Country 2 for 12,500 computers and have 50 additional cars (3,550) and 2,500 more computers (12,500), while Country 2 would have 50 more cars (1,450) and 1,500 more computers (7,500). It implies that trade is productive since it generates more output of both products.

2 Demand and Supply

Key concepts discussed in this chapter: market, quantity demanded, the Law of Demand, demand schedule, demand curve, inverse demand, market demand, horizontal summation, substitutes, complements, quantity supplied, the Law of Supply, supply schedule, supply curve, market equilibrium, point price elasticity, midpoint price elasticity, cross-price elasticity, total revenue, income elasticity of demand, normal good, inferior good

2.1 Demand as a function

Demand and supply are two sides of a market. Formal definition of a market as an economic institution is as follows:

Market: Any arrangement that brings buyers (demanders) and sellers (suppliers) together

Any market has two sides, consumption and production. Demand is associated with consumption. Therefore, demand summarizes behavior of buyers. In studying the behavior of buyers, we have to define some economic variables.

The quantity demanded: The amount of any good, service or resource that people are willing and able to buy during a specific period at a specified price

The quantity demanded is measured in units per time. Many things influence buying plans of consumers, and the most important of them is price. That is why we look first at the relationship between the quantity demanded and the price. In order to study this relationship, we keep all other influences on buying plans the same, and we ask the following question: How, other things being equal, does the quantity demanded of a good change as its price varies? The Law of Demand provides the answer:

The Law of Demand: Other things remaining the same, if the price of a good or service rises, the quantity demanded of that good or service decreases; and if the price of a good/service falls, the quantity demanded of that good/service increases

Why does the quantity demanded increase if the price falls, all other things being equal? Faced with a limited budget, people always have an incentive to find the best deals they can. If the price of one item falls and the prices of all other items remain the same, the item with the lower price is a better deal than it was before. So people buy more of this item. Here it is necessary to understand the difference between relative price and absolute (money) price.

Relative price: Any commodity's price in terms of another commodity

Absolute (money) price: The actual price that you pay in dollars and cents for any good or service at any point in time

The main thing about this difference is: In microeconomic analysis, only relative prices matter. Therefore, only changes in relative prices affect consumer decisions.

As the Law of Demand states, demand is the relationship between the quantity demanded and the price of a good/service when all other influences on buying plans remain the same. It means that the quantity demanded is one quantity at one price. In turn, demand is a list of quantities at different prices, as illustrated by a demand schedule and a demand curve.

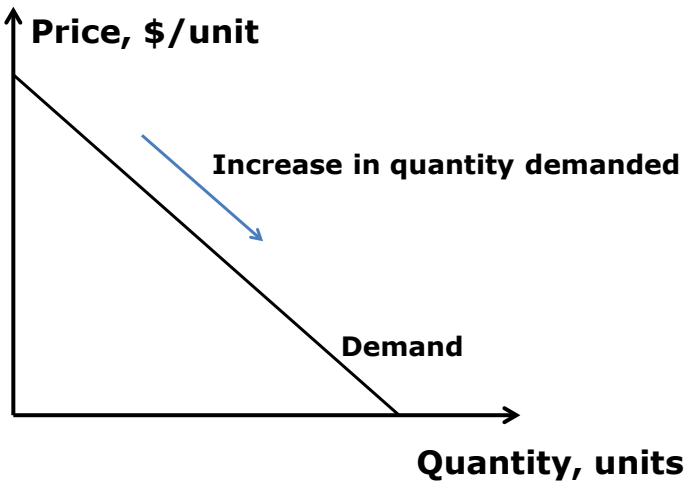
Demand schedule: A list of the quantities demanded at each different price when all other influences on buying plans remain the same

Demand curve: A graph of the relationship between the quantities demanded of a good/service and their prices when all other influences on buying plans remain the same

Demand schedule is usually presented in the form of a table as follows

Price of a good/service, $/unit	Quantity of a good/service, units/time

while demand curve is represented by a down-sloping curve as shown below:



Please note, that while the Law of Demand states the quantity demanded as dependent variable and the price as independent variable, the graph of demand as a function (the demand curve) shows the price as dependent variable on vertical axis with the quantity demanded as independent variable on horizontal axis. Therefore, mathematically the demand curve is *inverse demand*.

The difference between these two specifications can be understood from the following simple mathematical interpretation:

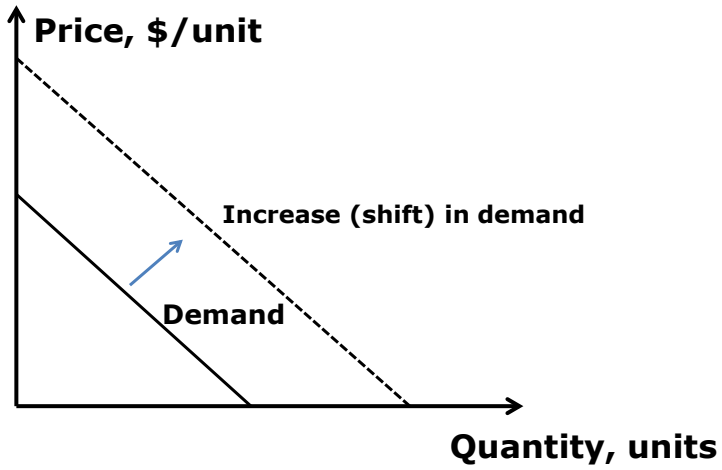
$$Q = a - bP$$

is an example of demand as a linear function with a and b being parameters (coefficients), Q as the quantity demanded and P as the price. If we re-arrange this equation with P on the left hand side, which means we solve it for P, we end up with inverse demand which in this case is

$$P = \frac{a}{b} - \frac{1}{b}Q$$

Remember, the Law of Demand produces the demand function while its graphical interpretation is given by the inverse demand function!

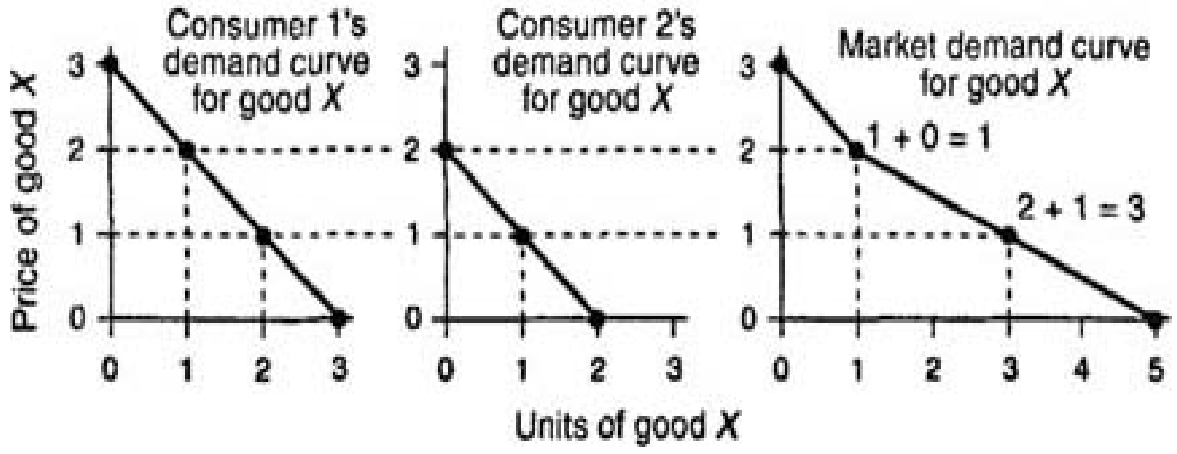
Next important point is: An increase in quantity demanded is given by the movement along the demand curve as shown above while an increase in demand is given by the upward, parallel shift in the whole curve as shown below:



One point on the demand curve corresponds to one row in the demand schedule.

2.2 Individual demand versus market demand

The individual consumer's demand for a particular good, let us call it good X, satisfies the Law of Demand and is depicted by a downward-sloping individual demand curve as already discussed. The individual consumer, however, is only one of many participants in the market for good X. The market demand curve for good X includes the quantities of good X demanded by all participants in the market for good X. The market demand curve is found by taking the *horizontal summation* of all individual demand curves. For example, suppose that there were just two consumers in the market for good X, Consumer 1 and Consumer 2. These two consumers have different individual demand curves corresponding to their different preferences for good X. The two individual demand curves and the resulting market demand curve are depicted in the following diagram:



Therefore, market demand is the sum of the demands of all the buyers in the market. At a given price, the quantity demanded on the market demand curve equals the sum of quantities demanded on the individual demand curves.

2.3 Determinants of demand

The demand curve shows how the quantity demanded changes when the price changes but all other influences on buying plans remain the same. When the price changes (remember, relative price!), we call the resulting change in buying plans a change in the quantity demanded, and we illustrate this change by a movement along the demand curve. As already mentioned, the price is the most important determinant of demand. However, it is not the only one.

When any influence on buying plans changes other than the price of a good/service, the demand curve shifts upwards – an increase in demand, or downwards – a decrease in demand.

The following is a list of major influences (determinants) on buying plans that increase/decrease demand:

- Preferences
- Income
- Prices of related goods and services
- Number of buyers
- Expectations

Consumer preferences or tastes influence demand. When preferences change, the demand for a good or services changes as well: Favorable change leads to an increase in demand, unfavorable change leads to a decrease

A rise in a person's income leads to an increase in demand (upward shift in demand curve), a fall leads to a decrease in demand for normal goods. Goods whose demand varies inversely with income are called *inferior goods*, which are explained in more detail at the end of this chapter.

A change in the price of one good or service can bring a change in the demand for a related good. Related goods are either substitutes or complements:

Substitute: A good that can be consumed in place of another good

The demand for a good increases if the price of one of its substitutes rises and the demand for a good decreases if the price of one of its substitutes falls. That is, the demand for a good and the price of one of its substitutes move in the same direction.

Complement: A good that is consumed with another good

The demand for a good decreases if the price of one of its complements rises; the demand for a good increases if the price of one of its complements falls. That is, the demand for a good and the price of one of its complements move in opposite direction.

The number of buyers affects demand as follows: the more buyers lead to an increase in demand; fewer buyers lead to a decrease.

Expected future income and prices influence demand as well. Future price: Consumers' current demand increases if they expect higher future prices; their demand decreases if they expect lower future prices. Future income: Consumers' current demand increases if they expect higher future income; their demand decreases if they expect lower future income.

2.4 Supply as a function

Very similar to the demand, we can define the quantity supplied, the Law of Supply, supply schedule and supply curve as follows:

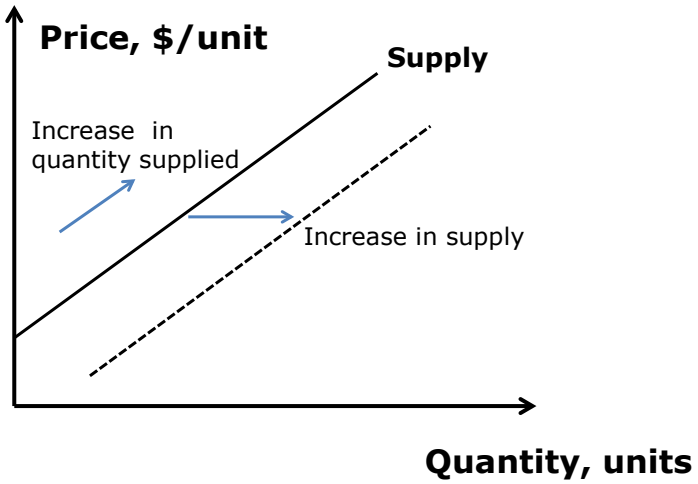
Quantity supplied: The amount of any good, service or resource that people are willing and able to sell during a specified period at a specified price

The Law of Supply: Other things remaining the same, if the price of a good or service rises, the quantity supplied of that good or increases; and if the price of a good or service falls, the quantity supplied of that good or service decreases

Supply schedule: A list of the quantities supplied at each different price when all other influences on selling plans remains the same

Supply curve: A graph of the relationship between the quantity supplied of a good/service and its price when all other influences remain the same

Hence, supply reflects a positive relationship between the quantity of a good or service supplied and the price. Again, by observing a seller's (producer's) behavior we can collect data on the quantity of a good/service supplied and its price, organize the data in a table and plot the relationship graphically. Graphically linear supply is represented by an upward sloping straight line which is actually the inverse supply in the way it was explained previously with regards to demand:



Increase/decrease in the quantity supplied corresponds to movements along the supply curve while increase/decrease in supply corresponds to the shifts in the whole line: An increase is associated with the rightward shift and a decrease is associated with the leftward shift.

2.5 Individual supply and market supply

Again we can use the analogy with demand. Above individual supply was presented graphically. Market supply is the sum of the supplies of all the sellers (producers) in the market. Again, in order to derive the market supply we have to horizontally sum up individual supply curves. The procedure is similar to the one we used to derive the market demand: In order to obtain a point on the market supply, we pick price from the vertical axis and sum up quantities that individual sellers are willing and able to supply at that price; we repeat this procedure for all possible prices.

2.6 Determinants of supply

As already explained, if the price changes, we call the resulting influence on selling plans the change in the quantity supplied, and we illustrate this change by a movement along the supply curve. Similar to the demand, although price is the most important determinant of the supply, it is not the only one. The following is a list of major influences (determinants) on selling plans that increase/decrease supply:

- Prices of resources or inputs (factors) of production
- Prices of related goods and services
- Technology (productivity)
- Number of sellers
- Expectations

Supply shifts if the price of a resource (for example oil) or other input (for example labor) used to produce a good or service changes. The reason is that resource and input prices influence the cost of production. Therefore, if the wage rate goes up, then the cost of production increases and a producer will produce less at every price which results in a leftward shift of the supply curve. A decrease in input prices shifts supply to the right.

A change in the price of one good can bring a change in the supply of another related good. Related goods can be classified as substitutes or complements in production:

Substitute in production: A good that can be produced in place of another

The supply of a good decreases if the price of one of its substitutes in production rises, and supply of a good increases if the price of one of its substitutes falls. That is, the supply of a good and the price of one of its substitutes move in opposite directions.

Complement in production: A good that is produced along with another good

The supply of a good increases if the price of one of its complements in production rises; the supply of a good decreases if the price of one of its complements in production falls. That is, the supply of a good and the price of one of its complement move in the same direction.

An increase in productivity due to technological advances lowers the cost of production and increases supply – rightward shift in supply

The greater the number of sellers (producers) in a market, the larger is supply – rightward shift in supply

Expectations about future prices have a big influence on supply today. Expectations of high future prices decrease the today's supply which results in a leftward shift in supply and vice versa.

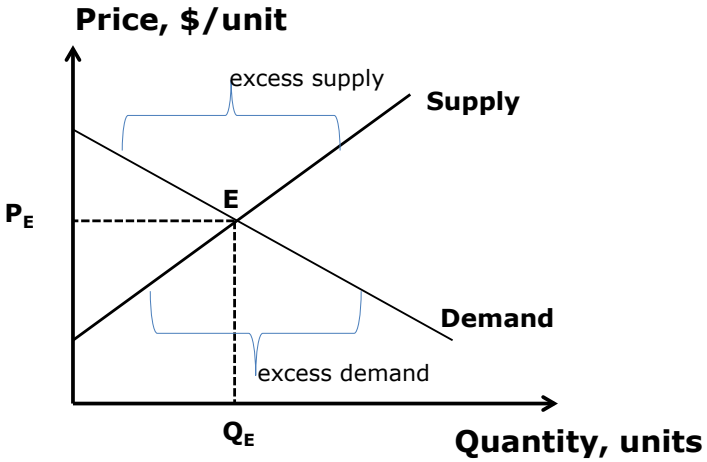
2.7 Market equilibrium

Market equilibrium occurs when the quantity demanded equals the quantity supplied – when buyers' and sellers' plans are consistent.

Equilibrium price, P_E: The price at which the quantity demanded equals the quantity supplied

Equilibrium quantity, Q_E: The quantity bought and sold at the equilibrium price

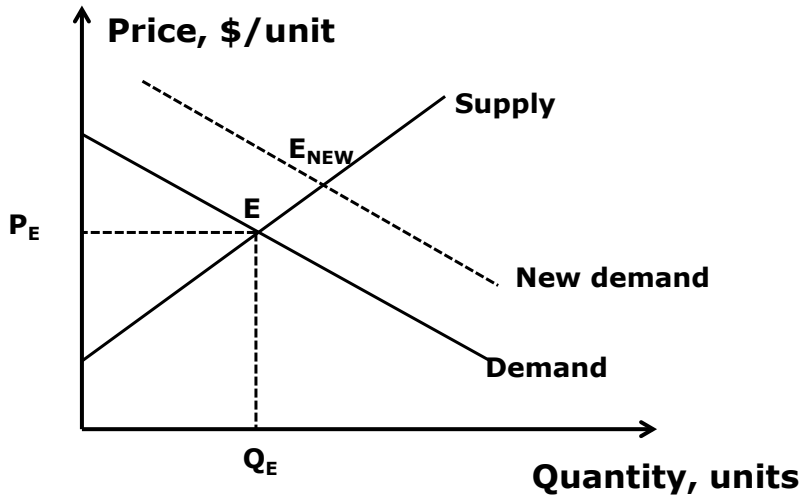
The following graph illustrates these concepts:



So, the market equilibrium occurs where the demand curve and the supply curve intersect. At equilibrium neither buyers (consumers) nor seller (producers) can improve their positions by changing either the price or the quantity.

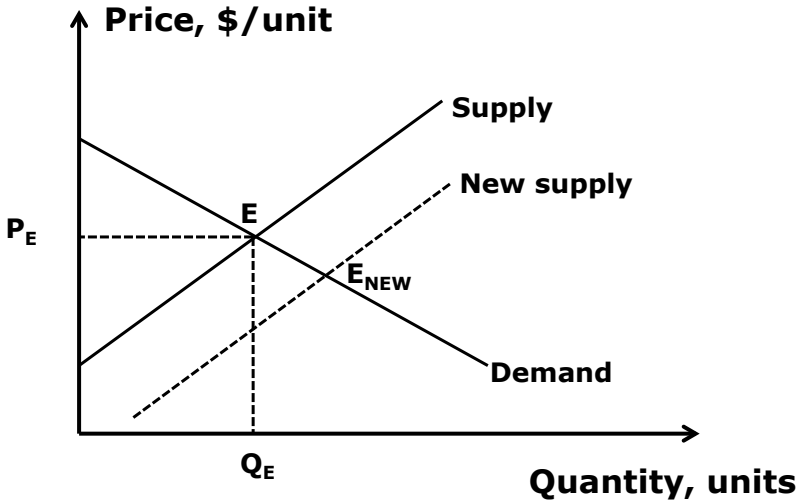
When equilibrium is disturbed, market forces restore it. Price is the regulator that pulls the market towards equilibrium. If the price is above the equilibrium price, there is a surplus or excess supply – the quantity supplied exceeds the quantity demanded. When there is a surplus, the price falls to restore the equilibrium. If the price is below the equilibrium price, there is a shortage or excess demand – the quantity demanded exceeds the quantity supplied. When there is a shortage, the price rises to restore equilibrium.

Markets are constantly hit by events that change demand and supply and bring change in the price and quantity. Some events change only demand, some change only supply and some change both demand and supply. For example, if demand increases due to a change in the influences on buying plans (e.g. change in tastes), the demand curve shifts to the right, and both equilibrium price and quantity increase as shown below:



Consequently, if demand decreases, the demand curve shift to the left, and both equilibrium price and quantity decrease.

If supply increases due to a change in the influences on selling plans (e.g. new technology), the supply curve shifts to the right, equilibrium price falls, but equilibrium quantity rises as shown below:



Consequently, if supply decreases, the supply curve shifts to the left, equilibrium price rises but equilibrium quantity falls.

If both demand and supply change simultaneously, the result is not so obvious, and it depends on the magnitude of the specific changes in demand and supply.

2.8 Price elasticity of demand

Changes in demand and supply bring changes in equilibrium prices and quantities. But by how much do prices and quantities change? *Elasticity* of demand (supply) is a powerful tool to predict the magnitudes of price and quantity changes.

Price elasticity of demand: A measure of the extent to which the quantity demanded of a good changes when the price of the good changes and all other influences on buyers' plans remain the same

In its most general form, the price elasticity of demand is a percentage change in quantity demanded due to a percentage change in the price, and it shows responsiveness or sensitivity of the quantity demanded to changes in the price. Mathematically it can be derived as follows:

Percentage change in price = $\frac{P_N - P_O}{P_O} \times 100\%$

where P_N is the new price and P_O is the old price.

Percentage change in quantity = $\frac{Q_N - Q_O}{Q_O} \times 100\%$

where Q_N is the new quantity and Q_O is the old quantity.

Finally, the price elasticity of demand is

$$e_D = [\frac{Q_N - Q_O}{Q_O} \times 100\%] / [\frac{P_N - P_O}{P_O} \times 100\%] = \frac{\Delta Q}{\Delta P} \times \frac{P_O}{Q_O}$$

where

$\Delta Q = Q_N - Q_O$ is the change in quantity

$\Delta P = P_N - P_O$ is the change in price

This is known as *point price elasticity of demand*. The above interpretation of elasticity can be applied if the change in price is small, usually less than 5%. It is so because the price elasticity of demand in the above formula depends on the initial price or rather on the ratio $\frac{P_O}{Q_O}$.

In order to avoid this deficiency, the *midpoint* method is applied. We have to derive average price and average quantity given the initial price and quantity and the new price and quantity:

$$\text{Percentage change in price} = \frac{P_N - P_O}{(P_N + P_O)/2}$$

$$\text{Percentage change in quantity} = \frac{Q_N - Q_O}{(Q_N + Q_O)/2}$$

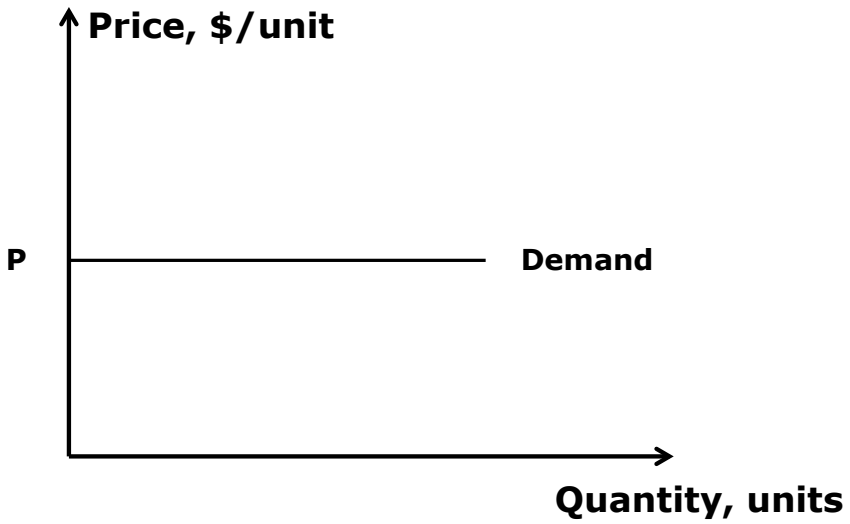
Mathematically, using definition of the price elasticity of demand as the percentage change in quantity demanded over percentage change in price, the following formula arises:

$$e_D = \frac{\Delta Q}{\Delta P} \times \frac{(P_O + P_N)}{(Q_O + Q_N)}$$

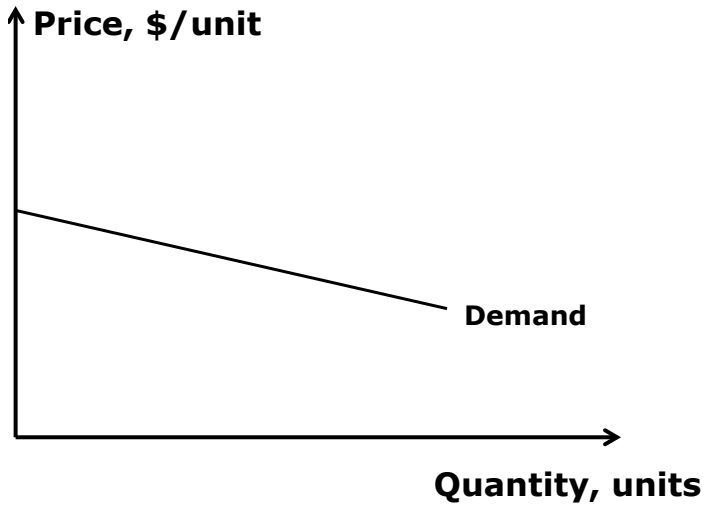
Normally this formula should be used whenever price change is greater than 5%. Note that the price elasticity of demand has a negative sign since the percentage change in price and the percentage change in quantity move in opposite directions due to the Law of Demand.

With respect to the value of the price elasticity of demand, it is possible to identify the following 5 specific cases:

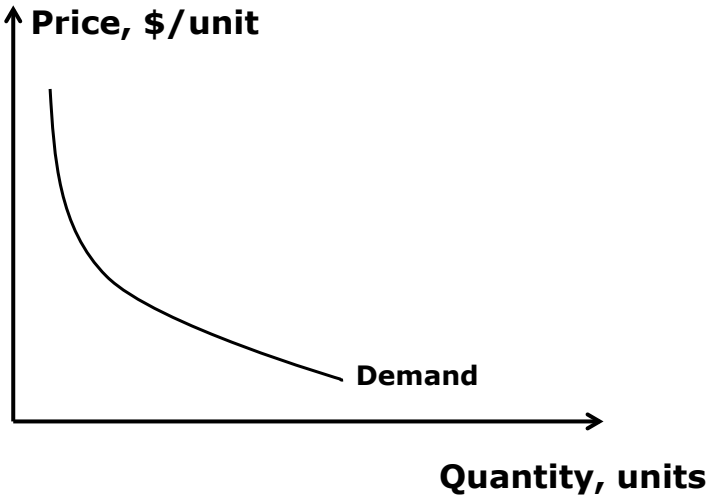
1. *Perfectly elastic demand:* Price elasticity of demand equals infinity. It happens if the quantity demanded changes by a very large percentage in response to almost zero change in price. Graphically it implies a horizontal demand:



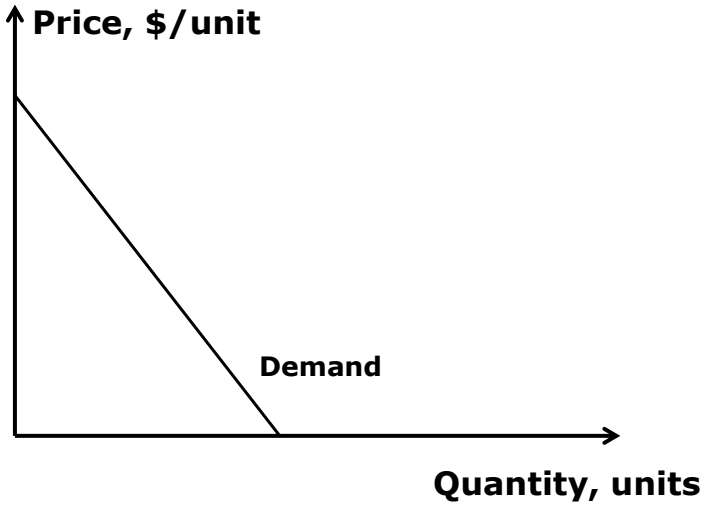
2. *Elastic demand:* Price elasticity of demand is greater than 1 in absolute value or $|e_D| > 1$. This happens if the percentage change in the quantity demanded exceeds the percentage change in the price. Graphically it implies a flat down-sloping demand curve:



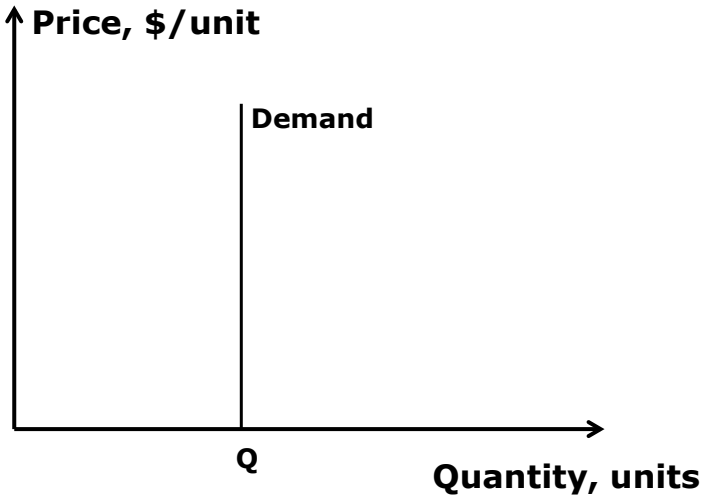
3. *Unit elastic demand:* Point price elasticity of demand is equal to – 1. This happens if the percentage change in the quantity demanded equals the percentage change in price. Graphically it is not a straight line, but rather a hyperbola of specific shape:



4. *Inelastic demand:* Price elasticity of demand is less than 1 in absolute value or $|e_D| < 1$. This happens if the percentage change in the quantity demanded is less than the percentage change in the price. Graphically it implies a steep down-sloping demand curve:



5. Perfectly inelastic demand: Price elasticity of demand is equal to 0. This happens if the quantity demanded remains constant as the price changes. Graphically it implies a vertical demand line:



What determines the price elasticity of demand is the substitution effect and income effect. *Substitution effect* is associated with the statement: "The demand for a good is elastic if a substitute for it is easy to find". Three main influences conform to this statement:

(i) Necessities versus Luxury

A necessity is a good that has few substitutes (after all, it is a necessity) and therefore, the demand for a necessity (e.g., food) is inelastic. A luxury has many substitutes, so the demand for a luxury (e.g., Russian caviar or diamonds) is elastic

(ii) Narrowness of definition

The demand for a narrowly define good is elastic since in such a case it is easy to find more substitutes. The demand for a broadly defined good is inelastic (e.g., apples versus food in general)

(iii) Time horizon

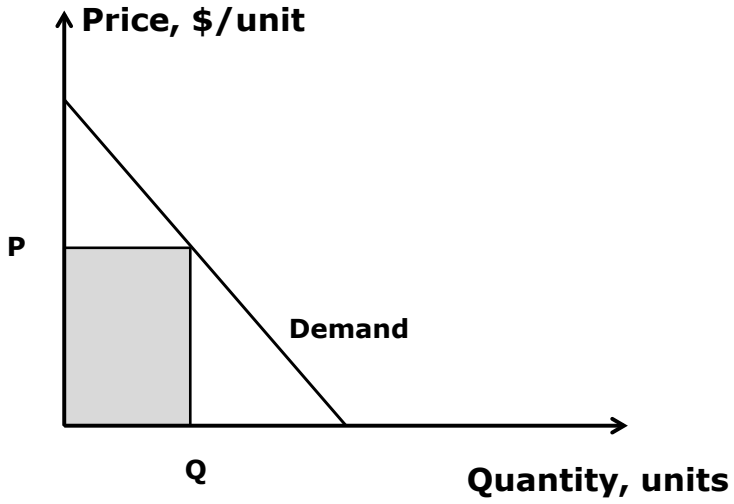
The longer the time that has elapsed since the price of a good changed, the more elastic is demand for the good. With passage of time, consumers can adjust to price changes and find new substitutes.

Income effect is associated with the following statement: "The greater the proportion of income spent on a good, the greater is the impact of a rise in its price on the quantities that people can afford to buy". It implies that the greater proportion of income spent on a good results in more elastic demand.

2.9 Total revenue and price elasticity of demand

Total revenue: The amount spent on a good and received by its sellers that equals the price of the good multiplied by the quantity of the good sold

With respect to demand, total revenue can be presented graphically as follows:



The shaded area on the above graph is total revenue. When the price changes, total revenue can change in the same direction, the opposite direction, or remain constant. Which of these outcomes occurs depends on the price elasticity of demand.

If demand is elastic, a given percentage rise in the price brings a larger percentage decrease in the quantity demanded, so total revenue decreases.

If demand is inelastic, a given percentage rise in the price brings a smaller percentage decrease in the quantity demanded, so total revenue increases.

If demand is unit elastic, a given percentage rise in the price brings the same percentage decrease in the quantity demanded, so total revenue remains unchanged.

2.10 Price elasticity of supply

Point price elasticity and midpoint price elasticity defined with respect to demand can be applied to supply as well. The only difference is the sign. Since supply reflects a positive relationship between the price of a good and its quantity, the price elasticity of supply is a positive number: Changes in prices and quantities supplied move in the same direction. Moreover, the formulas developed for the price elasticity of demand can be also used for the price elasticity of supply.

2.11 Other types of elasticity

The other two useful elasticities are:

- Cross price elasticity
- Income elasticity

Cross price elasticity of demand is a measure of the extent to which the demand for a good changes when the price of a substitute or complement changes, other things remaining the same. General expression is:

Cross elasticity of demand = (Percentage change in quantity demanded of a good)/(Percentage change in price of one of its substitutes or complements)

The sign of this indicator shows whether the two goods under study are substitutes or complements. The cross elasticity of demand for substitutes is positive while the cross elasticity of demand for complements is negative.

Mathematically it can be presented as

$$e_{12} = \frac{\Delta Q_2}{\Delta P_1} \times \frac{P_1}{Q_2}$$

where

ΔP_1 is the change in the price of good 1

ΔQ_2 is the change in the quantity of good 2

P_1 is the price of good 1

Q_2 is the quantity of good 2

Income elasticity of demand is a measure of the extent to which the demand for a good changes when income changes, other things remaining the same. General expression is:

Income elasticity of demand = (Percentage change in quantity demanded)/(Percentage change in income)

Mathematically it can be presented as

$$e_I = \frac{\Delta Q}{\Delta I} \times \frac{I}{Q}$$

where

ΔQ is the change in quantity;

ΔI is the change in money income;

Q is initial quantity;

I is initial money income

The income elasticity of demand falls into three ranges:

- Greater than 1 (normal good, income elastic)
- Between 0 and 1 (normal good, income inelastic)
- Less than 0 (inferior good)

3 Consumer Choice and Demand

Key concepts discussed in this chapter: rationality, budget constraint, budget line, utility, total utility, marginal utility, diminishing marginal utility principle, optimal consumption basket, utility maximizing rule, Marshallian demand, marginal willingness to pay, consumer surplus

3.1 Consumption and rationality

The Law of Demand emphasizes negative relationship between market price and quantity of a good or service consumed. However, it does not provide answers to the following questions:

- Why does the quantity demanded of a good increase when its price falls?
- Why is demand for one good influenced by the prices of other goods?
- Why does income influence demand?
- What makes demand elastic or inelastic?

In general, utility theory answers these questions. Economists do recognize that people have individual differences. However, in general they behave rationally. Therefore, in answering the question of why do consumers buy goods and services, economists conclude: Because they obtain pleasure or satisfaction from consumption. Rationality here comes in the form of a statement: The more you buy, the more pleasure or satisfaction you get which is known as *the more the better principle*. Therefore, it looks like a rational consumer would go for an infinite quantity of a good or service. However, in our world of scarcity all consumers are limited by the money they earn or as an economist would say there is a budget constraint.

3.2 Budget constraint and budget line

Let us start with the formal definition:

Budget line: A line that describes the limits to consumption choices and that depends on a consumer's budget (money income) and the prices of goods and services; it shows the various combinations of goods an individual can afford

Suppose that a consumer has a dollar amount I (money income) that he/she has to allocate between two goods given prices of these goods P_1 and P_2. If the consumer buys Q_1 of the first good and Q_2 of the second, his/her total spending is:

$$P_1Q_1 + P_2Q_2$$

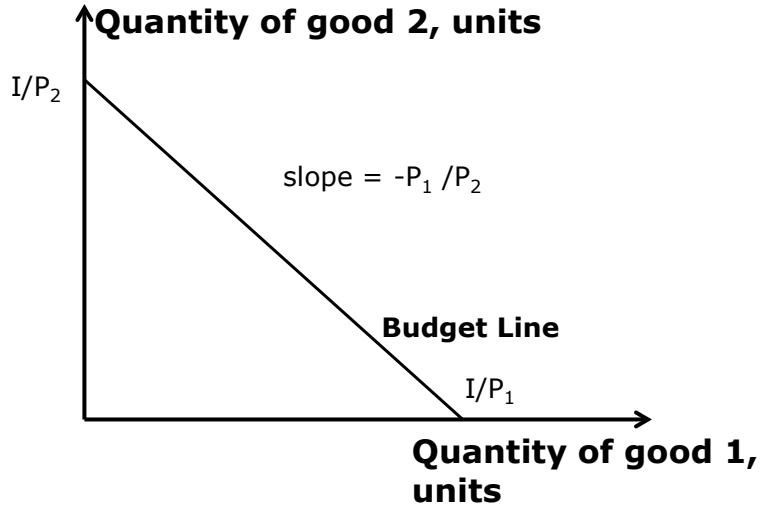
Given the consumer's money income *I*, his/her spending cannot exceed what he/she earns which means that the total spending (the consumer expenditures) has to exactly equal her money income or

$$P_1 Q_1 + P_2 Q_2 = I$$

Last expression is called the consumer's *budget constraint*. In order to derive the budget line given budget constraint, we have to solve the last equation for Q_2:

$$Q_2 = \frac{I}{P_2} - \frac{P_1}{P_2} Q_1$$

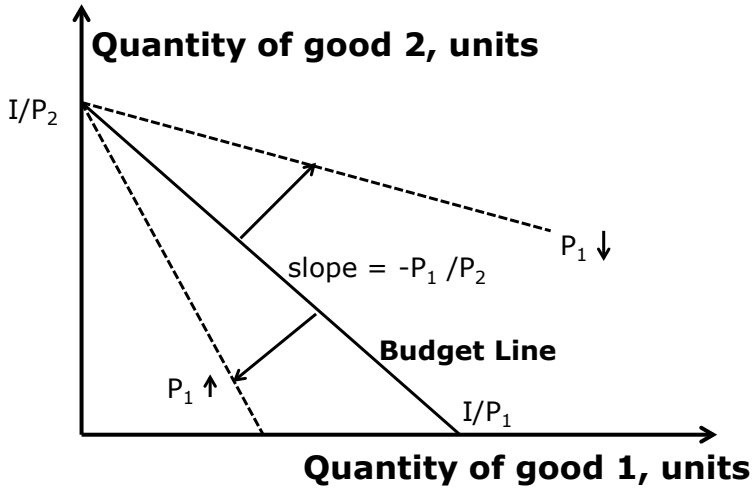
If we further assume that the prices of the two goods P_1 and P_2 are constant, and the consumer earns the same income every month, then it is easy to recognize that the budget line is a linear function in quantities. Graphically a linear function is represented by a straight line. In term of graphical interpretation, I/P_2 is the line's vertical intercept while P_1/P_2 is its slope. Moreover, since there is a minus sign before the price ratio, the line is down-sloping. Hence graphically the budget line can be presented as follows:



The consumer budget line is similar to the production possibilities frontier (PPF). Both show a limit to what is feasible. Therefore, we may call it the consumption possibilities frontier.

If the price of one good rises, when the prices of other goods and the money income remain the same, consumption possibilities shrink. If the price of one good falls, when the prices of other goods and the money income remain the same, consumption possibilities expand.

Suppose the price of good 1, P_1 falls while the price of good 2, P_2 remains unchanged. Since money income I remains constant as well, vertical intercept of the budget line I/P_2 is the same. However, its slope P_1/P_2 decreases which makes the budget line flatter or it rotates counter-clockwise around its vertical intercept:

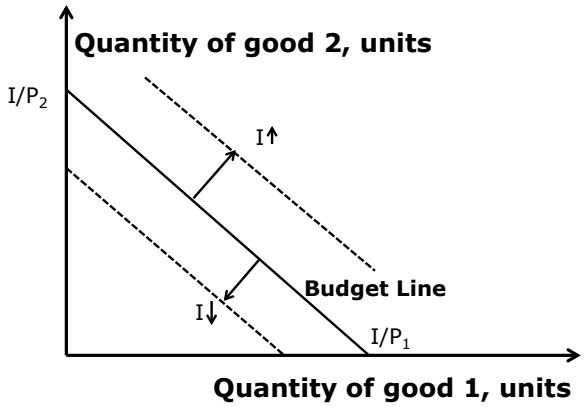


Consequently, if the price P_1 rises, the budget line rotates clockwise around its vertical intercept. It turns out that the budget line's slope reflects opportunity cost in terms of relative price:

Relative price: The price of one good in terms of another good – an opportunity cost

It equals the price of one good divided by the price of another good.

If the consumer's money income *I* increases, he/she will buy more of both goods at given prices. Graphically it means that the budget line shifts outward since vertical intercept of the line increases while its slope remains the same:



Consequently, if the consumer's income decreases, his budget line will shift inward.

3.3 Utility

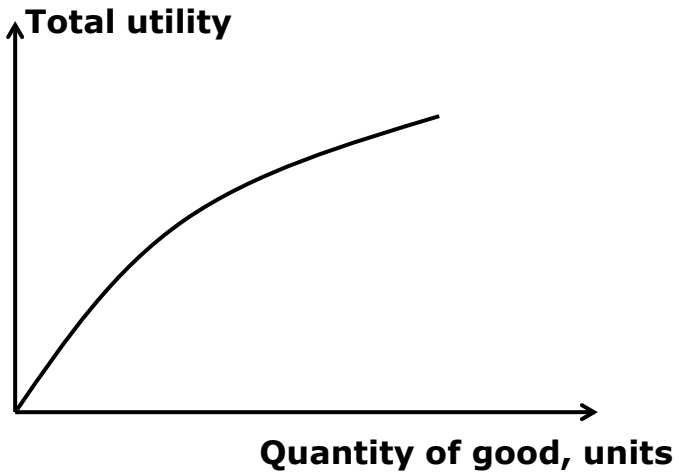
Economists define utility as follows:

Utility: The benefit or satisfaction that a person (consumer) gets from the consumption of a good or service

In economics utility is an index of how much a person wants something. It is a useful but abstract concept measured in arbitrary units. The measure of utility is ordinal which means that we can only rank an individual's preferences. Utility is personal, and it cannot be compared across individuals. First we have to distinguish between total utility and marginal utility.

Total utility: The total benefit that a person (consumer) gets from the consumption of a good or service

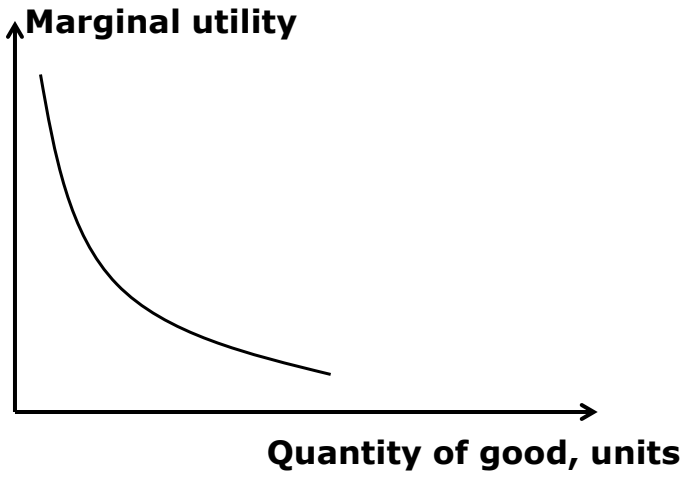
Total utility generally increases as the quantity consumed of a good increases:



Marginal utility: The change in total utility that results from a one-unit increase in the quantity of a good consumed

Marginal utility is subject to the diminishing marginal utility principle:

Diminishing marginal utility: The general tendency for marginal utility to decrease as the quantity of a good or service consumed increases



The principle is quite consistent with the observed consumer behavior. Definition of the total utility and the diminishing marginal utility principle imply:

- Total utility is an increasing function of the quantity of a good consumed, but at a decreasing rate
- Marginal utility is a decreasing function of the quantity of a good consumed

3.4 Maximizing total utility

Being a rational agent, a consumer's ultimate goal is to allocate the available budget (money income) in a way that maximizes total utility. The consumer achieves this goal by choosing the point on the budget line at which the sum of the utilities obtained from each good is as large as possible. This outcome occurs when a person follows *the utility-maximizing rule*:

- Allocate the entire available budget;
- Make the marginal utility per dollar spent the same for all goods/services.

The available budget is the amount available to an individual after he/she paid taxes and chose how much to save. Spending the entire available budget does not automatically maximize utility. Below we will see how the utility-maximizing rule works.

On the one hand, a consumer wants maximum utility due to the more the better principle, but on the other hand, the consumer has a limited budget. Actually, from a mathematical standpoint, the consumer solves a constrained optimization problem of the form: Maximize utility given budget constraint. Solution to the problem is the quantity of a good or service that the consumer is going to buy or *optimal consumption basket*.

With respect to rational choice it means that people choose goods and services in order to maximize total utility. In the case of consumption, the principle of rational choice works as follows:

Spend your money on these goods that give you the most marginal utility per dollar

Suppose that there are just two goods, *A* and *B,* with prices given by P_A and P_B and marginal utilities MU_A and MU_B

If $\frac{MU_A}{P_A} > \frac{MU_B}{P_B}$, a rational consumer will choose to consume an additional unit of *A*

If $\frac{MU_A}{P_A} < \frac{MU_B}{P_B}$, a rational consumer will choose to consume an additional unit of *B*

When do consumers stop adjusting their consumption? The principal of rational choice says that consumers keep adjusting their spending as long as marginal utility per dollar of two goods differs. The only time consumers do not adjust their spending is when there is no clear choice. When the ratios of marginal utility to price of the two goods are equal, consumers are said to maximize their utility. Therefore, expression

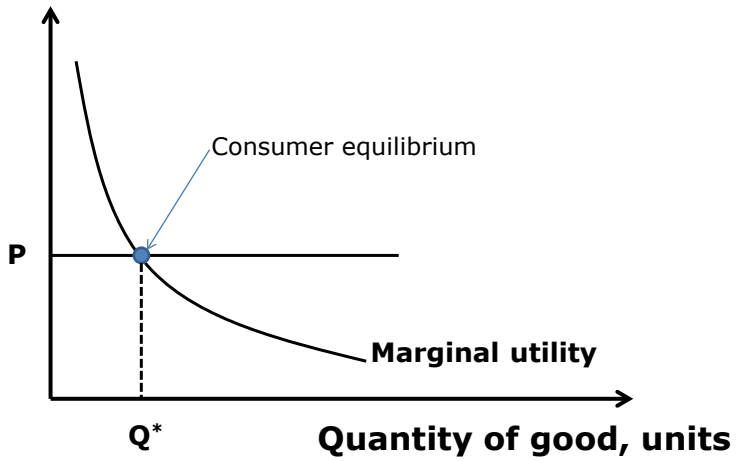
$$\frac{MU_A}{P_A} = \frac{MU_B}{P_B}$$

is known as the *utility-maximizing rule*. And when consumers are maximizing utility, they are in equilibrium. In order to understand how consumers can achieve equilibrium by adjusting spending, it is important to remember the principle of diminishing marginal utility. When a consumer consumes more of a good or service, the marginal utility of the last unit of that good or service decreases. In turn, as a consumer consumes less of a good, the marginal utility from the last unit consumed increases.

The two-good framework seems to be oversimplified. However, we can regard good *A* as a good under study, while treat good *B* as a composite good. We call the composite good *B* "All Other Goods". If expenditure on good *A* is a small portion of total income (which is almost always the case), then we can approximate it by interpreting good *B* as income (budget) itself. Since the price of a dollar of income is one (it is dollar itself) and the marginal utility of extra dollar is also one dollar, this interpretation reduces the consumer's utility-maximizing rule to:

$$\frac{MU_A}{P_A} = 1 \quad \text{or} \quad MU_A = P_A$$

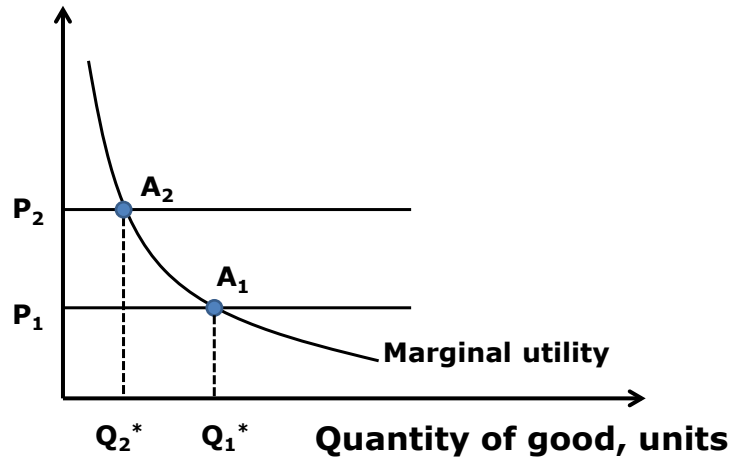
This is a single-good version of the utility-maximizing rule. As already explained, graphically marginal utility is represented by a downward sloping line. Let *P* be the current market price of a good. Since it is constant, graphically it can be represented by a horizontal line with the height equal to the value of the price in "Price-Quantity" space. Combining the two functions on one graph:



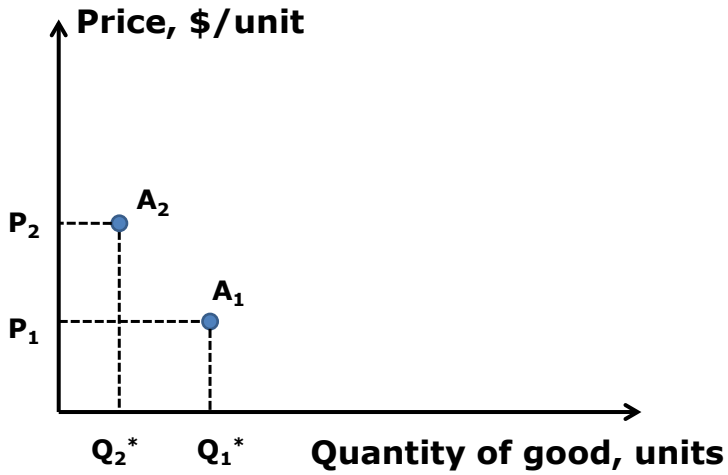
Price reflects the consumer expenditure: the more he/she buys, the more he/she spends. Therefore, the horizontal line reflects cost to the consumer. In turn, downward-sloping marginal utility shows benefit to the consumer. From the above diagram we can see that, for the first unit, benefit exceeds cost, and the consumer will definitely buy the unit. He/she will do so until benefit = cost at consumer equilibrium, which identifies Q^* as the optimal level of consumption because beyond this level cost exceeds benefit. Optimal point is associated with $MU = P$ which is the utility maximizing rule we derived earlier.

3.5 Deriving the demand curve

Suppose that the price of a good is P_1. According to a single-good version of the utility maximizing rule, at this price optimal level of consumption is Q_1^* as it is shown below:



If the price rises to P_2, then again according to the utility-maximizing rule optimal level of consumption becomes Q_2^* (point A_2). In fact we obtained two points on the demand curve: point A_1 is associated with the price P_1 and quantity demanded Q_1^*; point A_2 is associated with the price P_2 and quantity demanded Q_2^*. Graphically:



By changing price and applying the utility-maximizing rule each time, we can obtain as many points as we want. In fact, we are moving along the marginal utility line which means that the marginal utility line and demand curve coincide. So, eventually we would end up with the demand curve which is known as the Marshallian demand after a famous British economist of the end of 19th–the beginning of 20th century Allan Marshall. Therefore, we can now clearly see that the principle of diminishing marginal utility is behind the Law of Demand.

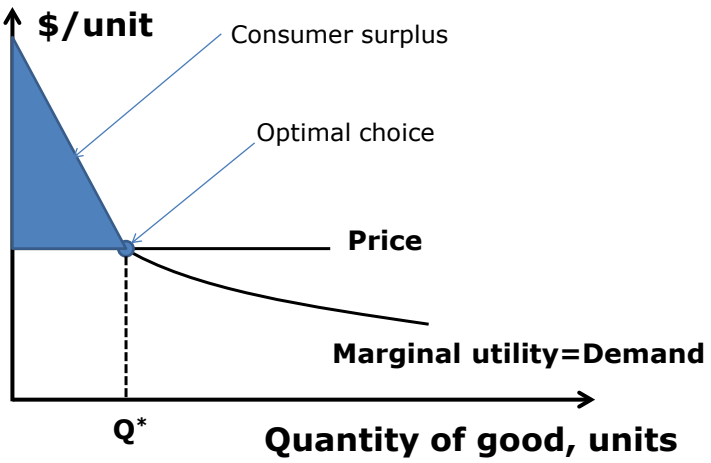
Note, that the price elasticity of demand is related to marginal utility. If, as the quantity consumed of a good or service increases, marginal utility decreases quickly, the demand for the good/service is inelastic. The reason is that for a given change in the price, only a small change in quantity of the good/service consumed is needed to bring its marginal utility per dollar spent back to equality with that of all other items in the consumer's budget. On the other hand, if, as the quantity of a good or service consumed increases, marginal utility decreases slowly, the demand for that good/service is elastic.

When a consumer allocates his/her limited budget to maximize his/her total utility, it is said that the consumer is using his/her resources efficiently. The derived demand curve describes the quantity demanded at each price *when total utility is maximized*. The demand curve is also a willingness-to-pay curve. It tells us a consumer's marginal benefit – the benefit from consuming an additional unit of a good as seen by the consumer himself/herself. In fact, definitions of the two are the same:

Marginal benefit/Marginal willingness to pay: The maximum price a consumer is willing to pay for an extra unit of a good or service when total utility is maximized

3.6 Consumer surplus

In terms of rational choice, when you think about what made you feel good about a recent purchase it is likely that you ended up paying less than the price at which you valued the product or service. Economist would say that you purchased this product at a price lower than your marginal willingness to pay. Let us discuss the following diagram in more detail:



In this diagram, the downward sloping Marginal Utility = Demand line shows a consumer's willingness to pay for a particular amount of a good under study. More precisely, the height of the marginal utility line is the consumer's marginal willingness to pay, associated with the last unit of the good consumed. As we found out, the amount Q^* was the optimal level of consumption, because it was associated with maximum utility. If we sum up marginal willingness to pay for all amounts between 0 and Q^*, we will end up with the consumer's Total Willingness to Pay for the amount Q^*. Graphically it is the area under the Marginal Utility = Demand line up to the optimal level of consumption Q^*. The area reflects the consumer's total benefits as seen by the consumer himself/ herself.

On the other hand, in order to consume the amount Q^*, the consumer has to pay amount equal to the price times quantity. This is the consumer's expenditures. Graphically the expenditures are represented by the area of a rectangle under the Price line up to the optimal level of consumption Q^*. If we now subtract the consumer's expenditures from total benefits, the result will be the so-called consumer surplus, defined as follows:

Consumer Surplus: The excess of total benefits, measured through total willingness to pay, over total expenditures

The consumer surplus is represented by the shaded area in the above diagram. Therefore graphically, the consumer surplus can be defined as the area under the demand curve and above the market price. We will later use this interpretation of consumer surplus analyzing individual markets or markets for various goods and services.

Fundamental conclusion from our discussion in this chapter is: We have to remember that the demand curve reflects four things at the same time

- Inverse (negative) relationship between the price and quantity of a good/service demanded
- Marginal utility
- Marginal willingness to pay
- Marginal benefit

In addition, we have to remember that the principle of diminishing marginal utility stands behind the Law of Demand, and that any point on the demand curve shows optimal combination of the price and quantity consumed at this price when utility is maximized.

4 Production and Costs

Key concepts discussed in this chapter: firm, sole proprietorship, partnership, corporation, profit, total revenue, total costs, accounting profit, economic profit, explicit costs, implicit costs, depreciation costs, opportunity costs of owner-provided capital, opportunity costs of owner-provided labor, normal profit, short-run, long-run, total product, marginal product, average product, marginal cost, average cost, increasing returns to scale, decreasing returns to scale, constant returns to scale, economies of scale, diseconomies of scale

4.1 A firm in economics

In general, a firm is any business such as a sole proprietorship, partnership or corporation:

A sole proprietorship: A business structure in which an individual and his/her company are considered a single entity for tax and liability purposes. A sole proprietorship is a company which is not registered with the state as a limited liability company or corporation. The owner does not pay income tax separately for the company, but he/she reports business income or losses on his/her individual income tax return. The owner is inseparable from the sole proprietorship, so he/she is liable for any business debts.

Partnership: A relationship of two or more entities (individuals) conducting business for mutual benefit

Corporation: The most common form of business organization, and one which is chartered by a state and given many legal rights as an entity separate from its owners.

From a standpoint of economic theory, a firm is an economic institution that transforms factors of production (inputs of production, resources) into goods and services. In general, any firm

- faces scarcity of all resources and makes a rational choice;
- faces the cost of something it must give up to produce it;
- responds to incentives.

The above principles force the firm to make decisions that maximize profit. A firm that does not seek to maximize profit is either eliminated by competition or bought by firms that do seek that goal.

Broadly speaking, profit is the difference between total revenue and total costs:

Total Revenue: The price received for selling a good or service multiplied by the quantity of the good/service sold at that price

Total Costs: Sum of all opportunity costs related to the production process

From above definition it turns out that economists have a special way of defining and measuring costs and consequently profit.

4.2 Accounting versus economic costs and profits

It is necessary to distinguish between two interpretations of costs and profit – accounting and economic. First of all, why do we need two interpretations? It turns out that they are complementary to each other, and they are associated with different goals.

Accountants measure costs and profit

- to ensure that a firm pays the correct amount of income tax;
- to show a bank how a firm's owner uses his/her bank loan;
- to calculate dividends paid to a firm's shareholders.

Accounting profit is the excess of a firm's income (total revenue in economic terms) over the firm's expenses (total costs in economic terms). The firm earns money after selling goods or services. If the money the firm earns is more than the money the firm spends for making/providing the goods/services, it is said that the firm has made an accounting profit. Accounting costs (expenses) do not only include the tangible money that was spent by the firm, but also include any provision for losses or depreciation that the firm makes over an accounting period. So once all these costs are subtracted from the total revenue earned by a firm, the remaining amount is an accounting profit

Economic profit is not just the excess of total accounting income (total revenue) over the total accounting expenses (total costs). It treats total costs as opportunity costs. In general, economists want to allocate all available resources in the most efficient way. It means that economists want to analyze the decisions that a firm makes to maximize its profit. These decisions respond to opportunity costs and economic profit.

In order to produce its output, a firm employs factors or inputs of production such as land, labor, capital and entrepreneurship. Another firm could have used these same resources to produce alternative goods and services. The highest-valued alternative forgone is the opportunity cost of a firm's production. From the viewpoint of the firm, this opportunity cost is the amount that the firm must pay the owners of the factors of production it employs to attract them from their best alternative use. So, a firm's opportunity cost of production is the cost of the factors of production it employs.

However, the amount that a firm pays to attract resources from their best alternative use is either explicit costs or implicit costs.

Explicit costs: Costs that are quantifiable and usually represent an actual payment in terms of money

Wages of workers, expenditures on equipment, rent payments are examples of explicit costs. Explicit costs are out-of-pocket costs or what is actually paid by a firm.

Implicit costs: Expenses that are difficult to quantify because there is no explicit payment made for the use of a factor of production; opportunity costs incurred by a firm when it uses the factor of production for which it does not make a direct money payment

Implicit costs are measured by what a factor of production (land, capital, etc) currently used in production could earn in other uses. Examples of implicit costs are:

- Depreciation costs
- Opportunity costs of owner-provided capital
- Opportunity costs of owner-provided labor
- Normal profit

Depreciation costs: Opportunity costs of a firm using capital that it owns measured as the change in the market value of capital over a given period of time

As a matter of fact, depreciation costs are usually included in accounting costs to approximate costs of capital assets like equipment, machinery, buildings, infrastructure, tools and others.

Opportunity costs of owner-provided capital are costs tied-up in capital goods (capital assets)

When a firm's owner buys a capital asset (e.g., a piece of equipment), he/she forgoes interest he/she could have earned by keeping these funds in a bank or invested in financial assets.

Opportunity costs of owner-provided labor are forgone costs of the owner's wage: Instead of managing his own business, a firm's owner could have earned some money working elsewhere

Finally, firms remain in an industry if they earn at least a normal rate of return. Owner of a firm often supplies entrepreneurship, the factor of production that organizes the business and bears the risk of running it. The return to entrepreneurship is normal profit defined as

Normal profit: The return to entrepreneurship which is a part of a firm's opportunity costs because it is the cost of not running another firm

As already mentioned, economic profit is the difference between total revenue and total costs. Total revenue is the amount received from the sale of the product. It is the price of the output multiplied by the quantity sold.

Total costs (or rather total economic costs) are the sum of the explicit and implicit costs, including normal profit, and are the opportunity costs of production.

The above discussion allows us to establish the following relationships:

- Accounting Profit = Total Revenue – Explicit Costs – Depreciation Costs
- Economic Profit = Total Revenue – Implicit Costs – Explicit Cost
- Opportunity Costs = Implicit Costs + Explicit Costs

Let us calculate the difference between accounting and economic profit:

Accounting Profit – Economic Profit = (Total Revenue – Explicit Costs – Depreciation Costs) – (Total Revenue – Explicit Costs – Implicit Costs) = Implicit Costs – Depreciation Costs

The difference is obviously positive since implicit costs include depreciation costs. It means that accounting profit is larger than economic profit which leads to the following possibility: Accounting profit is positive while economic profit is negative. In such a case, the firm's resources (factors of production) are not in their best use or, in other words, the business is not efficient.

4.3 The short-run and long-run

In microeconomic theory, traditionally all resources or factors (inputs) of production are divided into:

- Labor
- Capital
- Technology
- Entrepreneurship

In order to study the relationship between a firm's output decision and its costs, we have to distinguish between two time frames – the short-run and the long-run:

Short-run: The time frame in which the quantities of some resources are fixed

In the short-run, a firm can usually change the quantity of labor it uses but not the quantity of capital and technology. Organization of management or entrepreneurship is also fixed in the short-run. We call these fixed resources the firm's fixed inputs. In economic theory, the collection of such inputs of production as capital, technology and entrepreneurship is called a firm's plant. Hence in the short-run, the firm's plant is fixed. In agriculture and in some businesses land may be a fixed input as well.

In order to increase its output in the short-run, a firm must increase the quantity of variable inputs. Labor is usually the variable input (factor of production). In general, the short-run decisions are easily reversed: A firm can increase or decrease output by increasing or decreasing the labor hours it hires.

Long-run: The time frame in which the quantities of all resources can be changed

Therefore, the long-run is a period in which the firm can change its plant size. Long-run decisions are not easily reversed: Once a firm buys a plant, its resale value is usually much less than the amount the firm paid for it.

The short-run and long-run are management planning horizons that apply to decisions made by managers. The firm can operate only in the short-run in the sense that decisions must be made in the present. For example, investment is associated with a long- run decision made in the short-run.

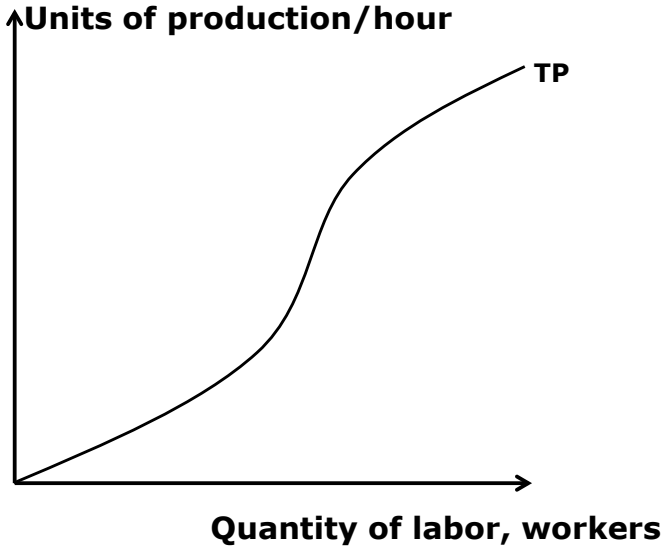
4.4 Short-run production

As already mentioned, in order to increase output of a fixed plant, a firm must increase the quantity of labor it employs. Three economic concepts are of most importance in order to understand this process:

- Total product, *TP*
- Marginal product, *MP*
- Average product, *AP*

Total product: The total output produced in a given period

It is measured in units produced per unit of time. Total product, *TP* increases as the quantity of labor employed increases and in general is presented by the following curve:



Like the production possibilities frontier, the total product (TP) curve separates attainable outputs from unattainable outputs. All the points above the TP-line are unattainable. Points below the TP-line are attainable but not efficient: They use more labor than it is necessary to produce a given output. Only points on the TP-line are efficient.

Marginal product: The change in total product that results from a one-unit increase in the quantity of labor employed

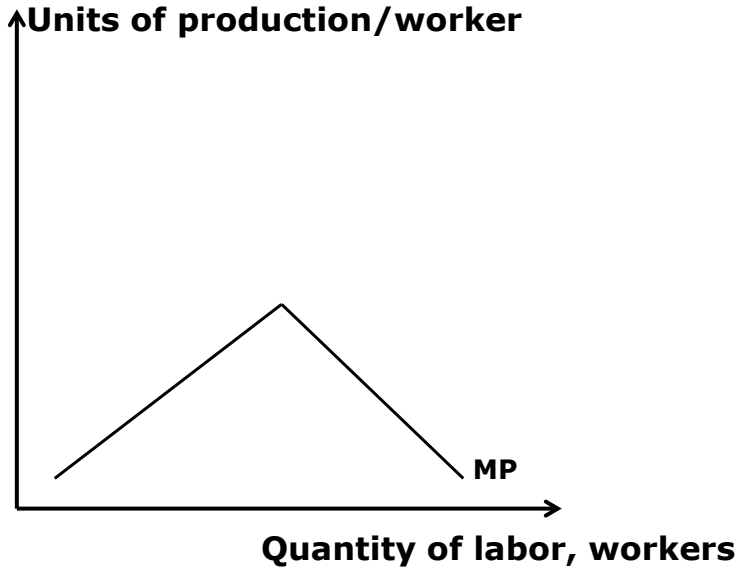
Hence:

Marginal product = Change in total product/Change in quantity of labor

or mathematically

$$MP = \frac{\Delta TP}{\Delta L}$$

where ΔTP is the change in total product while ΔL is the change in quantity of labor. Given the above definition and the TP-line, the following marginal product curve arises:



The MP-line reflects a feature that is shared by all production processes: Increasing marginal returns initially, and decreasing marginal returns eventually.

Increasing marginal returns occur when the *MP* of an additional worker exceeds that of the previous one. It is usually associated with a small number of workers when specialization and division of labor result in higher productivity of labor.

Decreasing marginal returns occur when the *MP* of an additional worker is less than that of the previous one. They arise from the fact that more and more workers use the same equipment and work space; in fact, congestion results. Decreasing marginal returns are so pervasive that they qualify for the status of a law:

The Law of Decreasing (Diminishing) Marginal Returns: As a firm uses more of a variable input with a given quantity of fixed inputs, the marginal product of the variable input eventually decreases

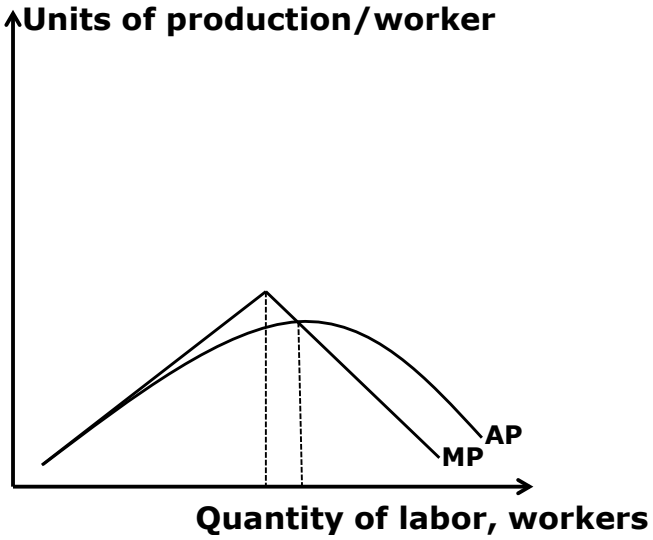
Note, that this Law is the short-run phenomenon only. And finally

Average product: Total product divided by the quantity of an input

For example, average product of labor is total product *TP* divided by the quantity of labor *L* employed or mathematically

$$AP = \frac{TP}{L}$$

The following graph shows the relationship between marginal product of labor, *MP* and the average product of labor, *AP*:



Some useful observations:

- Note that *AP* is at its maximum when *AP* = *MP*
- For employment levels at which *MP* > *AP*, the AP-line is upward-sloping
- For employment levels at which *MP* < *AP*, the AP-line is downward-sloping

As a matter of fact, the relationship between *AP* and *MP* is a general mathematical feature of the relationship between the average value and marginal value of any function.

4.5 Short-run costs

Since there are three types of product, there are three types of costs in the short-run:

- Total costs, *TC*
- Marginal cost, *MC*
- Average cost, *AC*

Total costs: The costs of all factors of production used by a firm

A firm's total costs are divides into two parts: Total fixed costs *TFC* and total variable costs *TVC*:

Total fixed costs: The costs of the fixed factors of production used by a firm

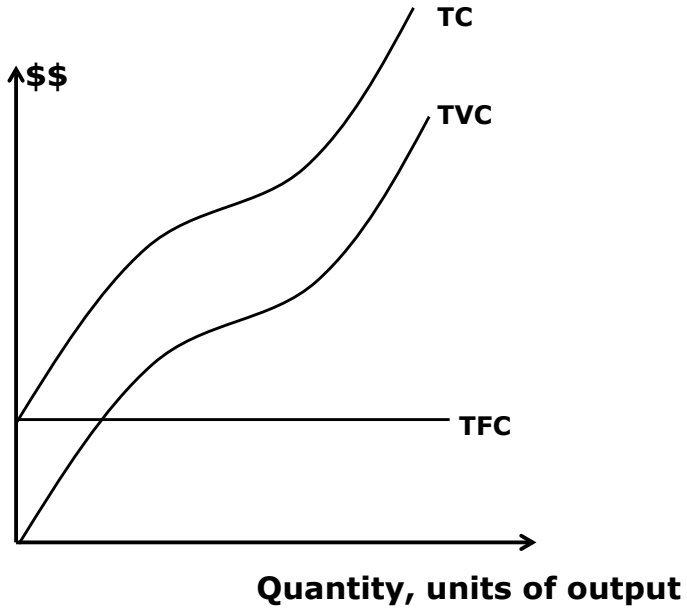
Usually *TFC* are the costs of land, capital, and entrepreneurship. Mathematical distinguishing feature of these costs is: They do not directly depend on the level of production (output).

Total variable costs: The costs of the variable factor of production used by a firm

Usually *TVC* are the costs of labor. These costs directly depend on output. Total costs *TC* are the sum of the two:

$$TC = TFC + TVC$$

The following graph illustrates the above defined costs as functions of quantity (output):



Shape of total cost is a mirror image of the total product which means that TC-line increases at a decreasing rate initially due to higher productivity but at an increasing rate eventually.

Marginal cost: The change in total costs that results from one-unit increase in output

Mathematically marginal cost *MC* is

$$MC = \frac{\Delta TC}{\Delta Q}$$

where ΔTC is the change in total costs while ΔQ is the change in quantity (output).

There are three types of average costs:

- Average fixed cost, *AFC*
- Average variable cost, *AVC*
- Average total cost, *ATC*

Average fixed cost: Total fixed costs per unit of output or $AFC = \frac{TFC}{Q}$

Average variable cost: Total variable costs per unit of output or $AVC = \frac{TVC}{Q}$

Average total cost: Total costs per unit of output which equals average fixed cost plus average variable cost or $AVC = \frac{TVC}{Q}$

Previously we defined total costs *TC* as

$$TC = TFC + TVC$$

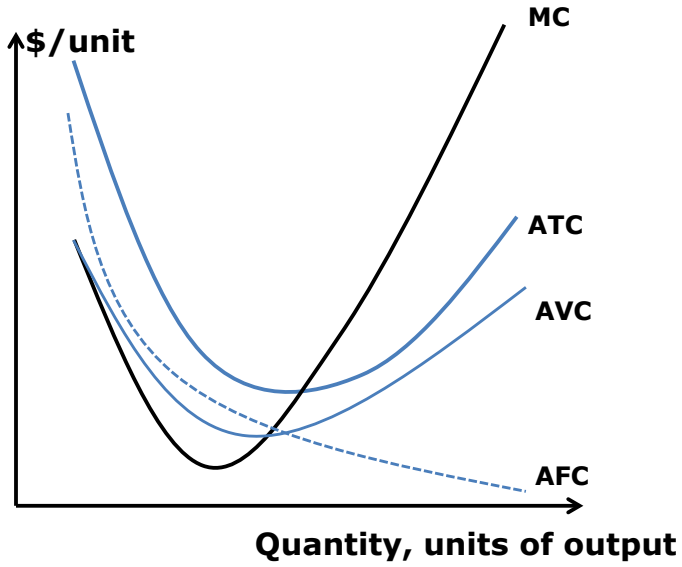
If we divide the last expression throughout by output *Q* then

$$TC/Q = TFC/Q + TVC/Q$$

Using the above-presented definitions of *ATC, AFC* and *AVC*

$$ATC = AFC + AVC$$

The following graph illustrates general shapes of the marginal *MC* and average cost functions:

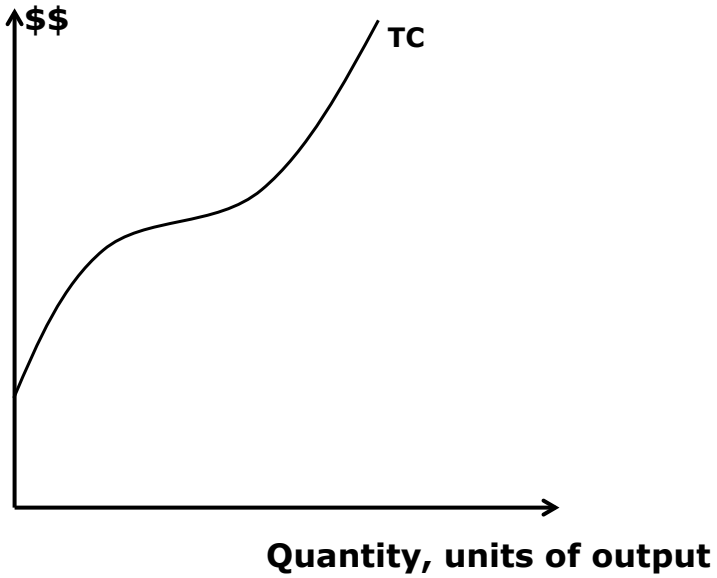


Some useful observations:

- MC, ATC and AVC are U-shaped
- MC intersects the AVC and the ATC at their minima
- If *MC* < *AVC*, then AVC is decreasing; if *MC* > *AVC*, then AVC is increasing
- If *MC* < *ATC*, then ATC is decreasing; if *MC* > *ATC*, then ATC is increasing

4.6 Simple algebra of the short-run costs

As discussed previously, in the short-run production is subject to the Law of Diminishing Marginal Returns. Graphically it means that total cost function is an increasing function in quantity. It increases at a decreasing rate initially and at an increasing rate eventually which is presented in the following graph:



Mathematically this shape can be captured by a cubic function as follows:

$$TC = a_0 + a_1 Q + a_2 Q^2 + a_3 Q^3$$

where Q is output in units of a product. In this specification, total fixed costs are captured by coefficient a_0 or

$$TFC = a_0$$

while total variable costs are captured by the rest of the expression or

$$TVC = a_1 Q + a_2 Q^2 + a_3 Q^3$$

Given these expressions, we can now define *AVC* and *AFC* as follows:

$$AVC = \frac{TVC}{Q} = a_1 + a_2 Q + a_3 Q^2$$

$$AFC = \frac{TFC}{Q} = \frac{a_0}{Q}$$

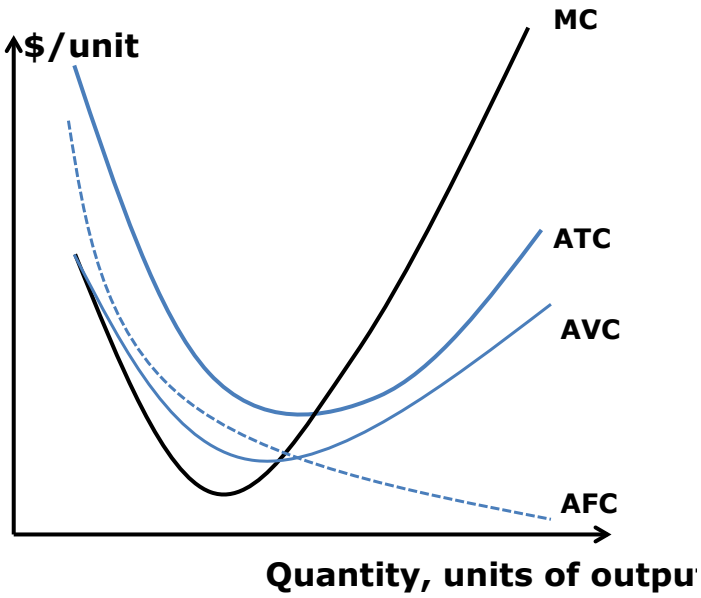
From the above specifications, we can easily see that *AVC* is a quadratic function or graphically it is represented by a parabola that has U-shape. In turn, *AFC* is represented graphically by a hyperbola. Marginal cost was defined as the change in total costs per unit of output which is just a derivate of the total cost function with respect to quantity of output or:

$$MC = \frac{dTC}{dQ} = a_1 + 2a_2Q + 3a_3Q^2$$

Therefore, *MC* is also a quadratic function. Moreover, if $Q = 0$, then

$$MC = AVC = a_1$$

which means that for the first unit of a product, marginal cost and average variable cost coincide. Now taking into account the above algebra, take a second look at the general shapes of the cost functions in the short-run:



4.7 Long-run production and costs

The first thing that happens in the long-run is that the distinction between fixed cost and variable cost disappears: All costs are variable in the long-run.

In the way we defined a plant, in the long-run, a firm can change its plant size. When the firm changes its plant size, its costs of producing a given output change as well. Three outcomes are possible:

- Increasing Returns to Scale (IRS) or economies of scale
- Decreasing Returns to Scale (DRS) or diseconomies of scale
- Constant Returns to Scale (CRS)

Increasing Returns to Scale (IRS) are observed when an increase in a firm's inputs, its plant size and labor employed, by some percentage *m*, increases its output by a larger percentage. The firm experiences what is frequently called economies of scale:

Economies of scale: Decreases in long-run average costs resulting from increases in output

The main source of economies of scale is greater specialization of factors of production.

Decreasing Returns to Scale (DRS) are observed when an increase in a firm's inputs, its plant size and labor employed, by some percentage *m*, increases its output by a smaller percentage. The firm experiences what is frequently called diseconomies of scale:

Diseconomies of scale: Increases in long-run average costs that occur as output increases

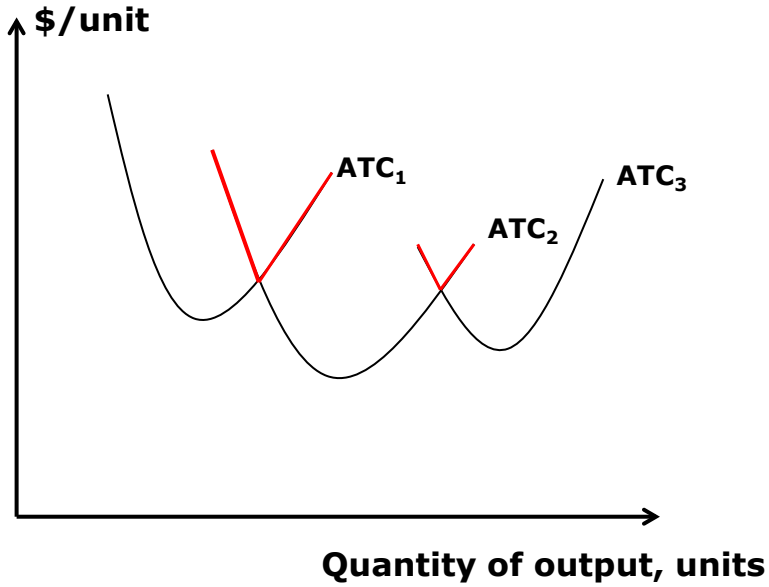
Diseconomies of scale arise from the difficulty of coordinating and controlling a large enterprise. Diseconomies of scale occur in almost all production processes but in some, perhaps, only at a very large output rate.

Constant Returns to Scale (CRS) are observed when an increase in a firm's inputs, its plant size and labor employed, by some percentage *m*, increases its output by the same percentage.

Constant Returns to Scale: No change in long-run average costs when output increases

Constant returns to scale occur when a firm is able to replicate its existing production facility including its management system.

The long-run average cost curve, LRAC shows the lowest cost at which it is possible to produce each output when the firm has had sufficient time to change both its labor and its plant size. The following graph illustrates the LRAC:



The long-run average cost curve, LRAC is derived from the short-run average total cost curves ATCs for different possible plant sizes. In the above diagram, it is helpful to think of ATC_1 as being associated with a small-size plant, ATC_2 as being associated with a medium-size plant and ATC_3 as being associated with a large-size plant. LRAC is an envelope of the three ATCs. In the above diagram it is shown by a black segment that includes portions of each of the individual ATC.

A decreasing portion of the LRAC curve displays increasing returns to scale (economies of scale), flat portion of the LRAC displays constant returns to scale, and an increasing portion of the LRAC displays decreasing returns to Scale (diseconomies of scale).

In other words, long-run average cost curve is the locus (set) of points representing the minimum unit cost of producing any given rate of output given current technology and resource prices. This long-run average cost curve is sometimes called the planning curve since this curve represents the various average costs attainable at the planning stage of the firm's decision making.

5 Perfect Competition and Efficiency of Markets

Key concepts discussed in this chapter: perfect competition, monopolistic competition, oligopoly, monopoly, profit-maximizing condition, level analysis, marginal analysis, short-run supply curve, producer surplus, total social surplus, social welfare, invisible hand, fairness, symmetry principle, conjecture

5.1 Market structures

Economic theory identifies four potential market structures:

- Perfect competition
- Monopolistic competition
- Oligopoly
- Monopoly

Perfect competition or *perfectly competitive market* is a benchmark market structure against which all other market structures are compared. It is rather an ideal market that has to satisfy the following four conditions:

1. Many firms sell identical product to many buyers
2. There are no restrictions on entry into (or exit from) the market
3. Established firms have no advantage over new firms
4. Sellers and buyers are well informed about prices

Monopolistic competition: *A market structure characterized by many firms selling similar but differentiated product (service) with free entry and exit*

Theory of monopolistic competition was originally developed by Chamberlin in 1933. In a sense, monopolistic competition violates the first condition of a perfectly competitive market presented above: Product differentiation versus homogeneous product. Therefore, all products in this market are close but imperfect substitutes. The differences are in "qualities" and circumstances such as type, style, quality, reputation, appearance and location that tend to distinguish goods and services.

Oligopoly: A market structure in which a market or industry is dominated by a small number of sellers (oligopolists) with a great deal of interdependence

Since there are few sellers, each oligopolist is likely to be aware of the actions of the others. The decisions of one firm influence, and are influenced by, the decisions of other firms. Strategic planning by oligopolists needs to take into account the likely responses of the other market participants. This causes oligopolistic markets and industries to be a high risk for collusion.

Monopoly: A market for a good or service that has no close substitutes and in which there is one supplier that is protected from competition by a barrier preventing the entry of new firms

Monopoly like perfectly competitive market is another extreme of market structure. Monopoly exists when a specific individual or an enterprise has sufficient control over a particular product or service to determine significantly the terms on which other individuals should access to it. Monopolies are thus characterized by a lack of economic competition for the good or service that they provide and a lack of viable substitute goods. In a monopoly there is one seller of the monopolized good or service that produces all the output. The firm and industry are identical.

In this chapter, we are going to first analyze perfectly competitive market and a representative firm in this market. The above specified four conditions for perfect competition usually arise when the market demand for the product is large relative to the output of a single producer and when economies of scale are absent so the efficient scale of each firm is small.

5.2 Profit-maximizing choices of a perfectly competitive firm

The main motive of any firm in any market is profit maximization. The key decisions that a firm in a perfectly competitive market or a perfectly competitive firm makes are:

- In the short-run, the firm achieves profit maximization by deciding the quantity to produce;
- In the long-run, the firm achieves profit maximization by deciding whether to stay in a market, enter or exit the market

A perfectly competitive firm does not choose the price at which to sell its output. It is a price taker which means that a perfectly competitive firm cannot influence the price of its product.

Under perfect competition, market demand and market supply determine the price of a good or service. By definition, total revenue equals the market price multiplied by the quantity sold. A firm's marginal revenue is the change in total revenue that results from a one-unit increase in the quantity sold.

It appears to be that in a perfectly competitive market, marginal revenue earned by a perfectly competitive firm is equal to price. The reason is that the firm can sell any quantity it chooses at ongoing market price. So, if the firm sells one more unit, its total revenue increases by

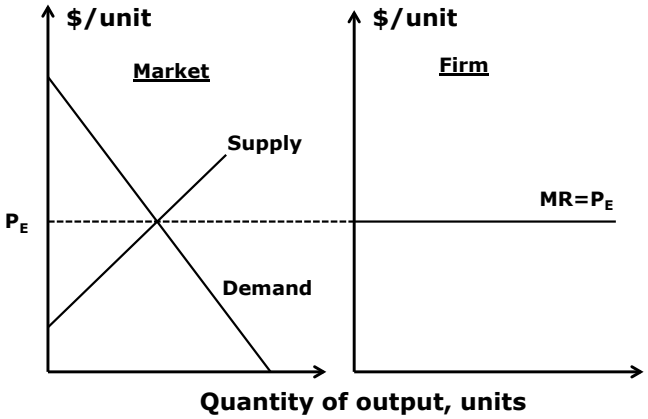
$$1\text{unit} \times \text{Price} = \text{Price}$$

which, in fact, is the firm's marginal revenue. It is constant for all output ranges.

So, forces of market demand and market supply determine market price; a perfectly competitive firm takes the price as given and decides on the level of output to maximize its economic profit. In a sense, the condition

$$\text{Marginal revenue} = \text{Price} = \text{constant}$$

defines the competitive firm's demand for its product. Graphically the above discussion can be summarized as follows:



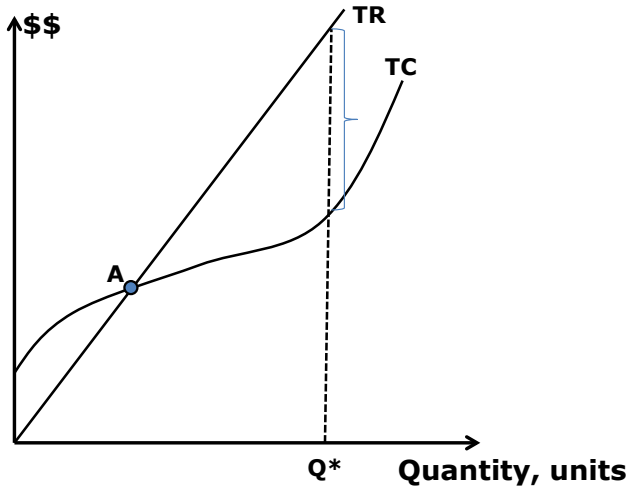
As we can see in the above diagram, demand for the competitive firm's product is a horizontal line (perfectly elastic). It is easy to see that average revenue of a perfectly competitive firm is also equal to the price and consequently to the marginal revenue. By definition, average revenue *AR* is

$$AR = \frac{TR}{Q} = \frac{P \times Q}{Q} = P$$

where *TR* is total revenue, *P* is the price and Q is the quantity of a good or service.

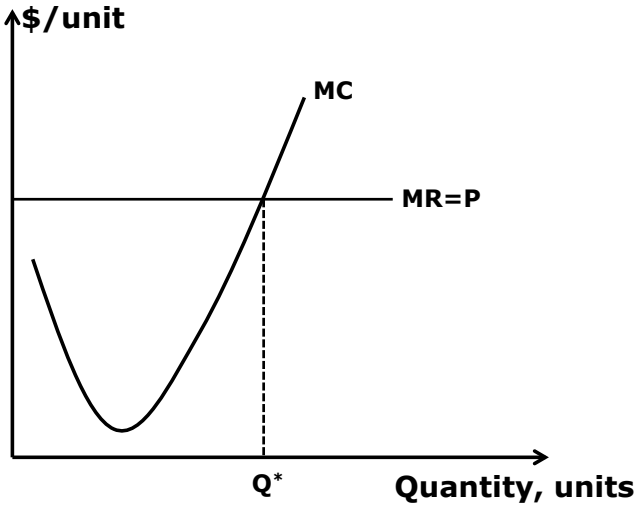
5.3 Profit-maximizing output of a perfectly competitive firm in the short-run

In the short-run, the only variable input of production is labor. Therefore, an increase/decrease in a perfectly competitive firm's output is achieved by an increase/decrease in quantity of labor. As output increases, total revenue increases as well. However, total costs also increase with an increase in output. Due to the law of diminishing marginal returns, total costs eventually increase faster than total revenue. Hence, there is one output level that maximizes economic profit, and a perfectly competitive firm chooses this level. The following graph presents total costs *TC* and total revenue *TR* as functions of output, Q:



Graphically profit is the gap between the TR-line and TC-line in the above graph. The graph shows that the distance between TR-line and TC-line is the largest when output is equal to Q^* which is the profit maximizing quantity. Point A is known as the break-even point: at this point $TC = TR$ and economic profit is zero; to the left of this point the firm incurs losses since $TC > TR$. This is the so-called *level analysis*.

Another way to find the profit-maximizing output is to use *marginal analysis* which compares marginal revenue and marginal cost. As output increases, marginal revenue *MR* remains constant and equal to the price because of the reasons already discussed. Marginal cost *MC* eventually increases with an increase in output.



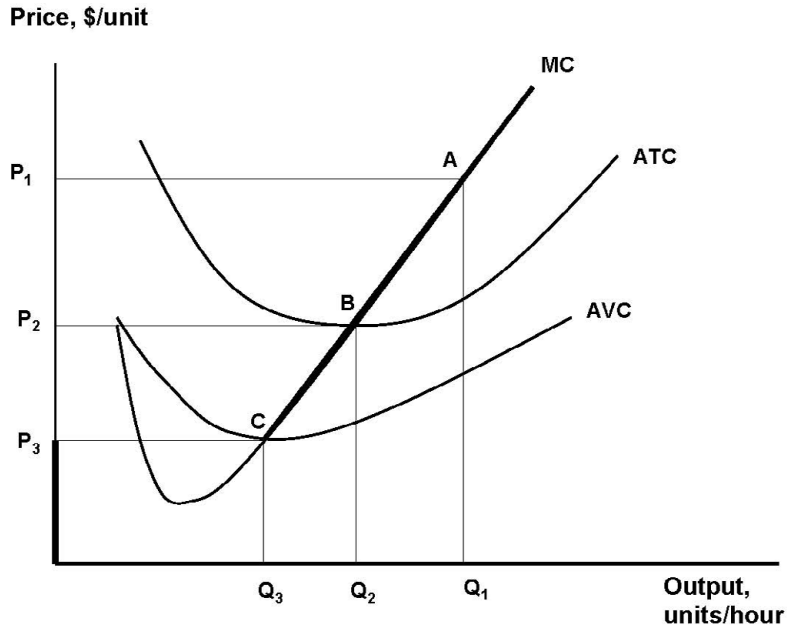
If *MR* > *MC*, then the extra revenue from selling one more unit exceeds extra cost incurred to produce it. It means that economic profit increases if output increases. However, if *MR* < *MC*, then the extra revenue from selling one more unit is less than extra cost incurred to produce it. In this case, economic profit increases if output decreases. It means that economic profit is maximized whenever *MR* = *MC*.

There is a strong link between two types of analysis. According to mathematical interpretation of marginal values, marginal function is a derivative of the original function. Geometric interpretation of derivative is slope of a function. Therefore, *MR* is the slope of TR-line while *MC* is the slope of TC-line. Profit-maximizing condition *MR* = *MC* implies: the slope of TR-line is equal to the slope of TC-line at the value of output that maximizes profit.

Since for a perfectly competitive firm *MR* = *P*, profit maximizing condition can be re-written as *P* = *MC*.

5.4 Short-run supply curve of a perfectly competitive firm

Supply curve of a perfectly competitive firm shows how the firm's profit-maximizing output varies as the price varies, other things remaining the same. The supply curve is based on the marginal analysis and the so-called shutdown decision. The following graph summarizes this analysis:



If market price is P_1, then the firm chooses point A based on profit maximizing condition $P_1 = MC$. The firm makes positive profit in the short-run since $P_1 > ATC$. It is so because of the following interpretation of profit:

$$\Pr ofit = TR - TC = Q \times (\frac{TR}{Q} - \frac{TC}{Q}) = Q \times (AR - ATC) = Q \times (P - ATC)$$

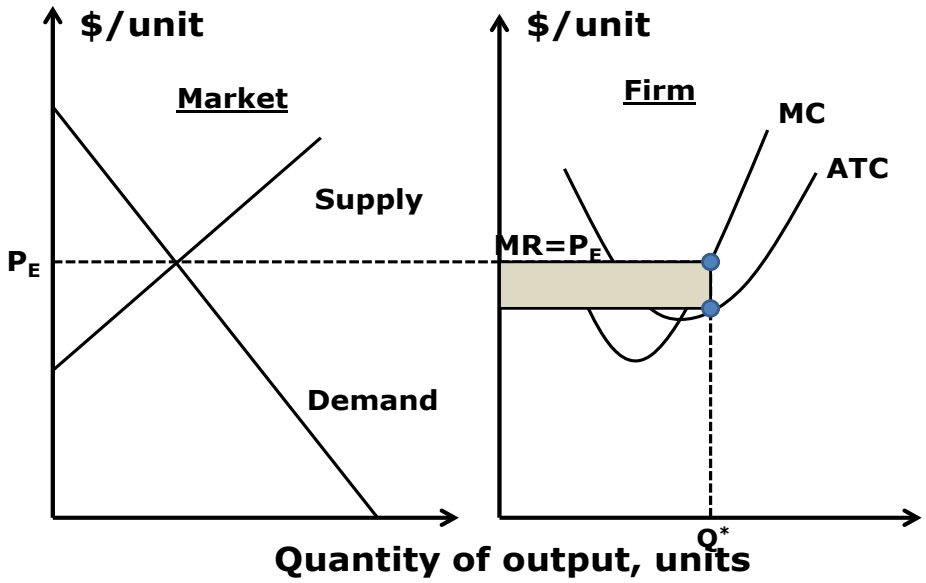
If the price is P_2, then the firm chooses point B and it earns zero economic profit in the short-run since $P_2 = ATC$. However, if the price falls below P_2, the firm makes negative profit (loss). It will operate in the short-run because in such a case it minimizes loss covering some of fixed cost, but will exit the market in the long-run. However, if the price falls below P_3, the firm will shut down right away. The above discussion gives us the following conditions:

- If $P > ATC$ produce in the short-run (positive profit)
- If $AVC \leq P \leq ATC$ still produce in the short-run but exit the market in the long-run (negative profit but loss is smaller than fixed costs)
- If $P < AVC$ shut down right away

One more very important conclusion is: The firm's MC-curve is its individual supply curve. It is so because when the firm makes its decision about optimal level of output given some price, it actually traces points along its MC-curve. Moreover, it traces points along vertical portion of the MC-curve – the portion above the AVC-curve. Therefore, the solid line in the above diagram is the firm's short-run supply curve.

5.5 Output, price and profit of a perfectly competitive firm in the short-run

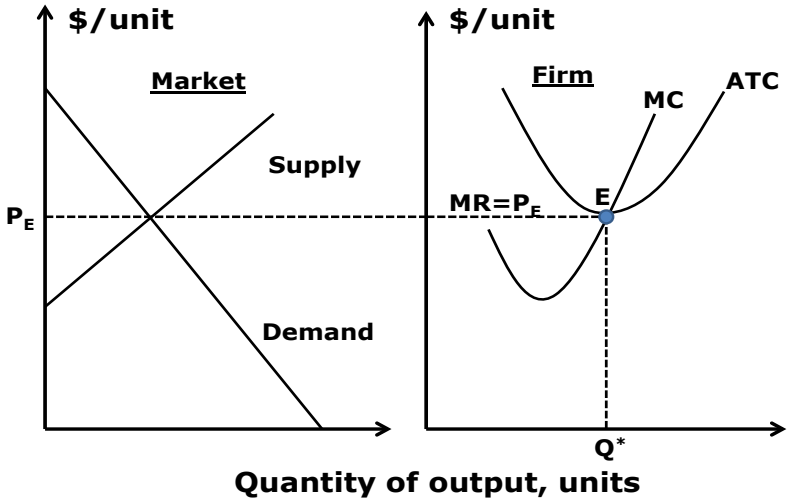
Market supply in the short-run shows the quantity supplied at each price by a fixed number of perfectly competitive firms like the one we already analyzed in the previous section. The quantity supplied at a given price is the sum of the quantities supplied by all firms at that price, obtained through horizontal summation. Interaction between market demand and the short-run market supply determines market price. A perfectly competitive firm takes the price as given and determines optimal output based on a profit-maximizing condition $P = MC$. The process is shown below graphically:



Market demand and market supply determine market price P_E. Profit-maximizing firm sets $MR = P_E = MC$ which determines the firm's short-run profit-maximizing output Q^*. At price P_E and output Q^*, the firm earns positive profit, represented by the shaded rectangle. It is so because the price is above the ATC-curve as discussed in the previous section. If the price falls below the ATC-line, the firm will incur loss.

5.6 Output, price and profit of a perfectly competitive firm in the long-run

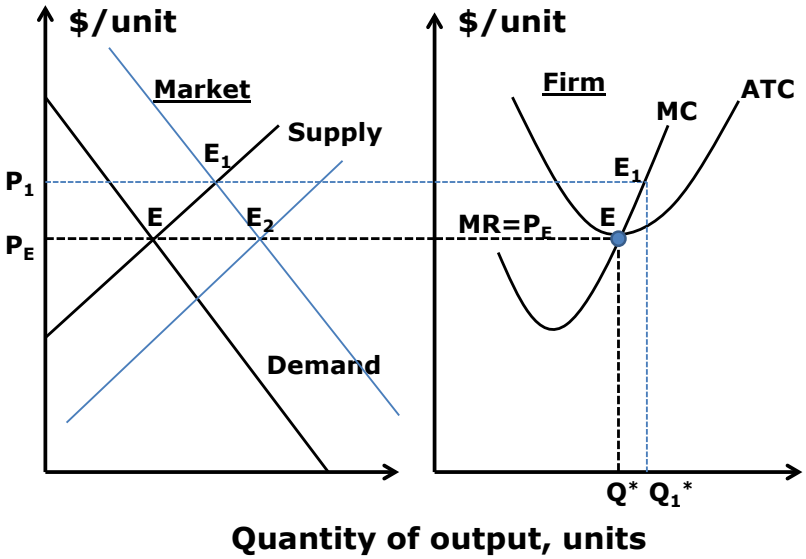
In the long-run, a perfectly competitive firm earns normal profit. It means that economic profit, in the way we defined it previously, is zero. The long-run equilibrium is shown below:



Again, market demand and market supply determine market price P_E. Long-run equilibrium of a perfectly competitive firm is point E. At this point, price equals average cost which implies total revenue equals total costs or economic profit is zero. However, since the firm's total costs include normal profit, in the long-run equilibrium the firm earns positive accounting profit which is exactly equal to the normal profit.

It is necessary to emphasize that in the short-run a perfectly competitive form can make profit or incur loss while in the long-run the firm always earns normal profit. In fact, in the long-run the firm responds to positive profit or loss by either entering or exiting the market.

If firms in competitive market earn positive economic profits in the short-run, then new firms would enter the market. Suppose that there is an increase in demand for some product. The following diagram traces the consequences of this event:



Increase in demand leads to the rightward shift of the demand line (blue demand in the left-hand graph), and equilibrium moves from point E to point E_1 in both graphs. In this short-run equilibrium, a representative perfectly competitive firm earns positive economic profit since the new price P_1 is above ATC-line. This fact sends a signal to new firms to enter the market. Since entry is easy and free, new firms enter the market which shifts the market supply to the right (blue supply in the left-hand graph). Supply shifts rightwards until all profit opportunities are eliminated and old equilibrium price P_E is restored.

Eventually, in the long-run market equilibrium at point E_2, the price is the same as before, but market quantity has increased (more firms in the market). Individual quantity of a representative firm returns to its initial value at point E. So again, the firm just earns normal profit. The same mechanism is initiated if there is a decrease in demand, or increase/decrease in supply due to external shocks or other man-made and natural events.

5.7 Efficiency of markets

First, we have to recall that demand curve reflects marginal benefits, *MB*: The demand curve for a good or service tells us the dollars' worth of other goods and services that people are willing to forgo to consume one more unit of the good of interest. Technically it means that the height of demand function is *MB*.

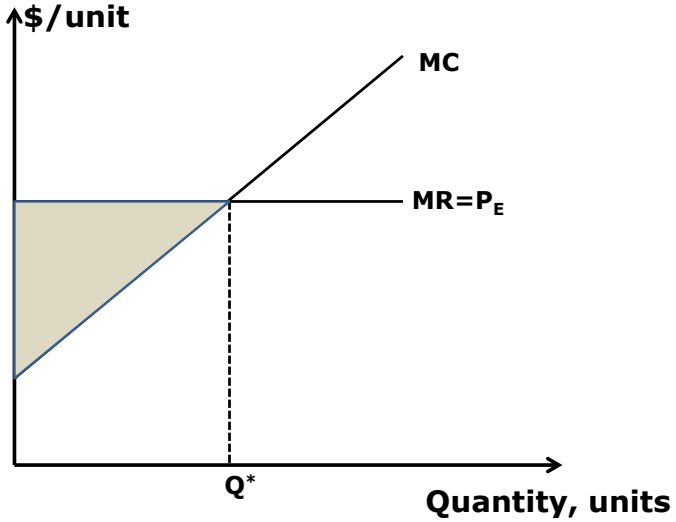
Second, as discussed in this chapter, supply curve of a perfectly competitive firm is just marginal cost curve or the height of supply is marginal cost, *MC*. The supply curve of a good or service tells us the dollars' worth of other goods and services that firms must forgo to produce one more unit of the good of interest.

We introduced the concept of consumer surplus in chapter 3. At large, consumer surplus is the marginal benefit from a good or service minus the price paid for it, summed over the quantity consumed.

It is possible to introduce a similar concept on production side. When the price exceeds marginal cost, the firm obtains a producer surplus:

Producer surplus: The price of a good minus the marginal cost of producing it, summed over the quantity produced

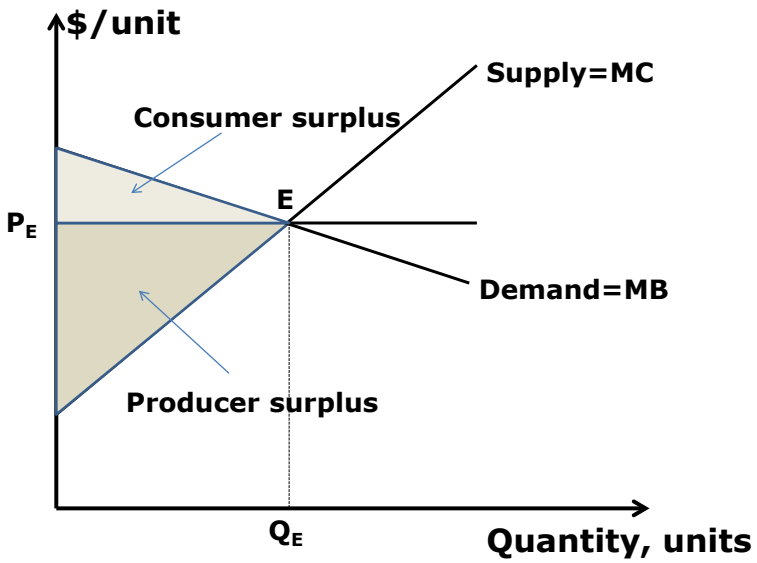
In a sense, producer surplus is the firm's short-run profit. The following graph shows the producer surplus according to its definition (shaded area):



So graphically, producer surplus is the area under market price and above the MC-line or marker supply in general since market supply is the sum of individual MC-lines.

Since the demand curve reflects marginal benefits while the supply curve reflects marginal costs, then when the demand curve and the supply curve intersect (perfectly competitive market equilibrium), marginal benefits equal marginal costs or *MB = MC*. As we learned previously, the condition *MB = MC* is the condition for allocative efficiency.

Therefore, a competitive equilibrium is efficient. In a competitive equilibrium, resources are used efficiently to produce goods and services that people value the most. And when the competitive market uses resources efficiently, the sum of consumer surplus and producer surplus, which is called total *social surplus* or *social welfare*, is maximized. The following graph illustrates the above discussion:



The great insight of the early economists was that an economy based on free-market transactions is self-organizing. Scottish economist and political philosopher Adam Smith (1723–1790), was the first to develop this insight fully.

Smith wrote that a society whose economy is organized by free markets produces ordered behavior that makes it appear as if people are guided by "*an invisible hand*". He did not literally mean that a supernatural presence guides economic affairs. Instead, he referred to the amazing order that emerges out of so many independent decisions.

The key to the explanation is that all individuals respond to the same set of prices, which are determined in markets that respond to overall conditions of scarcity. Free markets contain many buyers and sellers, and all of them are interested primarily in their own well-being. Yet, despite decentralized decision-making and self-interested decision-makers, the result is not chaos but efficiency. So, all participants in the economy, consumers (buyers) and producers (sellers) are motivated by self-interest, and the "invisible hand" of the market guides this self-interest into promoting general economic well-being.

By tradition, a free-market economy is usually compared with a centrally planned economy like the Soviet Union's before August of 1991. In the former Soviet Union, all prices were set by the government. They were not a result of interaction between market demand and market supply, and therefore, the prices were not subject to the 'invisible hand" framework. It means that economy of the Soviet Union was predominantly inefficient.

In order to illustrate how the invisible hand works, let us consider the following example. There was the so-called first oil shock in 1973–1974. As a result, the price of oil increased in four times. Consequently price of gasoline increased as well. It resulted in people buying smaller economic cars. Demand for large cars therefore decreased which resulted in lower number of large cars produced at a lower price (can you present this situation graphically using concepts of demand, supply and equilibrium?). However, at the same time demand for small cars increased resulting in a greater number of small cars produced at a higher price (again, try to present this situation graphically). The loss of the sale of large cars caused some shut-downs and layoffs in that market, but those resources were immediately employed in the market for small cars. Resources moved from a declining sector to an expanding one not due to a notorious central planner, but rather due to people's self interest. As a result, the economy found efficient allocation automatically.

Let us summarize all advantages of a perfectly competitive market discussed in this chapter:

- Consumers face the lowest price
- Producers produce at the least costs
- Some short-run opportunities arise but in the long-run all producers earn normal profit
- Social welfare as the sum of consumer surplus and producer surplus is maximized

5.8 Fairness of markets

Does a competitive, free market provide people with fair share of resources? Economists agree that increasing efficiency or making a larger "economic pie" and baking it at a lower cost per slice add to economic prosperity of a society as a whole. However, they do not agree on what fair shares of the economic pie for all people who make it should be. The reason is that ideas of fairness are not exclusively economic ideas. They involve subjective values, ethics, and politics. All ideas associated with fairness are based on a fundamental principle that seems to be hard-wired into human brain called the symmetry principle:

Symmetry principle: The requirement that people in similar situations be treated similarly

It is the principle that lies at the centre of all main moral codes that tell us, in some form or another, to behave towards other people in the way we expect them to behave towards ourselves. Unfortunately, the symmetry principle does not deliver universally accepted prescriptions for fair arrangements and leaves the big question, formulated at the beginning f this section, unanswered.

Therefore, it turns out that economic efficiency is a well defined concept with clear criteria while fairness is a value judgment which depends on subjective opinion. Most economists eventually came to a conclusion: What is efficient most likely is fair. However, again it is just a conjecture. A *conjecture* is a proposition that is unproven but is thought to be true and has not been disproven.

6 Monopoly and Market Power

Key concepts discussed in this chapter: barriers to entry, natural monopoly, public utilities, legal monopoly, single-price monopoly, price-discriminating monopoly, level analysis, marginal analysis, deadweight loss, first degree price discrimination, second degree price discrimination, third degree price discrimination, marginal cost pricing, average cost pricing, second best, market power, predatory pricing, product tying, overcapacity, Lerner Index, Herfindahl-Hirschman Index

6.1 Reasons for monopoly

Monopoly: A market for a good or service that has no close substitutes and in which there is one supplier that is protected from competition by a barrier preventing the entry of new firms

Monopoly like perfectly competitive market is another extreme of market structure. Monopoly exists when a specific individual or an enterprise has sufficient control over a particular product or service to determine significantly the terms on which other individuals should access to it. Monopolies are thus characterized by a lack of economic competition for the good or service that they provide and a lack of viable substitutes. In a monopoly there is one seller of the monopolized good/service who produces all the output. The firm and industry are identical.

Based on the above definition, it is possible to claim that monopoly arises when there are:

- No close substitutes
- Barriers to entry

If a good or service has a close substitute, even though only one firm produces it, that firm effectively faces competition from the producers of substitutes. Hence the arrival of a new product can also create a monopoly.

Anything that protects a firm from the arrival of new competitors is a barrier to entry. In general, there are two types of barriers:

- Natural
- Legal

A natural barrier leads to a natural monopoly:

Natural monopoly: A monopoly that arises because one firm can meet the entire market demand at a lower cost than two or more firms

Classic definition of natural monopoly was given by William Baumol in 1977:

"[*a*]*n industry in which multiform production is more costly than production by a monopoly*"

Natural monopolies are common in markets for important products that require an expensive infrastructure (or in general fixed costs) to deliver the good or service, such as in the cases of water supply, electricity, natural gas, and other industries known as public utilities.

Public utilities: Businesses that provide the public with necessities, such as water, electricity, natural gas, and telephone and telegraph communication

In transportation, railways are often considered a typical example of a natural monopoly with high costs of laying tracks and building a network, as well as the costs of buying or leasing the trains.

In terms of economic theory, natural monopolies are associated with large economies of scale or what we previously defined as increasing returns to scale. Graphically this phenomenon means a decreasing long-run average cost function for all output ranges. Later in this chapter theoretical rationale for natural monopolies is discussed.

Legal barriers to entry create legal monopoly:

Legal monopoly: A market in which competition and entry are restricted by the concentration of ownership of a natural resource or by the granting of public franchise, government license, patent, or copyright

According to the definition of legal monopoly, the following are reasons for such monopoly to arise:

- Control over important natural resources or in general inputs of production
- Exclusive rights granted by the government
- Mergers and takeovers

For example, De Beers is a family of companies that dominate the market for diamonds. It is so because De Beers controls more than 80% of the diamond deposits. In many countries, postal services are provided by public franchises, which are legal monopolies according to our definition. Patents and copyrights are also examples of exclusive rights granted by the government. When a pharmaceutical company comes up with a new drug it is given exclusive rights to produce this drug for some period of time. Mergers happen all the time in different segments of economy as well as takeovers of smaller firms by larger ones.

Monopolist is the only producer in a market, and therefore the monopolist faces the entire down-sloping market demand. It means that in order to sell a larger quantity, the monopolist must set a lower price. From a theoretical standpoint, in a monopolistic market there are two price-setting possibilities that create different trade-offs:

- Single price
- Price discrimination

Price-discriminating monopoly: A monopoly that is able to sell different units of a good or service for different prices

The latter is possible only if the monopolist can separate consumers with different preferences and prevent a product re-sale.

6.2 Single-price monopoly: Price and marginal revenue

As defined previously, total revenue is the price multiplied by the quantity sold. Marginal revenue is the change in total revenue resulting from a one-unit increase in the quantity sold. However, when a monopolist increases the quantity by one-unit, it has to lower price to sell that unit. It comes from the notion that the monopolist faces the entire down-sloping market demand.

When the price is lowered to sell one more unit, two opposing forces affect total revenue:

- The lower price results in revenue loss (the price effect)
- The increased quantity sold results in the revenue gain (the quantity effect)

We know from our previous discussion of the link between the price elasticity of demand and total revenue that a price decrease increases total revenue only if demand is elastic. That is why a monopoly produces only along elastic portion of the market demand in terms of point-price elasticity (remember: point price elasticity of demand usually changes along the demand curve!).

Mathematically, total revenue TR is price P times quantity Q or

$$TR = P \times Q$$

The change in total revenue can be presented as follows (if we totally differentiate the above expression):

$$\Delta TR = Q\Delta P + P\Delta Q$$

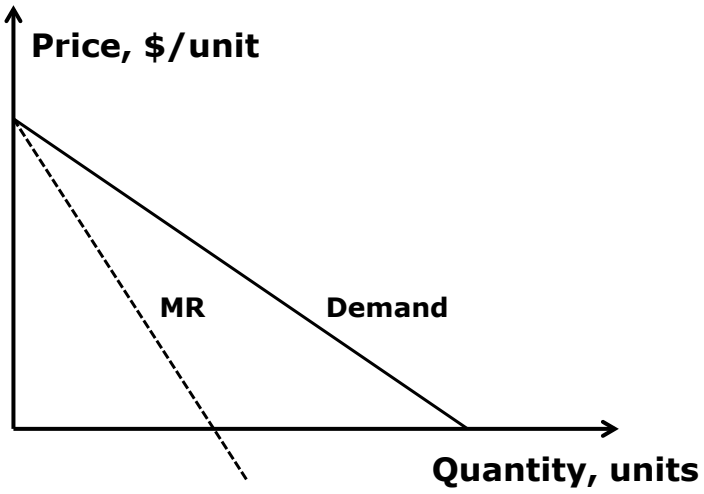
By definition, marginal revenue MR is the change in total revenue due to a one-unit increase in quantity. Therefore, if we assume $\Delta Q = 1$, previous expression becomes:

$$MR = Q\Delta P + P$$

The first term in the right-hand side of the above equation is definitely negative in the case of a price decrease because ΔP is negative. Two conclusions follow:

- Marginal revenue is always less than the price of a product
- Marginal revenue is positive at low levels of quantity produced and sold, but diminishing with an increase in the quantity

That is why marginal revenue curve is derived from the market demand using definitions of total revenue and marginal revenue as explained above. Graphically, the following relationship between market demand and marginal revenue is observed:



Moreover, if market demand is a linear function (a straight line like the one in the above diagram), *MR* is also a straight line with the slope in two times steeper than the demand's or the MR-line cuts horizontal intercept of the market demand line in half. In order to see this, let us present a simple mathematical explanation.

Linear inverse demand (remember, graphically we plot the inverse demand) can be presented as

$$P = a - bQ$$

Given definition of total revenue, in this case it can be written as

$$TR = P \times Q = (a - bQ) \times Q = aQ - bQ^2$$

or total revenue is a quadratic function in quantity given linear demand function. Marginal revenue is derivative of total revenue with respect to quantity or

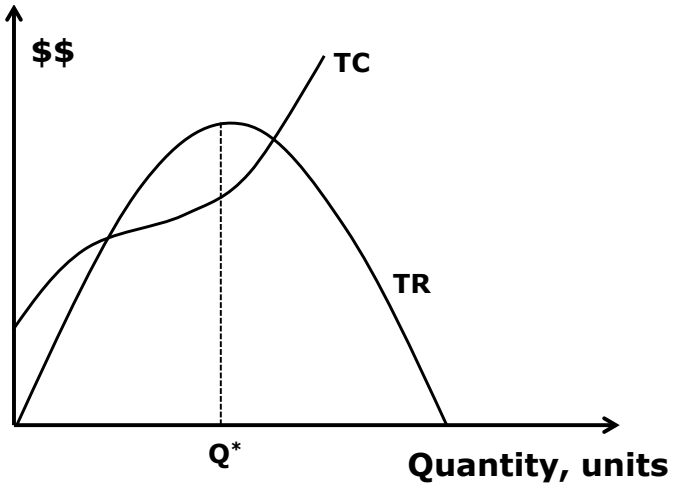
$$MR = \frac{dTR}{dQ} = a - 2bQ$$

It means that graphically MR-line starts with the same vertical intercept as the inverse demand but has slope in two times steeper.

6.3 Single-price monopoly: Output and price decision

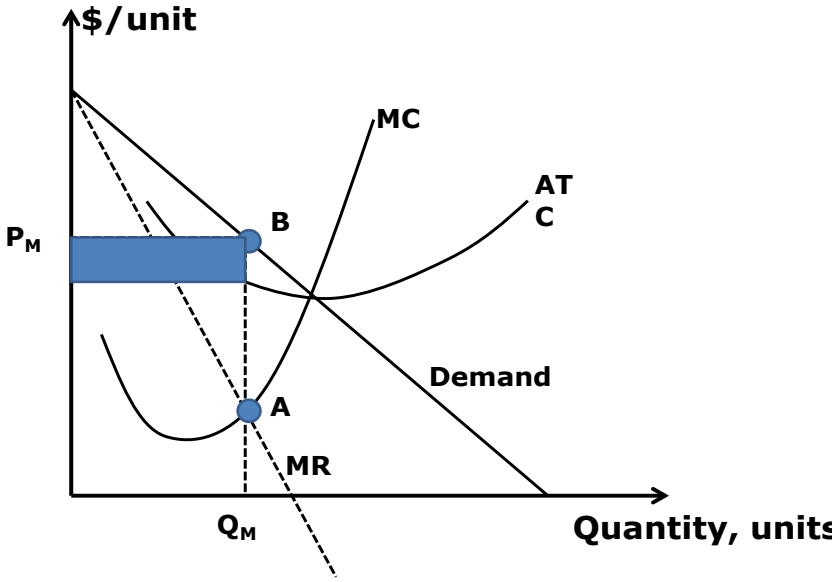
In order to determine the output level and price that maximize a monopoly's profit, two types of economic analysis can be applied: (i) level analysis, and (ii) marginal analysis.

In *level analysis*, we study the behavior of total revenue *TR* and total cost *TC*. Let us assume that market demand is linear function which means that total revenue is a quadratic function as discussed in previous section. Graphically it is presented by a parabola. Shape of the total cost function remains the same as before. This situation is depicted below:



From the above graph we can see that the profit-maximizing output is Q^*. At this level of output, the gap between the TR-line and TC-line is the largest which corresponds to the maximum profit.

Now let us employ *marginal analysis* and plot demand, marginal revenue and marginal cost curves:



For a perfectly competitive firm, marginal cost equals price or $MC = P$ is a profit-maximizing condition. The monopolist must take into account the fact that its production decision affects price. Profit of the monopolist is maximized if $MC = MR$ (point A). In a mathematical sense, the profit-maximizing condition of a perfectly competitive firm $MC = P$ is a special case of the general condition $MC = MR$ since for a perfectly competitive firm $P = MR = const$.

Why $MC = MR$ is the profit-maximizing condition of the monopolist?

If $MR > MC$, the monopolist gains profit by increasing output

If $MR < MC$, the monopolist gains profit by decreasing output

That is why adjustment in output eventually leads to $MR = MC$.

In the above diagram, the profit-maximizing condition produces profit maximizing output Q_M. Monopoly price P_M is determined by demand at the profit-maximizing quantity (point B). And finally, the shaded area represents profit since the area of the shaded rectangle can be expressed as follows:

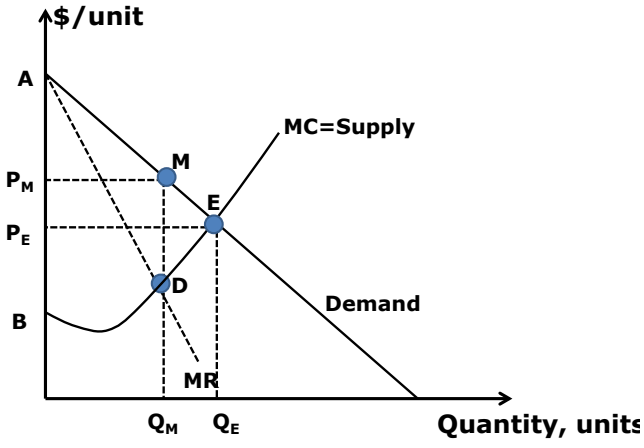
$$(P_M - ATC) \times Q_M = P_M \times Q_M - ATC \times Q_M = TR - TC = Profit$$

where P_M is the monopoly's price, Q_M is the monopoly's quantity, ATC is average total cost, TR is total revenue and TC is total costs.

So, we can see now that the monopolist earns positive economic profit. Positive economic profit is an incentive to enter the market. However, barriers to entry prevent that from happening. Therefore, in the monopoly case, the firm can make economic profit indefinitely. Sometimes a monopoly can incur a loss in the short- run because of an increase in fixed cost (e.g., research and development costs), but it is a temporary phenomenon. In the long-run, monopoly always earns positive economic profit.

6.4 Single-price monopoly and perfect competition compared

Let us analyze the following graph:



If there is a competitive market, there are many small firms producing identical product. Supply is a result of horizontal summation of the firms' individual marginal cost curves since they are individual supply functions. Hence, the supply curve in the above diagram reflects this fact. If we take into account demand, then the perfectly competitive market equilibrium occurs at point E with P_E as equilibrium (market) price and Q_E as equilibrium market quantity.

Now suppose that a single firm takes over or it buys all the firms in the market. Demand curve does not change, but since the firm becomes a monopoly, it now sets its quantity according to the monopoly's profit maximizing rule *MR= MC*. Price P_M comes from the demand evaluated at quantity Q_M.

Conclusion: Monopoly produces less and chargers a higher price than a perfectly competitive market would.

Now let us apply some concepts from the welfare economics in order to analyze the situation in more detail. In doing so, we have to identify consumer surplus and producer surplus:

In the case of a perfectly competitive market, consumer surplus is the area of triangle P_EAE. Producer surplus is the area P_EEB (recall definitions of the consumer and producer surplus and their geometric interpretations).

In the case of monopoly, consumer surplus is area P_MAM and it is smaller than it was under perfect competition. Producer surplus is area P_MMDB and it is larger than it was under perfect competition. However, if we evaluate total surplus as the sum of the consumer surplus and producer surplus, we will see that total surplus or total welfare under monopoly is less by the area DME.

Therefore, there is a loss in total welfare in the case of a monopoly. The identified triangle DME is called the *deadweight loss* which points to inefficiency of the monopoly since total welfare of a society is not maximized under monopoly.

6.5 Price-discriminating monopoly

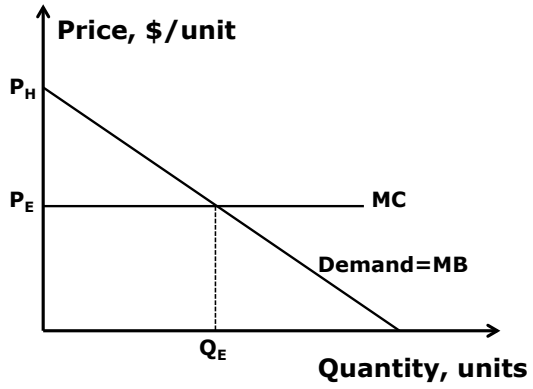
So far we have considered monopolists that charge the same price to all consumers. However, if a monopolist is able to identify and separate different types of buyers (consumers), the monopolist is going to price discriminate them.

Price discrimination: A practice of selling the same good at different prices to different consumers

In economic literature, three degrees of price discrimination are discussed.

In *first degree price discrimination*, price varies by customer's willingness or ability to pay. It is sometimes called perfect price discrimination: Monopolist knows exactly maximum willingness to pay (WTP) of each consumer and charges the consumers according to their maximum WTPs. By knowing the maximum WTPs, the monopolist is able to absorb the total surplus, thus taking the consumer's surplus from the consumers and transforming it into revenues.

Graphically this phenomenon can be shown as follows. Suppose that marginal cost *MC* is constant which means that MC-line is horizontal.



Perfectly price discriminating monopolist will charge prices between P_H and P_E since there are consumers with corresponding WTPs for the product bringing the total quantity produced and sold to Q_E. Obviously monopolist cannot charge $P < MC$ since it would make losses in such a case.

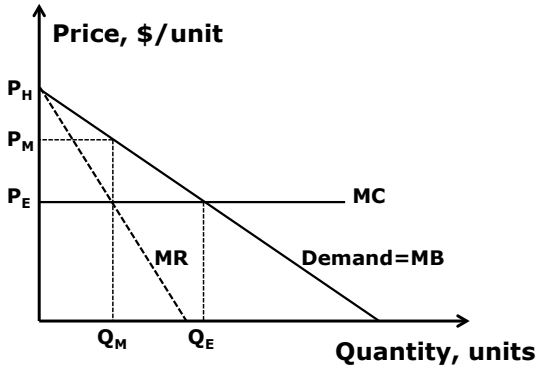
It appears to be that perfectly discriminating monopolist is efficient since it produces at the efficient level associated with allocative efficiency condition $MB = MC$. However, total surplus is transferred to the monopolist.

One example of the first degree price discrimination would be a tuition fee based on everyone's ability to pay. If University authorities knew exactly each student's WTP, they would charge each student exactly this amount. However, this can only happen theoretically since the monopolist does not know WTPs of its consumers.

Hence most monopolists cannot charge a separate price for each unit of output. Instead, they sell blocks of output for different prices. This is consistent with what is called *second degree price discrimination*:

Second degree price discrimination: *A practice of charging different prices depending on how much of a good or service the consumer buys*

Graphically this situation can be depicted as follows. Again constant marginal cost is assumed:



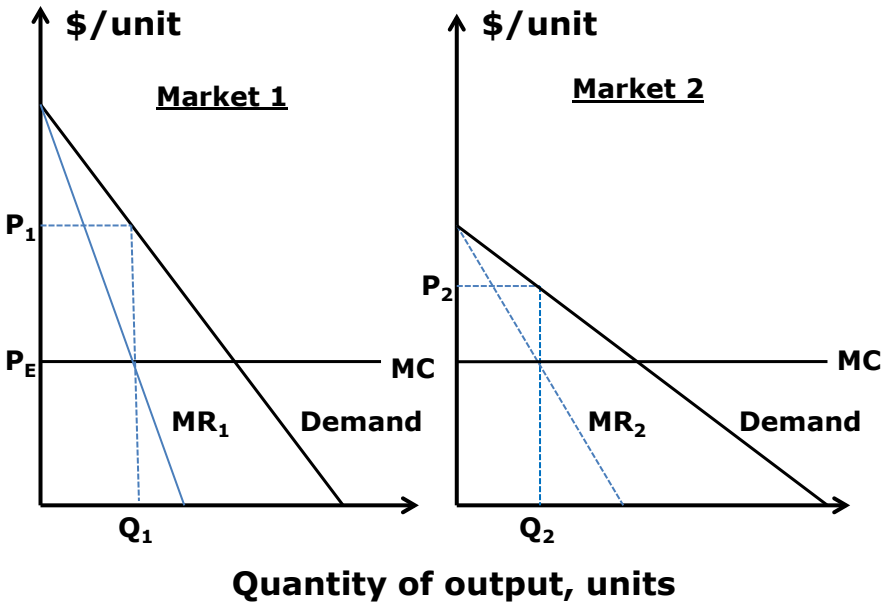
In this case, price discriminating monopolist will charge prices between P_M and P_E capturing a fraction of the consumer surplus and deadweight loss. Again it is an improvement from efficiency standpoint.

Example of this technique is in-store discounts like "buy one and get another one for 50% of the price". Another example is electricity prices. Utility companies charge more for the first block of electricity consumed than for subsequent blocks of electricity used in the same month. This is known as a two-tariff pricing. It reflects the fact that consumers typically place a higher marginal benefit on the first block of output than on subsequent blocks of output.

There is also *third degree of price discrimination*:

Third degree price discrimination: *A practice of charging different prices for the same good or service in different markets*

In business practice this is known as price segmentation. Graphically it is possible to depict this situation as follows:



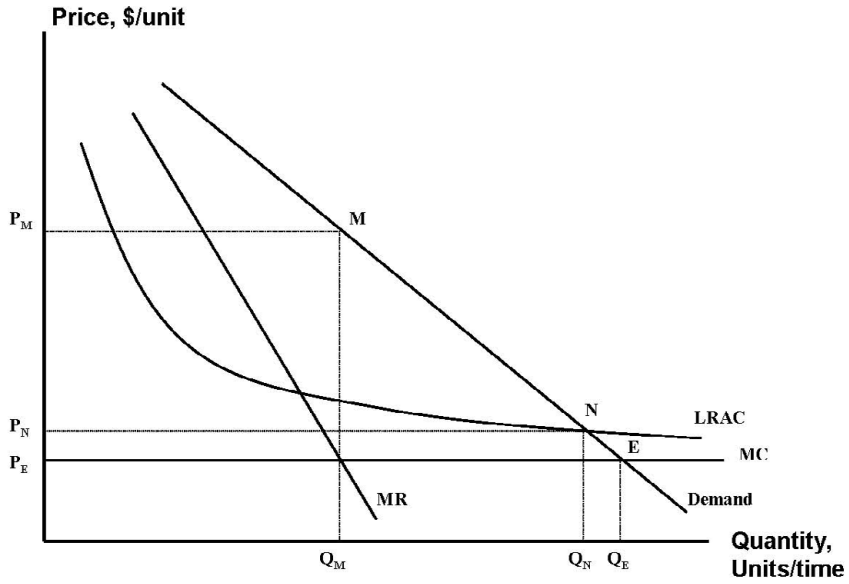
As seen from the above diagram, prices charged in two markets differ because demand functions of the consumers differ while MC-line remains the same since it is the same producer/provider. In fact, demand functions are associated with different price elasticities which reflect different consumer tastes.

Example is air travelling: In general, vacationers pay less buying tickets in advance while business people pay higher price buying tickets at the last moment. The monopolist (air transportation provider) sets a lower price for vacationers because they have a more elastic demand than businessmen. The monopolist is once again capable of capturing more surplus than would be possible without price discrimination.

In all cases discussed, a price discriminating monopolist tries to capture consumer surplus partially (second and third degrees) or entirely (first degree) plus a fraction of deadweight loss. A price discriminating monopolist can increase output to the point at which price equals marginal cost. This output is identical to that of perfect competition. Overall, from an efficiency viewpoint, the outcome of a price discriminating monopolist is better than that of a single-price monopolist because it is associated with lower deadweight loss.

6.6 Natural monopoly

The following diagram shows the demand curve, the marginal revenue curve MR, the long-run average cost curve LRAC, and the marginal cost curve MC in the case of a natural monopoly:



If the firm is unregulated and it maximizes profit, it will set *MC* = *MR* which corresponds to the quantity Q_M in the above diagram and the price P_M. This is a single-price monopoly price setting (point M).

Efficient use of resources is achieved when marginal benefit *MB* equals marginal cost *MC*. Since the demand curve reflects *MB*, then point E is associated with the efficiency condition. This is known as the marginal cost pricing. In the case of natural monopoly, this rule leaves the firm with negative profit since at quantity Q_E we observe $LRAC > P_E$.

Therefore, first best solution, which is based on *MB*=*MC* condition, is not available while the monopoly situation is not desirable. The latter is not desirable since in such a case price is too high while quantity produced is too low. That is why markets subject to natural monopolies are regulated.

Regulators almost never use marginal cost pricing rule in the case of a natural monopoly. Instead, they use an average cost pricing rule, which sets the price equal to LRAC (point N) to enable the firm to cover its costs and earn normal profit. In the graph above, such pricing leads to P_N with the quantity produced Q_N. The average cost pricing rule is known as the second best solution.

In order to set average cost pricing rule correctly, an authority must know the cost structure of a private producer. This information is not always available which creates a problem of asymmetric information.

6.7 Market power

The simplest definition of market power is:

Market power is the ability of a seller (producer) to maintain prices above competitive levels for a significant period of time.

In other words, it is the ability to set prices above marginal costs. As we saw in the monopoly case, it is not only inefficient but also leads to some undesirables consequences such as:

- Price discrimination
- Predatory pricing
- Product tying
- Creation of overcapacity

Price discrimination has been already discussed. *Predatory pricing* is the practice of selling a product or service at a very low price, intending to drive competitors out of the market, or create barriers to entry for potential new competitors. Under *product tying*, consumers are required to buy something they do not want as a condition of buying something they do want. *Creation of overcapacity* is associated with the situation in which an industry or business cannot sell as much as it produces.

There are different sources and degrees of market power. Monopoly possesses the highest market power. However, although a monopoly's market power is high, it is not absolute. A monopoly faces a negatively sloped market demand curve. Consequently, any price increase will result in the loss of some customers. It means that a monopoly has the power to set prices or quantities although not both. In general, monopolies derive their market power from barriers to entry – circumstances that prevent or greatly impede a potential competitor's entry into the market or ability to compete in the market. From a quantitative perspective, the two primary factors determining monopoly's market power are the market demand curve and the firm's cost structure.

Monopolistically competitive firm has some degree of market power. In this case, market power means that the firm has control over the terms and conditions of exchange. The firm can raise it prices without losing all its customers. The firm can also lower prices without initiating a "price war" with competitors. A monopolistically competitive firm derives its market power from the fact that it has relatively few competitors, competitors do not engage in strategic decision making, and the firms sell differentiated product.

Oligopolistic markets are characterized by strong market power. It follows from the facts that oligopolies are price setters rather than price takers and barriers to entry are high. The most important barriers are economies of scale, patents, access to expensive and complex technology and strategic actions by current firms designed to discourage or destroy emerging firms. Unlike monopolistically competitive firms oligopolies can retain long run abnormal profits. High barriers of entry prevent sideline firms from entering market to capture excess profits.

The starting point in looking for market power in any market is the competitive price level. Pricing above the marginal, or incremental, cost of a good or service cannot be regarded per se as the only evidence of market power. In real world markets, the competitive price level will often be higher than marginal cost. In industries with high fixed costs, such as, for example, transportation, prices must include mark-ups over marginal costs for firms to break even across their whole business.

Regulated prices like in the case of natural monopolies may also be an inappropriate starting point for detecting market power, as they may differ from competitive price levels. For example, in many countries prices for certain "basic" telephone services are regulated and set below their economic cost to meet universal service goals. In these circumstances market power cannot, and should not, be inferred by comparing any given firm's price to the regulated price level.

Another point for the existence of market power is that the price increase must be sustainable. Firms may be able to temporarily increase prices above competitive levels, for example due to opportunistic behavior or based on innovation. However, in the absence of market power, such price increases are unsustainable. True market power requires that the firm be able to profitably implement the price increase for a significant period of time.

A high market share also does not necessarily infer market power. Firms may gain high market shares through means other than market power. A firm's market share may increase, at least temporarily, due to a successful new invention or better customer service.

Alternatively, a firm may have a high market share for historical reasons. For example, incumbent telecommunications firms were once monopoly franchises in most countries and have high market shares as a result. As competition emerges, an incumbent's market share cannot guarantee it the ability to charge prices higher than its competitors.

Market share in itself is neither necessary nor sufficient for market power. Firms with high market shares may be constrained from raising prices by a range of factors, including:

- Competition from other suppliers already in the market
- The potential for competition from new entrants, and
- The "countervailing power" of customers in the market, for example their willingness to do without the service if the price increases.

Therefore, in general all of the above features should be taken into account in order to make a decision about market power.

In addition, several quantitative measures exist that can help assess whether a firm may have market power. Lerner Index is a measure of price and it is defined as

$$L = \frac{P - MC}{P}$$

where P is the market price set by the firm and MC is the firm's marginal cost. The index ranges from a high of 1 to a low of 0, with higher numbers implying greater market power. For a perfectly competitive firm $L = 0$ since $P = MC$. Such a firm has no market power. The main problem with this measure, however, is that it is almost impossible to gather the necessary information on prices and particularly marginal costs.

The *Herfindahl index*, also known as *Herfindahl-Hirschman Index* or HHI, is a measure of the size of firms in relation to the industry and an indicator of the amount of competition among them. This economic concept is widely applied in competition law, antitrust and also technology management. It is defined as the sum of the squares of the market shares of the 50 largest firms (or summed over all the firms if there are fewer than 50) within the industry, where the market shares are expressed as fractions. The result is proportional to the average market share, weighted by market share. As such, it can range from 0 to 1, moving from a huge number of very small firms to a single monopolistic producer. Increases in the HHI generally indicate a decrease in competition and an increase in market power.

7 Market Failures and Government Intervention

Key concepts discussed in this chapter: market imperfections, market failures, externalities, positive externalities, negative externalities, marginal private costs, marginal external costs, marginal social costs, internalization of externality, direct regulation, incentive policy, voluntary solution, property rights, emission charges, Pegovian tax, marketable permits, marginal abatement costs, marginal private benefits, marginal external benefits, marginal social benefits, rivalness, excludability, private goods, public goods, common property resources, open access resources, club goods, free riding, asymmetric information, adverse selection, moral hazard, principle-agent problem, signaling, screening, marginal tax on producers, tax incidence, buyer's price, seller's price, tax revenue, excise tax, income tax, payroll tax, excess burden of taxation, price ceiling, price floor

7.1 Market failures versus market imperfections

We have already learned that in general the invisible hand of a market automatically brings equilibrium which is efficient. However, there are some situations when the invisible hand of the market breaks down. In economic theory these situations are known as market failures.

Market failure: Inability of a free market to allocate resources in the most efficient way

However, it is necessary to distinguish between market failures defined above and market imperfections defined as follows:

Market imperfections: Deviations from perfect competition

In both cases, natural market forces of demand and supply do not function to create efficient market equilibrium. In the case of market imperfections, the key is the existence of some market structures other then perfectly competitive market. Monopolistic competition, monopoly and oligopoly discussed previously are examples of market imperfections. As we learned before, all these market structures produce inefficiencies associated with market power.

In the case of marker failures, there is competitive market but it is not enough to achieve overall economic efficiency. In general, economic theory identifies three sources of market failures:

1. Externalities
2. Public goods
3. Asymmetric information

7.2 Externalities

An important requirement for the invisible hand to guide markets in society's interest is that market transactions have no side effects on non-market participants. Such side effects are called *externalities* which are the effect of a decision on a third party that is not taken into account by the decision maker. Formal economic definition of externality is:

Externality: A cost or benefit that arises from production that falls on someone other than producer; or a cost or benefit that arises from consumption that falls on someone other than consumer

As seen from the above definition, externalities are associated with situations in which economic agents other than market participants – producers and consumers – are affected by the side effect of market activity.

Externalities can arise from either a production activity or a consumption activity. As well, an externality can be either a negative externality, which increases costs, or a positive externality, which increases benefits.

Air pollution by producers is the best example of a *negative production externality*. For example, a chemical plant produces fertilizer which is consumed by farmers. As a result, the plant emits nitrogen oxides which affect a nearby community. People who live there suffer from the pollution. Since pollution is a by-product of the fertilizer production it is not a part of the producer's decision resulting in some external costs associated with consequences of air pollution. Producer's marginal costs are associated with marginal costs of producing fertilizer, not nitrogen oxides. And in general, people who live nearby are neither producers nor consumers of fertilizer.

A classic example of a *positive production externality* is an owner of an orchard who grows apples and an owner of bees who produces honey. Close location of the two causes positive externalities since bees collect nectar from apple trees and pollinate them at the same time. As a result, the production of both, apples and honey increases. Thus, there is an external production benefit not taken into account by the market.

Smoking tobacco is an example of a *negative consumption externality*. It affects people other than smokers, and there are some external costs imposed on non-smokers in such a case.

Education usually creates *positive consumption externalities*. Society that has more people with university diplomas benefits in terms of new knowledge spread across various sectors of an economy. This results in some external consumption benefits.

So, as the above examples show, market activity may result in by-products that affect people who directly do not participate in the market activity. In such cases, economists say: "External costs or external benefits are generated". In order to understand economic consequences of externalities, it is necessary to distinguish between private costs and benefits and social costs and benefits.

7.3 Private costs versus social costs

It turns out that in general social costs involve both private costs and external costs. Since the fourth microeconomic principle states that we have to apply marginal analysis let us define the relevant costs in terms of marginal costs:

Marginal private cost: The cost of producing an additional unit of a good or service that is borne by the producer of that good or service

It turns out that marginal cost *MC* we discussed so far has been marginal private cost.

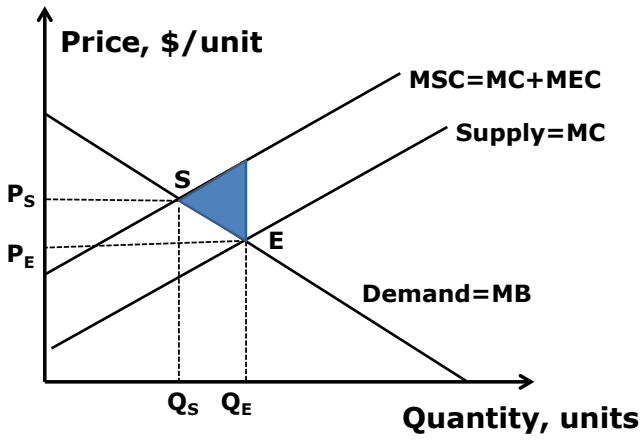
Marginal external cost: The cost of producing an additional unit of a good or service that falls on people other than the producer

Marginal social cost: The marginal cost incurred by the entire society – by the producer and everyone else on whom the cost falls

Hence, marginal social costs, *MSC* is the sum of marginal private costs, *MC* and marginal external costs, *MEC* or

$$MSC = MC + MEC$$

As pointed out earlier, air pollution is a good example of a negative production externality that results in marginal external costs. Let us analyze this situation graphically given the above definitions:



Market supply reflects marginal private costs, and it is marked as MC in the above diagram. The demand curve is the marginal benefit curve MB (to be precise, it reflects marginal private benefits discussed later). In the above diagram, market equilibrium occurs at point E with equilibrium price P_E and quantity Q_E. However, at this equilibrium, marginal social costs *MSC* exceed marginal benefits *MB* or *MSC* > *MB* by the amount of marginal external costs *MEC* (graphically, this is vertical distance between MSC line and supply). From the standpoint of a society, equilibrium at point E is not efficient while equilibrium at point S with price P_S and quantity Q_S is.

Private producer would produce at point E which causes the deadweight loss (from the standpoint of a society) represented by a shaded triangle in the diagram. As we know, the deadweight loss is a measure of inefficiency. In turn, society as a whole is better off at point S – social equilibrium which is the result of intersection between market demand and marginal social costs MSC line.

In general, the presence of an externality results in inefficient resource allocation by a free market. When such situation occurs, there is need for government intervention to correct it. Economic strategy that is usually applied in these circumstances is called *internalization of externalities*. Internalization of externality means changing the incentives of the parties involved so that they begin to act as if there is a market for the external costs or benefits. As a result, the overall efficiency of the market improves.

There are different ways to internalize externalities:

- Direct regulation
- Incentive policies
- Voluntary solutions

Under *direct regulation*, the amount of a good or service people are allowed to use is directly regulated by the government. In general, direct regulation is not efficient from standpoint of economic theory. That is why economists tend to like *incentive policies* to deal with externalities. At large, two types of such policies are popular: (i) tax incentive policies, and (ii) market incentive policies. Below examples of incentive policies are discussed with respect to air pollution. *Voluntary solutions* to address externalities are based on individual choice of the parties involved through negotiations.

Sometimes internalization of externalities can be achieved by establishing property right where the one did not exist before:

Property rights: Legally established titles to the ownership, use, and disposal of factors of production and goods and services that are enforceable in the courts

In our previous example of a chemical plant producing fertilizer, suppose that the polluter owns homes affected by his/her emissions. Housing rent would go down because of emissions, and therefore, the producer is confronted by the costs of its own pollution. In this case, property right (ownership) would make external costs a part of the producers decision or external costs would become internal to the polluter.

7.3.1 Air pollution as an example of negative production externality

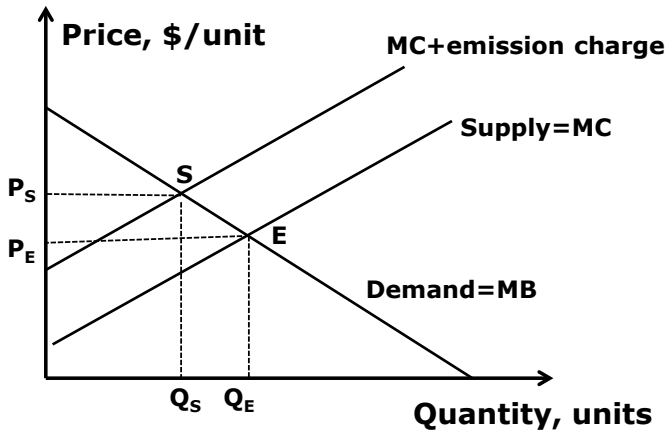
In general there are three types of incentive policies a government can implement to internalize this externality:

- Emission taxes (charges)
- Marketable permits
- Taxes

Emission charges confront a polluter with the external cost of pollution and provide an incentive to seek new technologies that are less polluting. This method of dealing with environmental externalities has been used only modestly in North America, but it is common in Europe.

Basic idea of the method is: You can pollute as much as you want but you will pay for it! The more pollution a firm creates, the more it pays in emission charges. This charge creates an incentive for firms to produce less and to use at some point a technology that creates less pollution in order to reduce the emission charge.

Basic requirement in terms of efficiency is: Efficient charge should be exactly equal to the marginal external

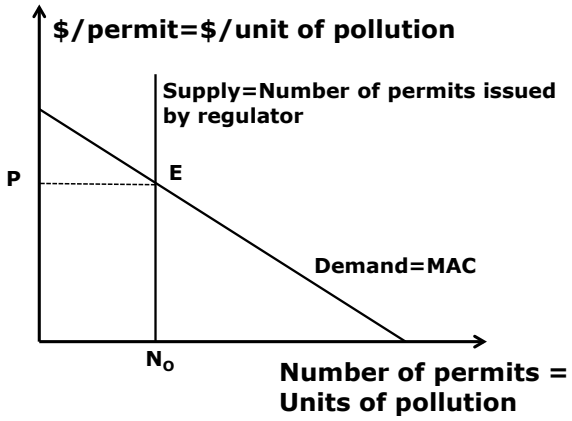


In the above diagram, if emission charge is set equal to the vertical distance between Supply and MSC lines in the previous diagram, which is in fact marginal external cost MEC, then the resulting price P_S and quantity Q_S will be efficient from the standpoint of a society. This is known as *Pegovian tax* after the name of a British economist Arthur Pegou (1920). In order to impose it in efficient way an authority (regulator) has to know marginal costs of a polluter.

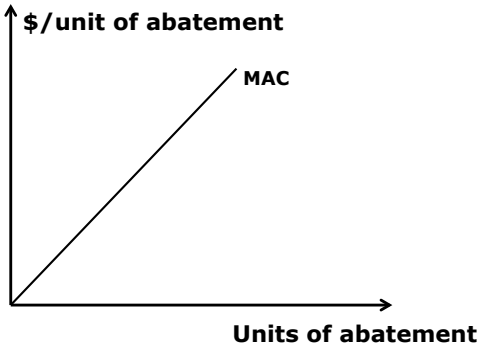
Marketable permits are another example of incentive policies. They are a clever way of overcoming the need for the regulator to know every firm's marginal cost schedule. The following three steps should be performed in this regard:

1. It is necessary to introduce a new right – the right to pollute
2. Issue permits according to the rule: One permit = one unit of pollution; set a goal in terms of desired level of pollution
3. Auction the permits and let the holders trade them

In this framework, initial number of permits N_O issued by the regulator is a fixed supply of permits. Graphically it is represented by the vertical line:



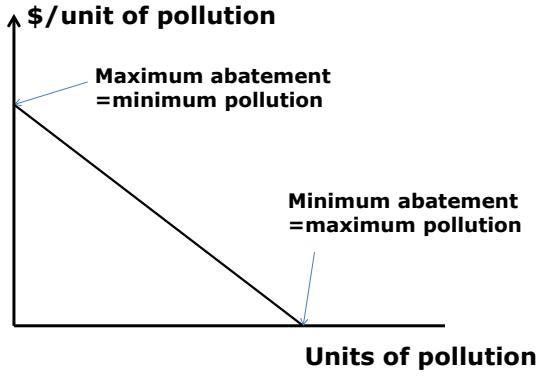
Demand for permits coincides with the polluter's marginal abatement cost MAC since it reflects each polluter's willingness to buy permits. Demand for permits is constructed from the original MAC which is



and the fact that pollution and abatement are inversely related in the following way

$$Pollution = Initial\ pollution - Abatement$$

Therefore, in terms of pollution MAC becomes



Price of a permit *P* arises as a result of market forces of demand and supply. In the end, the price plays the role of emission charge.

If a polluter finds a cheaper way to deal with pollution, the polluter will sell some permits. When government wants to introduce a strict control, it will buy some permits shifting supply of permits to the left. As well, "green" groups can participate in the market by buying some permits and reducing their supply.

This method of dealing with pollution provides an even stronger incentive than emission charges to find technologies that pollute less because the price of a permit rises as the demand for permits increases.

The government can also use *taxes* as an incentive to cut back on an activity that creates an external cost. By setting the tax rate equal to the marginal external cost, firms can be made to behave the same way as they would if they bore the cost of externality directly. Theoretically taxes work the same way as emission charges. Practically, an emission charge is imposed per unit of pollution, while a tax is imposed per unit of a good or service produced.

7.4 Private benefits versus social benefits

Previously we analyzed private, external and social costs. Benefits can be defined in the same way:

Marginal private benefit: The benefit of an additional unit of a good or service that the consumer of that good or service receives

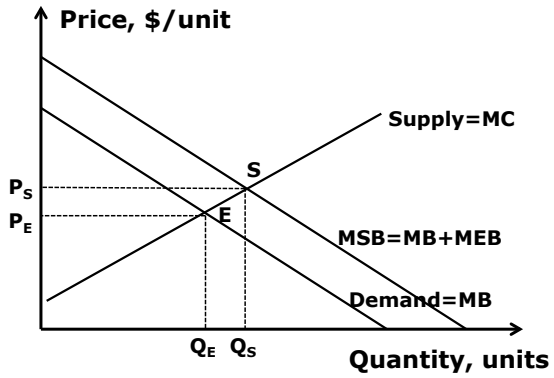
Marginal external benefit: The benefit from an additional unit of a good or service that people other than the consumers of the good or service enjoy

Marginal social benefit: The marginal benefit enjoyed by society as the sum of marginal private benefit and marginal external benefit

Market demand reflects marginal private benefits *MB*. Presence of a positive externality results in marginal social benefits *MSB* which is the sum of marginal private benefits *MB* and marginal external benefits *MEB*:

$$MSB = MB + MEB$$

Graphically *MSB* can be represented by a higher demand curve and technically analysis in this case is similar to the one associated with the social costs: Two equilibria arise – private equilibrium given by interception of *MC* (market supply) and *MB* (market demand) at point E and social equilibrium given by interception of *MC* and *MSB* at point S:



7.5 Public goods

First of all, we have to be able to distinguish between private goods and other possible types of goods. In doing so it is useful to group all goods according to two characteristics associated with the following questions:

- Is a good excludable? Can people be prevented from using the good?
- Is a good rival? Does one person's use of the good diminish another person's enjoyment of it?

It appears to be that that there are two basic properties – excludability and rivalness – that define a good under question:

Excludability: The property of a good whereby a person can be prevented from using it

Rivalness: The property of a good whereby one person's use diminishes other people's use

Based on these two properties the following classification of different goods arises:

	Excludable	**Non-excludable**
Rival	Private goods	Common property resources
Non-rival	Club goods	Public goods

Private good: A good or service that can be consumed by only one person at a time and only by those people who have bought it or own it

Public good: A good or service that can be consumed simultaneously by everyone and from which no one can be excluded

Common property resources: Resources owned and managed collectively by a community or society rather than by individuals

Sometimes common property resources are called open access resources.

Club goods: A type of good in economics, sometimes classified as a subtype of public goods that are excludable but non-rival, at least until reaching a point where congestion occurs

Goods and services discussed so far have been private goods/services according to the above definition. Examples of public goods are national defense, public TV, healthcare, street lights, highways, bridges and others. Examples of club goods include private golf courses, cinemas, cable television, access to copyrighted works, and the services provided by social or religious clubs to their members. Examples of common property resources include irrigation systems, fishing grounds, pastures, forests, water or the atmosphere.

It is important to realize that because users cannot be excluded from enjoying a public good, it is unlikely that they, acting in their own self-interests, will pay enough to cover the costs of providing the good. Since most people try to free ride on others, private profit-oriented firms do not find it in their interests to provide public goods.

Free rider: A person who enjoys the benefits of a good or service without paying for it

The only way to produce a public good is to have the government to pay a share or total costs through the levying of taxes. However, when the government finances a public good, it does not mean that it must provide it. For example, garbage collection: A municipal government finances the service which is provided by a private company. Another example is construction of a highway: Government can hire private company to do it or it can set a partnership with a private company. On the other hand, in the case of the police, government directly provides these services.

Society needs public goods. However, if private markets were allowed to provide and allocate these goods, we would end up with zero of the goods because in general the following relationship would be observed:

Total Willingness to Pay for the good < Total Cost of providing the good

Why is this true? It is very difficult to reveal the true willingness to pay for the good because of the "free riding" problem defined above. In general, if people think they will have to pay for a public good, they may not reveal their true willingness to pay. This problem can be resolved through voting, referendum, associations of users, etc. However, all of these are not market solutions and that is why public goods are market failures.

7.6 Asymmetric information

Asymmetric information arises when one party in a transaction has more or superior information compared to another. This often happens in transactions where the seller knows more than the buyer, although the reverse can happen as well. Potentially, this could be a harmful situation because one party can take advantage of the other party's lack of knowledge.

Asymmetric information usually leads to the following two problems:

1. *Adverse selection*: Behavior that takes advantage of asymmetric information before a transaction. For example, a person who is not in good health may be more inclined to purchase life insurance than someone who feels good.
2. *Moral Hazard*: Behavior that takes advantage of asymmetric information after a transaction. For example, if someone has fire insurance they may be more likely to commit arson to reap the benefits of the insurance.

Most commonly, asymmetric information is studied in the context of the *principal-agent problem*. The principal-agent problem is about the difficulties that arise under conditions of asymmetric information when a principal hires an agent. Such situation usually causes the problem of potential moral hazard and conflict of interest since the principal hires the agent to pursue the principal's interests. The principal-agent problem arises when a principal compensates an agent for performing certain acts that are useful to the principal and costly to the agent, and where there are elements of the performance that are costly to observe.

Asymmetric information in general and principle-agent problem in particular are very important in any bargaining process. They are especially important in any contract negotiations.

There are some methods to address information asymmetry. *Signaling* and *screening* are the most popular. Michael Spence (1973) originally proposed the idea of signaling. He proposed that in a situation with information asymmetry, it is possible for people to signal their type, thus believably transferring information to the other party and resolving the asymmetry. Joseph Stiglitz pioneered the theory of screening. In this way the under-informed party can induce the other party to reveal their information. They can provide a set of choices in such a way that the choice depends on the private information of the other party.

7.7 Government intervention

Even though the concept of invisible hand states that markets can automatically find an efficient price and quantity of a good or service under question, in real life we frequently observe intervention by the government. As we have already seen, in the case of natural monopolies, externalities and public goods government intervention is justified by economic theory because eventually it improves economic efficiency. However, in some cases this intervention can destroy the invisible hand and the associated efficiency. In this section, we discuss two types of government intervention:

- Taxation
- Price control

7.7.1 Taxation

Almost everything we buy is taxed. In general, tax revenue collected by the government is used to provide public goods and services. However, there are some important economic questions associated with taxation. First of all, how taxation affects economic efficiency, and second, who really pays the tax – consumers or producers?

When we work, we pay income taxes on our earnings. We and our employers pay some amount of tax to finance our pension plan. In many cases, employers pay taxes on workers they employ. Of course, we are interested in how these taxes are split between consumers and producers or between you and your employer. *Tax incidence* addresses this question.

Tax incidence: The division of the burden of a tax between the buyer and the seller

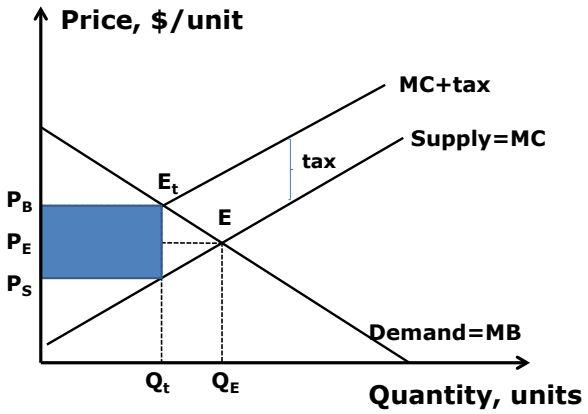
In general, when a tax is imposed, the following two prices arise:

- Price that includes the tax; buyers respond to this price and therefore, it is called the buyer's price P_B
- Price that excludes the tax; sellers respond to this price and therefore, it is called the seller's price P_S

Let us analyze the following, most popular three types of taxes:

1. Marginal tax on producers (sellers)
2. Income tax
3. Payroll tax

A *marginal tax on the sellers* of a good will shift the supply curve to the left until the vertical distance between the two supply curves is equal to the per unit tax. Other things being equal, this will increase the price paid by the consumers to P_B, and decrease the price received by the sellers to P_S:



Relationship between the two prices can be expressed by

$$P_B = P_S + tax$$

Market activity is reduced because the after-tax quantity Q_t is less than the before-tax quantity Q_E. The tax results in tax revenue to the government equal to $(tax \times Q_t)$ represented by the shaded blue area. Fraction of this tax revenue is paid by the consumers (the portion of the shaded rectangle above P_E line) while the other fraction is paid by producers (the portion of the shaded rectangle below P_E line).

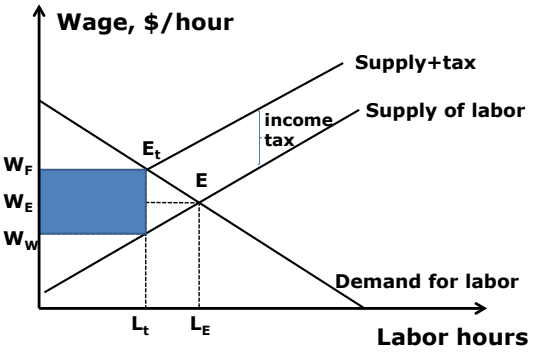
In general, the division of the burden of the taxation between the buyer and the seller depends on the price elasticities in the following way:

Rule of tax incidence: Tax burden falls more heavily on the side of the market that is less elastic

An excise or excise tax is an example of the mechanism we just discussed. This tax is commonly referred to as an inland tax on the sale, or production for sale, of specific goods; or, more narrowly, as a tax on a good produced for sale, or sold, within a country. Excise taxes are distinguished from customs duties, which are taxes on imports. If an excise tax is imposed on production it works exactly as discussed above. For example, in Canada both the federal and provincial governments impose heavy excise taxes on inelastic goods such as cigarettes, gasoline, alcohol, and for vehicle air conditioners. In economic sense, this tax is imposed per unit of a good produced which is subject to the framework depicted by the above diagram.

It appears to be that basic principles that we have just learned apply to all types of taxes. Let us analyze the other two types of taxes – income tax and the payroll tax.

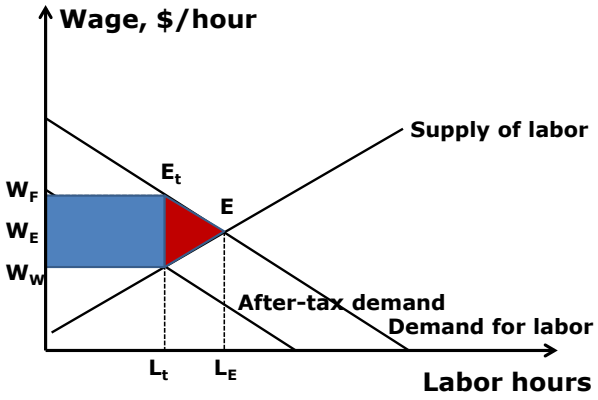
Suppose the government imposes a flat 20% income tax. It means that for every $1 of wage earned, people must pay 20 cents to the government. Because of the 20% income tax, the supply of labor shifts to the left as shown below:



As a result, workers (suppliers of labor) receive wage W_W and firms (consumers of labor) pay W_F. Compared to the before tax wage rate of W_E, the workers' wage decreases and the wage paid by the firms increases. As before, the shaded area shows total income tax, collected by the government. Again, despite the income tax was imposed on workers, both economic agents – workers and firms – suffer. We can conclude that technically this tax works exactly as the tax on producers discussed above.

Payroll tax: A tax on employers based on the wages they pay their workers

Since the payroll tax is imposed on employers, it affects demand (not supply as in the case of income tax) for labor as shown below:



In this case, the demand for labor shifts downwards (extra costs to producers decrease the quantity of labor demanded by firms) but the result is the same: Both economic agents suffer from the tax. Again the burden of the tax falls more heavily on more inelastic side of the market. In goods and services market, sales tax is an example of the tax that shifts demand downwards. However, technically it works the same way we just discussed.

As it was mentioned at the very beginning, taxation destroys the invisible hand. Tax imposes a wedge between marginal benefits of buyers and marginal costs of sellers. Compared to the before-tax situation, in general per unit tax results in the deadweight loss represented by the red area in the above diagram. In the case of taxation, this deadweight loss is associated with the so-called *excess burden of taxation*. The excess burden of taxation, also known as the distortionary cost or deadweight loss of taxation, is one of the economic losses that society suffers as the result of a tax.

From efficiency viewpoint, taxes discussed in this section, affect it negatively. On the other hand, since the government uses the tax revenue to provide public goods and services that people value, the excess burden might be worth bearing to obtain the benefits of government-provided public goods and services such as for example, environmental protection, education, medical care, etc.

7.7.2 Price control

Price control is usually realized in two forms:

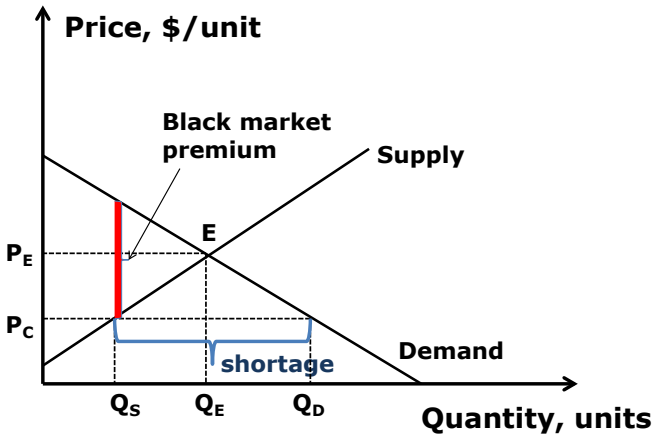
- Price ceiling
- Price floor

Let us analyze the price ceiling first.

Price ceiling: The legal maximum on the price at which a good or service is sold

It appears that the price ceiling results in a price, set below equilibrium level. It makes no sense to set the maximum price above equilibrium price because in such a case interaction between buyers and sellers will bring the price to its equilibrium value. Why? Remember, price ceiling does not allow price to be higher than the one, specified by the government; however, it does not say that it cannot be lower. Buyers and seller will choose the lower equilibrium price because this price satisfies both groups of economic agents.

On the other hand, if the government sets the maximum price below equilibrium, the equilibrium price becomes unattainable in principle. That is why in all textbooks the price ceiling is presented by the price set below equilibrium level. Let us analyze the situation graphically:

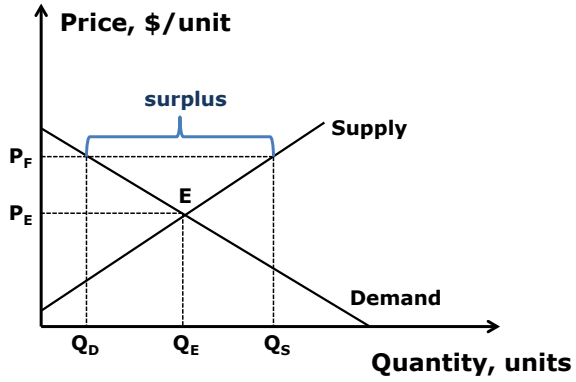


At price *Pc*, which is below equilibrium price P_E, the quantity demanded Q_D by consumers is greater than the quantity supplied Q_S by producers. It is not surprising since consumers want to buy more at lower price while producers sell less at such a price. The result is excess demand in the amount of $Q_D - Q_S$, which is known as shortage.

That is exactly what happened in the former Soviet Union with respect to consumer goods. Authorities (central planners) used to fix prices of almost all consumer goods below their equilibrium values, and as a result, consumers faced shortages of these goods. Consumers were shopping elsewhere to buy what they wanted, which caused development of the so-called secondary economy or "black market". In market economies, similar situation arises with regard to the fixed prices of tickets on sport events, rock concerts, and others, when demand exceeds supply. Another example of the price ceiling is rent control in housing market. As a result of this price setting the so-called black market premium arises which stimulate illegal actions by sellers and buyers (see the premium in the above diagram).

Price floor: The legal minimum on the price at which a good or service can be sold

By the same reason as in the case of the price ceiling, price floor results in a price set above equilibrium level. This situation is depicted in the graph below:



At price P_F, which is set above the equilibrium price P_E, the quantity supplied Qs exceeds the quantity demanded Q_D. As a result, there is surplus in the amount of $Qs - Q_D$. The best illustration of the price floor is the Minimum Wage Law. The good under question here is labor, measured in labor hours or number of workers. Price of labor is wage per hour. As a result, there is surplus of labor at minimum wage which is in fact unemployment. Therefore, according to our framework, the Minimum Wage Law leads to higher unemployment. However, there are many pros and cons regarding this issue discussed in economic literature.

In general price control reduces economic efficiency since it results in either shortage or surplus. In the case of price ceiling, people spend money, time and expend effort in order to deal with the shortages. They waste a lot of time looking for a good while if the market works efficiently, consumers can find quickly the goods they are looking for. In the case of price floor, those who would be willing to sell a good at the lowest price are not always those who actually manage to sell it. Therefore, like price ceiling, a price floor generates inefficiency by wasting resources.

8 National Accounting, Unemplyment and Inflation

Key concepts discussed in this chapter: macroeconomics, Keynes, neoclassical synthesis, national accounting, national accounting principles, gross domestic product (GDP), final goods, intermediate goods, expenditure approach, income approach, consumption, investment, government expenditure, exports, imports, net exports, aggregate expenditures, Gross National Product, nominal GDP, real GDP, GDP deflator, Labor Force Survey, working age population, labor force, employed unemployed, unemployment rate, labor force participation rate, discouraged worker, full-time workers, part-time workers, average labor hours, job losers, job leavers, entrants/reentrants, frictional unemployment, structural unemployment, seasonal unemployment, cyclical unemployment, business cycle, natural rate of unemployment, potential GDP, general price level, inflation rate, consumption basket, Consumer Price Index, real variables, nominal variables, Fisher's equation

8.1 Macroeconomics and national accounting

While microeconomics studies individual behavior of various economic agents, macroeconomic analysis is associated with the overall performance of an economy as a system:

Macroeconomics: The study of the determination of economic aggregates such as total output, total employment, the price level, and the rate of economic growth

The key in the above definition is the word *aggregate* as opposite to the *individual* behavior of economic agents studied by microeconomics. According to this definition, macroeconomics studies three big issues:

1. The standard of living
2. Unemployment
3. The cost of living

Economic growth is also a focus of macroeconomics but we will give it separate consideration later. As a matter of fact, modern macroeconomic analysis requires understanding of the short-run fluctuations along with the long-run economic growth. We will address short-run issues first.

Macroeconomics as a branch of economic theory in its modern form appeared in 1930s due to John Maynard Keynes's "General Theory of Employment, Interest and Money" published in 1936. That is what we call now Keynesian approach to macroeconomics. The generation following Keynes combined the macroeconomics of the "General Theory…" with neoclassical microeconomics to create the *neoclassical synthesis*. By the 1950s, most economists had accepted the synthesis view of the economy as a macroeconomic system. Economists like Paul Samuelson, Franco Modigliani, James Tobin, and Robert Solow developed formal Keynesian models, and contributed formal theories of consumption, investment, and money demand that fleshed out the Keynesian framework. Later other important schools of economic thought such as monetarists, new classicals, new Keynesians and some others have appeared. References to these schools are given in corresponding chapters.

In macroeconomics, activity of any economy is analyzed in terms of national accounts. National accounts are the implementation of complete and consistent accounting techniques for measuring the economic activity of a nation. Fundamental economic concepts are derived from national accounting tables.

National accounting has developed in tandem with macroeconomics from the 1930s with its relation of aggregate demand to total output through interaction of such broad expenditure categories as consumption and investment. Economic data from national accounts are also used for empirical analysis of economic growth and development.

A central concept in macroeconomics is the standard of living. It is associated with the value of total production in an economy rather than production of each individual good or service. This value is called *Gross Domestic Product GDP)*. There are some basic accounting principles incorporated in defining this concept:

1. The value of total output equals the value of total income (Fundamental national accounting identity); it also means that when something is produced in economy some income is earned.
2. What is produced in economy is consumed

8.2 Total production, national income and aggregate expenditures

Formal definition of GDP is:

Gross Domestic Product: The market value of all the final goods and services produced within a country in a given period of time

So, GDP reflects the market value of total production in an economy. The above definition can be presented in mathematical way as follows

$$Y_t = \sum_{i=1}^{N} P_{it} Q_{it}$$

where Y_t is gross domestic product in year t, Q_{it} is the quantity of the i-th good or service produced in the economy in year t, P_{it} is the price of the i-th good or service in year t, and N is the number of goods and services produced in the economy in year t.

In this definition, only the value of final goods and services should be taken into account:

Final good or service: A good or service that is produced for its final user and not as a component of another good or service

Therefore, it is necessary to distinguish between final and intermediate goods and services:

Intermediate good or service: A good or service that is produced by one firm, bought by another firm, and used as a component of a final good or service

In order to avoid doubling counting, value-added method can be used to evaluate GDP. According to this method:

$$Y_t = \sum_{i=1}^{M} VA_{it}$$

where Y_t is GDP in year t, VA_t is value added by producer i in period t, M is the number of producers.

Next, only goods and services produced within a country should be counted as a part of that country's GDP. Finally, the phrase *in a given period* of time usually means one year.

Previously formulated accounting principles define two methods of calculating GDP:

1. the expenditure approach
2. the income approach

Under the *expenditure approach*, traditionally economists identify four aggregate economic groups associated with consumption of the final goods and services:

- Households
- Firms (businesses)
- Government
- The rest of the world

Consequently, there are four types of expenditures associated with these groups:

- Consumption expenditure, *C*
- Investment spending, *I*
- Government expenditure on goods and services, *G*
- Net exports of goods and services, *NX*

Consumption expenditure: The expenditure by households on consumption goods and services

Investment spending: The purchase of new capital goods (tools, instruments, machines, buildings and other construction) by firms and additions to the firms' inventories

Government expenditure on goods and services: The expenditure by all levels of government on goods and services and investment in equipment and structures

Net exports of goods and services: The value of exports goods and services minus the value of imports goods and services

Exports: Items produced domestically and sold to the rest of the world (abroad)

Imports: Items produced abroad and sold domestically

Hence

$$\textit{Total (Aggregate) Expenditures} = C + I + G + NX$$

The described framework is associated with the expenditure method

Under the income approach, national income is the total income earned by citizens and businesses of a country. It consists of

- Wages earned by workers
- Interest earned by investors
- Rents earned by landlords
- Profits earned by entrepreneurs

Income is based on what factors (inputs) of production earn. From a microeconomic viewpoint, there are four basic factors of production: labor, capital, land and entrepreneurship. Labor earns wages, capital earns interest, land earns rent and entrepreneurship earns profits.

At large, from a theoretical viewpoint the following is true:

Total Value of Production = National Income = Total (Aggregate) Expenditures

It means that from a theoretical standpoint GDP reflects:

- Market value of total output
- National income
- Aggregate expenditures

National accounting also defines Gross National Product (GNP) as follows:

Gross National Product: The total value of all final goods and services produced within a nation in a particular year, plus income earned by its citizens (including income of those located abroad), minus income of non-residents located in that country

Basically, GNP measures the value of goods and services that the country's citizens produced regardless of their location.

8.3 Nominal GDP versus real GDP

Given definition of GDP as the market value of all final goods and services produced in an economy, an increase in GDP can be due to:

- An increase in prices of goods and services
- An increase in quantities of goods and services produced

If there is an increase in prices, our standards of living do not increase. That is why it is necessary to divide an increase in GDP in two parts:

- Change in production
- Change in prices

An increase in production is measured by real GDP:

Real GDP: The value of the final goods and services produced in a given year when valued at constant prices

In turn, an increase in prices is measured by nominal GDP:

Nominal GDP: The value of the final goods and services produced in a given year, valued at prices that prevailed in that year

To calculate real GDP, the so-called base year is chosen. Prices of the final goods and services in that year are used as constant prices to evaluate real GDP:

$$Y_R = \sum_{i=1}^{N} Q_{it} P_{ib}$$

where Y_R is real GDP and P_{ib} is the price of the i-th good or service in base year. Nominal GDP in year t is calculated according to formula

$$Y_N = \sum_{i=1}^{N} P_{it} Q_{it}$$

Since nominal GDP reflects price change while real GDP is evaluated at constant prices, the ratio of the nominal GDP to real GDP can provide us with the measure of the general price level. Such a measure is called *the GDP deflator*:

$$P = \frac{Y_N}{Y_R} \times 100\%$$

where P is GDP Deflator, Y_N is nominal GDP and Y_R is real GDP. Formal definition of the GDP deflator is:

GDP deflator: An average of current prices expressed as a percentage of base-year prices

Therefore, GDP deflator is a measure of the general price level in an economy and is widely used by economists in their empirical work.

8.4 Employment

Labor Force Surveys are statistical surveys conducted in a number of countries designed to capture data about the labor market. Some countries conduct them annually some monthly. For example, every month Statistics Canada conducts the Labor Force Survey (LFS) to analyze the structure of labor force in Canada.

According to the LFS, usually total population in a country is divided into different categories. First, total population is divided in the following groups:

- Working age population
- Others

Working age population: The total population in a region, within a set range of ages, that is considered to be able and likely to work

Second, the working age population is divided into:

- Labor force
- Not in labor force

Labor force: The number of people employed plus the number unemployed

Finally, the labor force is divided into:

- Employed
- Unemployed

Employed: Persons above a specified age who define the supply of labor for the production of goods and services

When measured for a short reference period (of one week or one day), it refers to all persons who worked for pay, profit or family gain during that period. It also includes all persons who had a job or enterprise but were absent from that job or enterprise during that period on a temporary basis: persons who during the reference period were sick, on vacation, maternity leave, strike or were temporarily laid off.

Unemployed: Persons above a specified age who are available to, but did not, compose the supply of labor for the production of goods and services

When measured for a short reference period, it relates to all persons not in employment who would have accepted a suitable job or started an enterprise during the reference period if the opportunity arose, and who had actively looked for ways to obtain a job or start an enterprise in the near past.

Those who do not fall into one of the above two groups are counted as not in the labor force.

The following are major labor market indicators:

- Unemployment rate
- Labor force participation rate

Unemployment rate: The percentage of the people in the labor force who are unemployed

If we define the unemployment rate as u, E as the number of employed, U as the number of unemployed and LF is the size of the labor force, then:

$$u = \frac{U}{LF} \times 100\%$$

Labor force participation rate: The percentage of the working-age population who are members of the labor force:

If we define working age population as WAP, then

$$Labor\ Force\ Participation\ Rate = \frac{LF}{WAP} \times 100\%$$

Neither unemployment rate nor labor force participation rate includes an important group of unemployed people called discouraged workers.

Discourage worker: A person who does not have a job, is available and willing to work, but has not made specific efforts to find a job within the previous four weeks

As well, in many countries the LFS distinguishes between full-time workers and part-time workers:

Full-time workers: People who usually work 30 hours or more a week

Part-time workers: People who usually work less than 30 hours per week

In turn, part-time workers are divided in two groups:

- Voluntary part-time workers
- Involuntary part-time workers

Involuntary part-time workers: People who work part time but want a full-time job

In empirical work, in order to determine the total amount of labor employed, economists usually measure labor in hours rather than in jobs:

Average hours: The total number of hours worked by all the people employed, both full-time and part-time, during a year

Job creation and job destruction and the movement into and out of the labor force create unemployment. People who become unemployed are:

- Job losers
- Job leavers
- Entrants or re-entrants

People who were laid off are called job losers. People who voluntarily left their jobs are called job leavers. People who just left high school and entered the job market are called entrants. People who previously had jobs but quit and left the labor force and who have now decided to look for a job are called re-entrants.

From a standpoint of economic theory, there are four types of unemployment:

- Frictional unemployment
- Structural unemployment
- Seasonal unemployment
- Cyclical unemployment

Frictional unemployment: The unemployment that arises from normal labor turnover – from people entering and leaving the labor force and from ongoing creation and destruction of jobs

The amount of frictional unemployment depends on the rate at which people enter and re-enter the labor force and on the rate at which jobs are created and destroyed. The amount of frictional unemployment is also influenced by unemployment benefits: The more generous these benefits, the longer is the average time taken in job search.

Structural unemployment: The unemployment that arises when changes in technology or international competition change the skill needed to perform jobs or change the location of jobs

Structural unemployment is usually due to mismatch between the structure of the supply of labor and the structure of the demand for labor. It can be the result of an underlying shift in the economy that makes it difficult for certain segments of the population to find jobs. It is typically when there is a mismatch between the jobs available and the skill levels of the unemployed. Structural unemployment usually lasts longer than frictional unemployment

Seasonal unemployment: Periodic unemployment in any field or industry that is dependent upon the weather

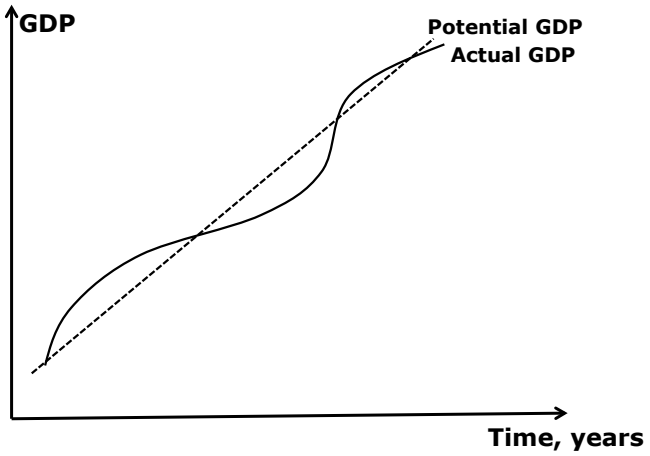
Usually seasonal unemployment increases in winter and decreases in summer time.

Cyclical unemployment: Fluctuation unemployment over the business cycle that increases during recessions and decreases during an expansion

Economists describe cyclical unemployment as the result of businesses not having enough demand for labor to employ all those who are looking for work. The lack of employer demand comes from a lack of spending and consumption in the overall economy. Cyclical unemployment is fluctuating unemployment over the business cycle:

Business cycle: The economy's short-run fluctuations in output and employment

It appears that real GDP follows a rather wavy pattern like the one presented by a solid line in the following diagram:



At a business cycle trough, cyclical unemployment is positive while at the business cycle peak it is negative. Hence, business cycles are associated with fluctuations of the real GDP.

A downswing (falling actual GDP) is referred to as recession while the upswing is referred to as expansion or boom when an economy is "overheated".

Since labor is one of the most important factors of production in an economy, one of the goals of economic policy is to achieve full employment. However, what is meant by full employment? Economists define full employment as follows:

Full employment: Situation when there is no cyclical unemployment or when all the unemployment is frictional, structural and seasonal

The unemployment rate at the full employment level of the economy is called *the natural rate of unemployment*. There is one more important macroeconomic indicator – potential GDP – presented by a dashed line in the above diagram.

Potential GDP: The level of real GDP that the economy would produce if it were at full employment

Since the unemployment rate fluctuates around the natural unemployment rate, real GDP fluctuates around its potential GDP.

8.5 Inflation and the cost of living

Economists use the term inflation to describe a situation in which the economy's general (overall) price level is rising. It is usually expressed in terms of the inflation rate:

The inflation rate: The percentage change in the price level from the previous period

If we define general price level as P, the inflation rate as π and time period (year) as t, then the above definition mathematically can be written as

$$\pi = \frac{P_t - P_{t-1}}{P_{t-1}} \times 100\%$$

Inflation is a closely watched aspect of macroeconomic performance, and is a key variable guiding macroeconomic policy. Measuring inflation is a difficult problem for government statisticians. To do this, a number of goods that are representative of the economy are put together into what is referred to as a "consumption basket". The cost of this basket is then compared over time. This results in a price index, which is the cost of the market basket today as a percentage of the cost of that identical basket in the starting year. The most popular index to measure inflation is the Consumer Price Index or CPI:

Consumer Price Index (CPI): A measure of the overall cost of the goods and services bought by a typical consumer

The goods and services are organized according to a classification system. Every product has a unique place in this classification. Products are grouped with other items either because they have a common end-use or because they are considered substitutes for each other. These families of products are joined together at different levels in the classification system to form a hierarchy.

In order to calculate the CPI, the following five steps should be performed:

Step 1. Define the consumption basket. Quantities for each commodity in the basket are set according to different surveys of consumer behavior. These quantities (weights) are adjusted periodically approximately once in every 5 years. Let us define quantity of the i-th item in the basket as Q_i.

Step 2. Find the prices. Statistician then collect data on prices of all commodities in the consumption basket over time. Let us define the price of the i-th item in the basket in year t as P_{it}.

Step 3. Compute the basket's cost. The cost of the basket in year *t* is the sum of the product "quantity×price" for all goods and services in the basket:

$$\textit{cost of consumption basket in year } t = \sum_{i=1}^{M} Q_i P_t$$

where *M* is the number of goods and services in the consumption basket.

Step 4. Choose the base year and compute the CPI. Let P_{ib} be the price of the *i*-th item in the base year. Then the CPI in year t is:

$$CPI_t = \frac{\sum_{i=1}^{M} Q_i P_{it}}{\sum_{i=1}^{M} Q_i P_{ib}} \times 100\%$$

Step 5. Compute the inflation rate according to its definition.

$$\pi = \frac{(CPI_t - CPI_{t-1})}{CPI_{t-1}} \times 100\%$$

where CPI_{t-1} is the CPI in the previous period. For example, if CPI was 105 in the previous year and is 108 now, then the current inflation rate is:

$$\pi = \frac{(108 - 105)}{105} \times 100\% = 2.86\%$$

There are four basic problems with the CPI as a measure of inflation:

- New goods bias
- Quality change bias
- Commodity substitution bias
- Outlet substitution bias

New goods bias: New goods keep replacing old ones but the CPI method uses a fixed basket to compute the price index. There is no sure way of making the necessary adjustment and most likely the arrival of new goods puts an upward bias onto the CPI and its measure of the inflation rate.

Quality change bias: With respect to quality, is the improvement in quality greater than the increase in cost? The CPI counts too much of any price rise as inflation than as improvement in quality and so overstates inflation.

Commodity substitution bias: Changes in relative prices lead consumers to substitute more expensive items in their consumption basket for less expensive. However, the CPI reflects cost of a fixed basket (unchanged quantities of goods and services).

Outlet substitution bias: When confronted with higher prices, people use discount stores more frequently and convenience stores less frequently. The growth of online shopping in recent years has provided an alternative to discount stores which makes outlet substitution even easier and potentially makes this source of bias more serious.

The main result of this discussion is: CPI overstates the inflation rate because of the above specified four biases. These biases in the CPI have two main consequences:

- They distort private contracts
- They increase government outlays

First, private contracts are distorted because usually wage contracts are linked to the CPI. Second, the CPI is used to adjust the incomes of people who receive income assistance from the government. Since the CPI overstates the inflation rate, both private contacts and government transfers are also over-valued.

That is why economists in their empirical research instead of CPI use another measure of the general price level discussed previously – the GDP deflator. The inflation rate in such a case is calculated as

$$\pi = \frac{(P_t - P_{t-1})}{P_{t-1}} \times 100\%$$

where P stands for the GDP deflator. Therefore, CPI and GDP deflator are two proxies for the general price level. The latter is preferred.

8.6 Nominal versus real values

The CPI provides a useful link between real and nominal values of macroeconomic variables. Nominal values are sometimes called values in actual dollars or dollars of current (not base!) year.

For example, to compare dollar amounts at different dates, we need to know the CPI at those dates. Suppose we want to compare prices in years t and n. The following relationship can be used:

$$P_{t,n} = P_n \times \frac{CPI_t}{CPI_n}$$

where $P_{t,n}$ is the price in year n in dollars of year t, P_n is the price in year n and CPI_t, CPI_n are the CPIs in year t and n respectively.

If the inflation rate is known and it is constant over some period of time, then the amount of dollars in some year t or P_t can be expressed in today's dollars as follows:

$$P_0 = \frac{P_t}{(1+\pi)^t}$$

where P_0 is amount in today's dollars and π is annual inflation rate

In fact, macroeconomics makes a big issue between nominal and real values. For example, in order to calculate real GDP economists use the GDP deflator to obtain real values of gross domestic product rather than the CPI:

$$Y_{R,t} = \frac{Y_{N,t}}{P_t}$$

where $Y_{R,t}$ is real GDP in year t, $Y_{N,t}$ is nominal GDP in year t and P_t is the GDP deflator in year t.

The other two major real-nominal relationships are associated with the following macroeconomic variables:

- Nominal wage versus real wage
- Nominal interest rate versus real interest rate

Nominal wage: The average hourly wage rate measured in current dollars

Real wage: The average hourly wage rate measured in dollars of a given base year

The relationship between the two is:

$$W_{R,t} = \frac{W_{N,t}}{CPI_t} \times 100$$

where $W_{R,t}$ is real wage in year *t*, $W_{N,t}$ is nominal wage in year *t* and CPI_t is the consumer price index in year *t*.

The real wage measures the change in the quantity of goods and services that an hourly work can buy or the so-called purchasing power of the wage. It measures the real reward for labor.

Nominal interest rate: The interest rate paid on a loan expressed as a percentage of the loan

Real interest rate: The interest rate payable on a loan expressed in the purchasing power of the interest received – the nominal interest rate adjusted for inflation

The approximate relationship between the two, which is known as Fisher's equation, is:

$$r = i - \pi$$

where r is real interest rate, i is nominal interest rate and π is the inflation rate. Exact relationship between the two is:

$$r = \frac{1+i}{1+\pi} - 1$$

9 Real Economy

Key concepts discussed in this chapter: real variables, nominal variables, classical dichotomy, aggregate demand/aggregate supply model, macroeconomic equilibrium, inflationary gap, recessionary gap, demand side shock, supply side shock, self-correction mechanism

9.1 Real economy and classical dichotomy

Economy as a system consists of various elements. By tradition, in macroeconomics the following basic elements are usually identified:

- Real sector or real economy
- Financial sector or money market
- Labor market
- Foreign exchange market

Real economy is the part of the economy that is concerned with actually producing goods and services, as opposed to the part of the economy that is concerned with buying and selling on the financial markets. In this chapter we are going to introduce concepts and models associated with the real sector of macroeconomic system.

However, first in order to study the real economy we have to understand the so-called classical dichotomy. As already discussed in chapter 8, macroeconomists have made progress in understanding how the economy works by dividing fundamental macroeconomic variables that describe macroeconomic performance into two lists:

- Real variables
- Nominal variables

Real variables: Items such as real GDP, real wage rate, real interest rate, and other real items such as levels of employment and unemployment

These variables describe the real economy and tell us what is really happening to production and consumption, saving and investment, work and leisure, all of which contribute to the standard of living.

Nominal variables: Items such as the price level (CPI or GDP deflator) and the inflation rate along with nominal GDP, the nominal wage rate, and the nominal interest rate

These variables describe the nominal economy and tell us how dollar values and cost of living are changing.

The separation of macroeconomic performance into a real part and a nominal part is the basis of a huge discovery called *the classical dichotomy*, which states:

When the economy is operating at full employment, the forces that determine the real variables are independent of those that determine the nominal variables

The classical dichotomy describes the economy at full employment. However, it does not hold over the business cycle as the economy fluctuates around full employment.

Business Cycle: Short-run fluctuations of real GDP around potential GDP

Potential GDP, introduced in chapter 8, is discussed in the next section in more detail. Therefore, it turns out that the forces that shape the real economy and those that shape the nominal economy interact to create the business cycle. Therefore, we model real economy in order to understand sources and consequences of the short-run fluctuations of GDP. In economics it is done with the help from the "Aggregate Demand/Aggregate Supply" (AD-AS) model.

As its name suggests, the AS-AD model applies to the economy as a whole – the aggregate or total economy – the ideas of supply and demand that you studied earlier.

In chapter 8 we introduced and discussed the following expression:

$$Y = C + I + G + NX$$

where Y is national income (total output), C is consumption, I is investment, G is government purchases of goods and services and NX is net exports.

In a mathematical sense, it is rather identity than equation since identity implies that the left-hand side and the right-hand side must be equal under any values of the variables involved. It makes perfect sense in economics, because we now know that the above identity reflects two possible ways to measure GDP: (i) through expenditure (the right-hand side), and (ii) through income or total output (the left-hand side). According to economic theory, the two measures should be equal in equilibrium. In fact, in the light of the AS-AD model, the above identity reflects the macroeconomic equilibrium when aggregate supply equals aggregate demand.

In general, the AD-AS model explains short-run fluctuations in real GDP and the price level.

9.2 Aggregate Supply

Aggregate supply is the total supply of all goods and services in the economy. The aggregate supply (AS) curve is a graph that shows the relationship between the aggregate quantity of output supplied by all firms in an economy and the overall price level.

When firms make long-term commitments to pay agreed dollar amounts, they make bets about future prices they will be able to obtain for their products. If their bets are accurate and the prices they obtain turn out to be what they expected, each firm operates at its desired output rate. The economy as a whole operates at full employment, and real GDP equals potential GDP:

Potential GDP: The level of real GDP that the economy would produce if it were at full employment

If the prices that firms can obtain for their product fall below the levels they expected, they take defensive actions. For some of the firms, the fall in prices brings lay-offs and a decrease in production. For other firms, the fall in prices is more serious: It shuts them down and destroys jobs. As a result, real GDP decreases.

In contrary, if the prices the firms obtain for their products rise above the levels they expected, they will act to take advantage of the higher prices. Existing firms hire new workers and increase production. New firms are set up, they hire new workers and start to produce. As a result, real GDP increases.

We have described the actions of firms that generate aggregate supply:

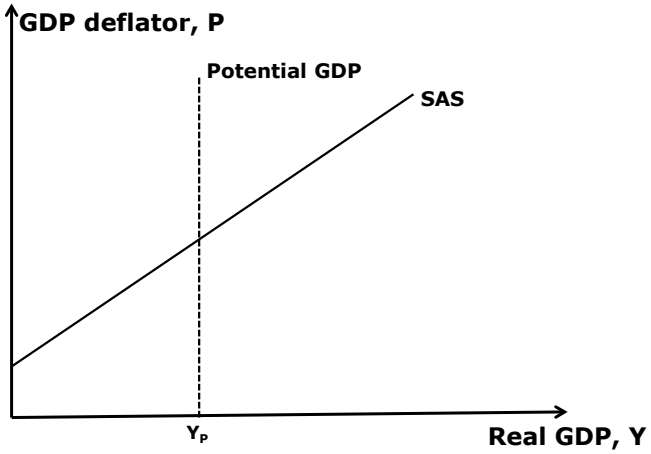
Aggregate Supply: Positive relationship between the price level and real GDP supplied on assumption that technology and all input prices are held constant

If in addition we can recall the definition of GDP as the market value of all final goods and services produced in an economy or:

$$Y = \sum_{i=1}^{N} P_i Q_i$$

where P_i is the price of the i-th good or service produced in the economy, Q_i is the quantity of the i-th good or service and N is the number of goods and services produced in the economy usually in a year. Again this analytical interpretation of aggregate supply shows a positive relationship between output Y (real GDP) and the price level P as noted before.

The following graph illustrates aggregate supply as the short-run aggregate supply (SAS) curve:



As the graph shows, aggregate supply SAS is an upward-sloping line in two-dimensional space "P – Y". As well, the vertical line on the graph illustrates potential GDP – a fixed level of real GDP, Y_p at full employment. In a physical sense, potential GDP at a point in time shows the economy's capacity at that time and it represents the state of an economy in the long-run.

The aggregate supply curve is not a market supply curve and it is not the simple sum of all the individual supply curves in the economy. One reason is that firms do not simply respond to market-determined prices, but they actually set prices. Price-setting firms do not have individual supply curves because these firms are choosing both output and price at the same time.

Another reason is that when we draw a firm's supply curve, we assume that input prices are constant. If the overall price level is rising, there will be an increase in at least some input prices. The outputs of some firms are the inputs of other firms. As wage rates and other input prices rise, the firms' individual supply curves are shifting, so we cannot sum them to get an aggregate supply curve.

Aggregate supply would change if potential GDP changes. If potential GDP increases, aggregate supply increases as well, and the AS-curve shifts rightward. Aggregate supply also changes when the money wage rate or any other money costs such as the price of oil (or other factors of production) changes. A rise in the money wage rate or the price of factors of production raises firms' production costs, decreases aggregate supply, and shifts the AS-curve leftward. Technology is another shift parameter of aggregate supply. In general, more advanced technology reduces cost of production and therefore shifts the AS-curve rightward – more will be produces at every price level. Therefore, in general factors that shift the SAS curve are:

- Potential GDP
- Input prices
- Technology

9.3 Aggregate Demand

In any economy, people have some money in banks, some debts, and some available funds to buy goods and services. All of these influence our spending plans. Firms also have funds available for spending on new capital goods which we call investment in physical capital. All of these funds and lines of credit are set in nominal, current dollar-value terms.

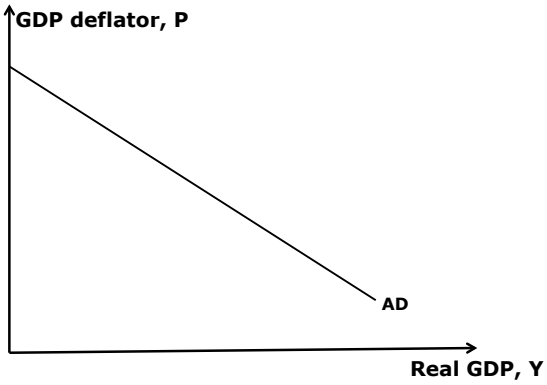
Suppose that all of a sudden, the prices that we pay for goods and services increase. With a fall in the purchasing power of money, people scale back their consumption plans, and firms scale back their investment plans. Government reduces its purchases of goods and services. Foreigners buy less of our production, so exports decrease as well.

In other times, the prices we pay for goods and services may decrease. With the following rise in the purchasing power of money, people increase consumption while businesses increase investment spending. Government increases its purchases of goods and services. Foreigners buy more of our products and exports increase.

We have just described a negative relationship between the price level and expenditures of different types – by households, firms, government and the rest of the world. This relationship is known as aggregate demand AD:

Aggregate Demand: Negative relationship between the quantity of real GDP and the price level when all other influences on expenditure plans remain the same

The following graph shows this negative (inverse) relationship in "P – Y" space:



As we can see, the AD-curve is down-sloping reflecting the inverse relationship between the price level *P* and real GDP, *Y*. The AD curve is not a market demand curve, and it is not the sum of all market demand curves in the economy. It is a more complex concept

It is possible to express aggregate demand in analytical form using our fundamental GDP identity. In that identity, the right-hand side reflects all types of expenditures incurred in the economy, and hence aggregate demand AD can be represented as

$$AD = C + I + G + NX$$

Many factors change aggregate demand and shift the AD-curve. First of all, any change in any of the four types of expenditures that appear on the right-hand side of the above expression would shift AD directly: an increase would shift demand upwards while a decrease would shift demand downwards.

In addition, fall in the interest rate, an increase in the quantity of money in the economy, a tax cut, and an increase in real GDP in the rest of the world would stimulate domestic spending and, as a result, would increase aggregate demand (upward shift of the AD-curve). Changes in these variables in the opposite direction would decrease aggregate demand (downward shift of the AD-curve).

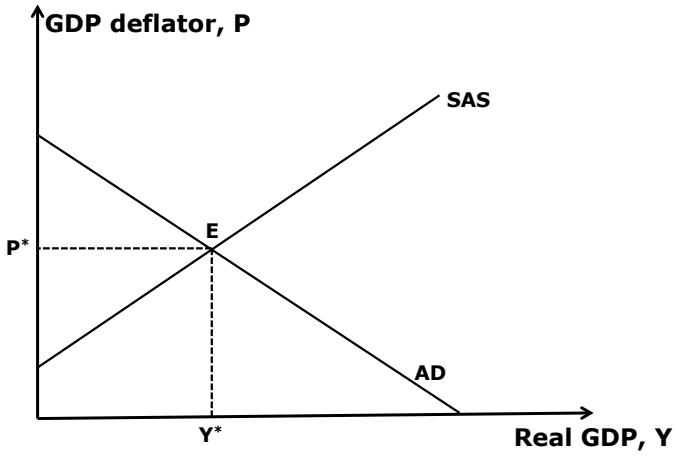
Therefore, in general factors that shift the AD curve are:

- C, I, G and NX
- Interest rate
- Quantity of money
- Taxes
- Foreign income

Also changes in the exchange rate which is discussed later affect AD.

9.4 Macroeconomic equilibrium

Interaction between aggregate demand AD and the short-run aggregate supply SAS determine real GDP, *Y* and the price level *P* in an economy in the short-run. Macroeconomic equilibrium occurs when the quantity of real GDP demanded equals the quantity of real GDP supplied at a point of intersection of the down-sloping AD-curve and the upward-sloping SAS-curve which is shown in the following graph.



Suppose that the price level *P* is above *P**. At this price level, the quantity of real GDP demanded is less than the quantity of real GDP supplied. Firms are unable to sell all their output and unwanted inventories pile up:

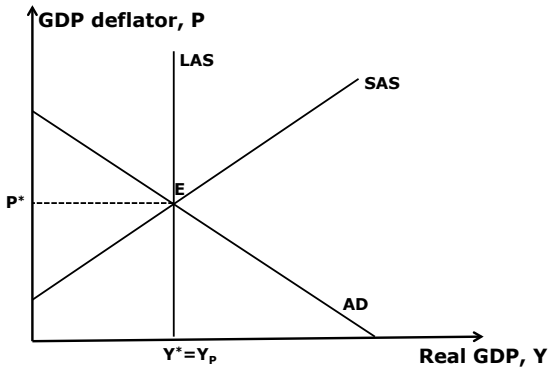
Inventories: A stock of finished but unsold goods

So, firms decrease production and prices until they can sell all their output, which occurs exactly at point E. Now suppose that the price level is below *P**. At this price, the quantity of real GDP demanded exceeds the quantity of real GDP supplied, and firms are unable to meet the demand for their output. Inventories decrease, and customers clamor for goods and services. As a result, firms increase production and raise prices until they can meet demand, which occurs exactly at point E.

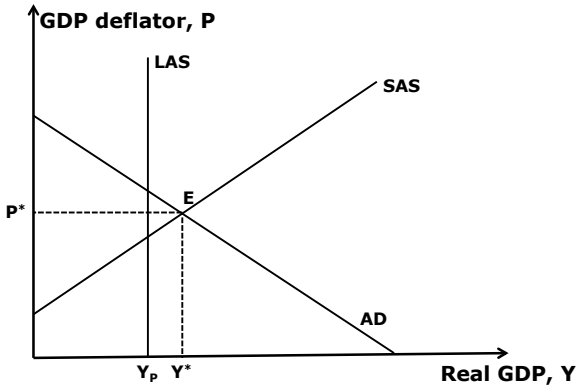
In macroeconomic equilibrium, the economy might be:

1. at full employment;
2. above full employment;
3. below full employment.

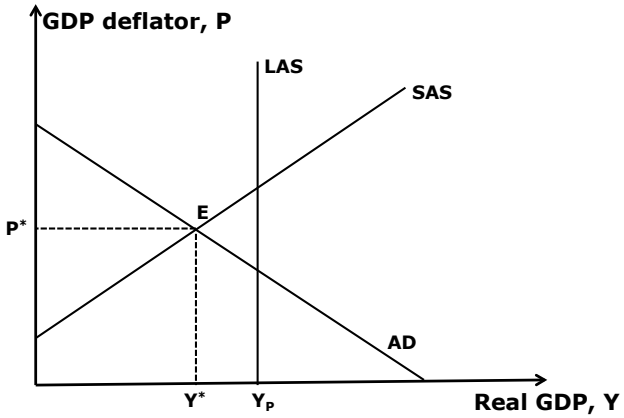
Full employment implies equilibrium real GDP is equal to potential GDP or $Y^* = Y_p$ as shown below:



When equilibrium real GDP exceeds potential GDP, the economy is above full-employment level, and it is said that there is an inflationary gap:



In turn, when potential GDP exceeds equilibrium level real GDP, the economy is below its full-employment level, and it is said that there is a deflationary or recessionary gap:



In this regard, it is vital to understand the forces that determine potential GDP. We produce the goods and services that make up real GDP by using factors of production as follows:

- labor and human capital;
- physical capital (equipment, machinery, tools, buildings, infrastructure, etc)
- land and natural resources;
- entrepreneurship.

At any given point in time, only quantity of labor is variable while all other factors of production are fixed. Hence, real GDP depends on the quantity of labor employed.

On the one hand, quantity of labor supplied depends on the choices that people make about the allocation of time between work and leisure. In turn, the quantity of labor demanded depends on firms' decision about how much labor to hire. Equilibrium level of labor employed comes from the labor market when the quantity of labor supplied by workers equals the quantity of labor demanded by firms. This equilibrium level of labor employed determines potential GDP.

9.5 Macroeconomic shocks and business cycles

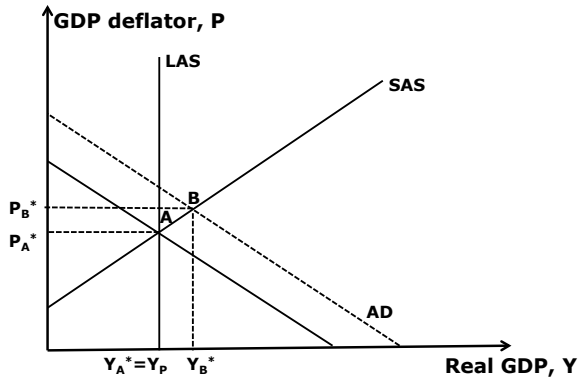
It is necessary to distinguish between demand side shocks and supply side shocks. First, let is discuss an example of the demand side shock. Suppose that an economy's trading partners abroad decided to eliminate their tariffs:

Tariff: A tax applied on imports of goods and services

As a result, the economy can now sell more of domestic goods and services to foreigners and consequently increase its exports. Net exports go up as well since:

$$NX = Exports - Imports$$

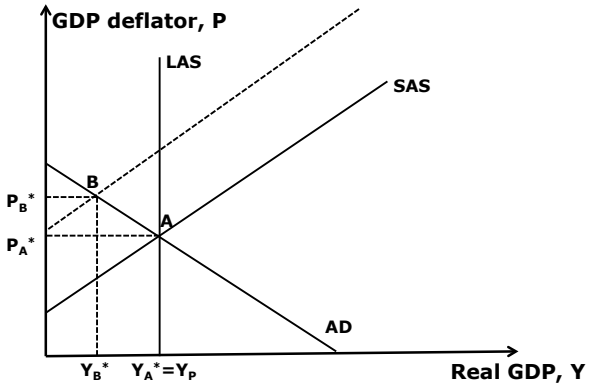
As discussed in the previous section, net exports *NX* is a component of aggregate demand, and hence AD-line shifts upwards:



The economy starts in a long-run equilibrium at point A. As a result of a positive demand side shock, caused by elimination of tariffs, it moves from point A to point B. At point B, $P_B > P_A$ which means inflation. On the other hand, the economy's output increases since $Y_B > Y_A$ which results in a decrease in unemployment. Therefore, because of the positive demand shock we observe a trade-off between inflation and unemployment: Unemployment goes down while inflation goes up.

In general, aggregate demand shocks cause price and real GDP to change in the same direction: Both rise with an increase in aggregate demand, and both fall with a decrease in aggregate demand.

Now suppose that the price of oil goes up. It is an example of the supply side shock. Oil price affects almost all input prices. As a result, costs of production increase, and aggregate supply shifts to the left since at the same price less will be produced:



The economy moves from point A to point B. At point B, $P_B > P_A$ which means inflation. However, output decreases since $Y_B < Y_A$. Hence, as a result of the negative supply shock we observe two negative effects – higher inflation and higher unemployment.

In general, aggregate supply shocks cause the price level and real GDP to change in opposite directions. With an increase in supply (rightward shift of SAS), the price level falls and GDP rises; with a decrease in supply (leftward shift of SAS), the price level rises and GDP falls.

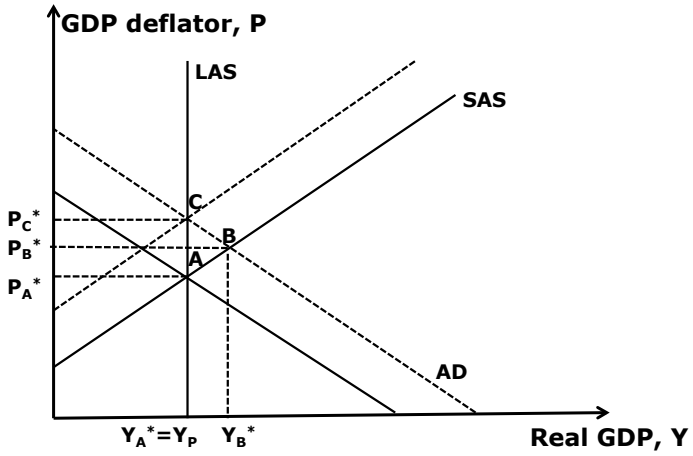
It appears to be that any economy is constantly hit by variety of different shocks, and that is why we observe fluctuations in real GDP (Y_B versus Y_A) all the time which we earlier defined as business cycles. Should the government intervene to correct the situation? We discuss the issue next.

9.6 Self-Correction Mechanism

In microeconomics, we introduced a concept of the "invisible hand". There exists an analog of the invisible hand in macroeconomics known as a *self-correction mechanism*. In principle, it can eliminate inflation and unemployment automatically.

Previously we have used the short run aggregate supply to trace the consequences of demand and supply shocks or existence of business cycles. We also assumed that in the long-run the economy should be at its potential GDP or at its full employment level. It means that the long-run aggregate supply LAS is a vertical line drawn at the level of potential GDP, Y_P .

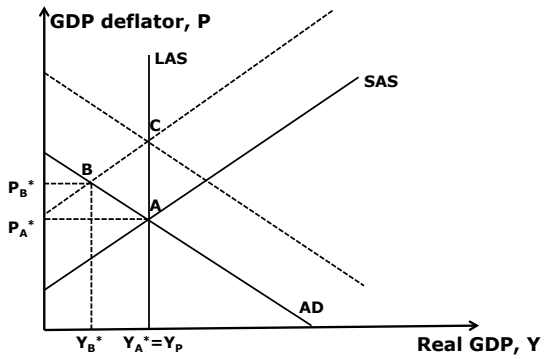
Now recall, when an economy produces more than potential GDP, we call that situation an inflationary gap. Economy can arrive at this position, for example, as a result of a positive demand side shock discussed in the previous section. This situation can be depicted as follows:



In the short-run the economy is at point B. At this position, factors of production are scarce, workers work overtime and equipment is over-utilized. Firms compete for the scarce factors of production pushing their prices up. Wages increase due to overtime work. Eventually increasing input prices increase costs of production in the economy, and the short-run upward-sloping aggregate supply SAS shifts to the left until it reaches the long-run vertical aggregate supply LAS at point C. As a result, the economy is at its full capacity Y_P (potential GDP), however, inflation is observed since $P_C > P_B$.

Conclusion: Whenever the economy starts off at a point that is beyond potential GDP, the process of wage and price inflation pulls it back to the full employment level Y_P.

Let us now analyze the opposite case known as a deflationary gap, which may occur, for example, as a result of a negative supply side shock as discussed in the previous section.



At the short-run equilibrium at point B, there are many unemployed who eventually accept jobs at lower wages. The economy's costs of production decrease and the short-run aggregate supply SAS (dashed line in the above diagram) shifts to the right until the full capacity level Y_P is restored at original point A.

Conclusion: Whenever the economy starts off at a point that is below potential GDP, the process of wage and price deflation pulls it back to the full employment level Y_P.

Some economists believe that the described self-correction mechanism works rather slowly, especially in the case of deflationary gap because workers resist wage cuts. They suggest government intervention that pushes aggregate demand upwards to reach point C instead of point A. We will discuss this situation in more detail later. On the other hand, in many cases free market economy is able to offset different shocks on its own.

10 Money and Monetary System

Key concepts discussed in this chapter: commodity money, fiat money, bank money, liquidity, money supply, M1, M2, M3, money substitutes, monetary system, central bank, commercial bank, monetary policy, inflation targeting, balance sheet, assets, liabilities, monetary base, desired reserve ratio, deposit multiplier, currency drain, money multiplier

10.1 Money and its functions

Money is always something that can be recognizable and that can be divided in small parts. It might be an actual commodity (silver, gold), a token (a $10 bill) or a virtual token such as an electronic record in a bank's database. In general money is defined as follows:

Money: Any commodity or token that is generally accepted as a means of payment

Economists differentiate among three different types of money:

- Commodity money
- Fiat money
- Bank money.

Commodity money is a good whose value serves as the value of money. Gold coins are an example of commodity money. In most countries, commodity money has been replaced with fiat money. *Fiat money* is a good, the value of which is less than the value it represents as money. Dollar bills are an example of fiat money because their value as slips of printed paper is less than their value as money. *Bank money* consists of the book credit that banks extend to their depositors. Transactions made using checks drawn on deposits held at banks involve the use of bank money.

There are 3 basic functions of money:

- Medium of exchange: An object that is generally accepted in return for goods and services
- Unit of account: An agreed-upon measure for stating the prices of goods and services
- Store of value: Any commodity or token that can be held and exchanged later for goods and services

One important property of any financial asset including money is liquidity defined as follows:

Liquidity: The degree to which an asset or security can be bought or sold in the market without affecting the asset's price

Liquidity is characterized by a high level of trading activity. Assets that can be easily bought or sold are known as liquid assets. In a sense liquidity is the ability to convert an asset to cash quickly. It turns out that money is the most liquid asset amongst all financial assets. It is so because primary function of money is being the economy's medium of exchange.

In economics, money is a broad term that refers to any financial instrument that can fulfill the functions of money discussed above. These financial instruments together are collectively referred to as *the money supply* of an economy. The money supply is the amount of financial instruments within a specific economy available for purchasing goods or services. Since the money supply consists of various financial instruments such as

- currency
- demand deposits
- various other types of deposits,

the amount of money in an economy is measured by adding together these financial instruments creating a monetary aggregate. Accordingly there are several measures of the money supply. The most popular are:

- M1
- M2
- M3

M1: Currency outside the banks plus demand deposits (chequing accounts or chequable deposits)) owned by individuals and businesses at chartered banks

M2: M1 plus personal savings deposits and non-personal time deposits under $100,000 at chartered financial institutions

M3: M2 plus time deposits larger than $100,000 plus deposits at credit unions, caisses populaires, and other depository institutions plus money market mutual funds and deposits at other non-chartered financial institutions

Currency (coins and banknotes) is money by definition. Chequing deposits can be transferred from one person to another electronically or by writing a cheque, and hence they are also money. Since M1 consists of currency and chequing deposits, it is money.

Some of savings deposits in M2 and M3 are just as much a means of payment as the chequing deposits in M1. However, other savings deposits such as, for example, term deposits cannot be transferred in the same way and thus, they are not means of payment. Hence only part of M2 and M3 is money.

Things that serve as medium of exchange but are not a store of value are sometimes called *money substitutes*. Credit cards are a prime example. With a credit card, many transactions can be made without either cash or cheque. A credit card serves the short-run function of a medium of exchange but this is only temporary: Money remains the final medium of exchange for all transaction when the credit card bill is paid.

10.2 Monetary system: Central bank

Modern monetary system consists of two fundamental elements:

- Central bank
- Other financial (monetary) institutions

Central bank such as, for example, Federal Reserve in USA or European Central Bank in European Union is the issuer of currency to the monetary system. Other financial institutions, mostly what we call commercial banks, can create credit but they cannot create net new financial assets.

At large, other financial institutions can be divided into:

- Commercial or chartered banks: Private firms that are chartered by the government to receive deposits and make loans
- Credit unions: Cooperative organizations owned by social or economic groups that accept deposits from and make loans to their members
- Other depository institutions such as mutual funds, trust companies, mortgage companies, hedge funds, etc.

The central bank is usually described as "the lender of last resort", which means that it is responsible for providing its economy with funds when commercial banks cannot cover a supply shortage. In other words, the central bank prevents the country's banking system from failing. However, the primary goal of central banks is to provide their countries' currencies with price stability by controlling inflation. A central bank also acts as the regulatory authority of a country's monetary policy and is the sole provider and printer of notes and coins in circulation. Time has proved that the central bank can best function in these capacities by remaining independent from government fiscal policy and therefore uninfluenced by the political concerns of any regime. The central bank should also be completely divested of any commercial banking interests.

Therefore, there are four basic functions of a central bank:

1. Printing currency
2. Serving as a banker to commercial banks
3. Serving as a banker to the government
4. Control over money supply (the most important function!)

At present in any market economy the central bank is government owned but separate from the country's ministry of finance. Although the central bank is frequently termed the "government's bank" because it handles the buying and selling of government bonds and other monetary instruments, political decisions should not influence central bank operations. Of course, the nature of the relationship between the central bank and the ruling regime varies from country to country and continues to evolve with time. To ensure the stability of a country's currency, the central bank should be the regulator and authority in the monetary system.

From a theoretical standpoint, a central bank has two fundamental functions: (1) macroeconomic: regulating inflation and price stability, and (2) microeconomic: functioning as a lender of last resort. We are more interested in the macroeconomic function of a central bank.

Macroeconomic price stability implies that the central bank must regulate the level of inflation by controlling money supplies by means of *monetary policy* defined as follows:

Monetary policy: The process by which the monetary authority of a country controls the supply of money, often targeting a rate of interest for the purpose of promoting economic growth and stability

The central bank performs open market transactions that either inject the market with liquidity or absorb extra funds, directly affecting the level of inflation. To increase the amount of money in circulation and decrease the interest rate (cost) for borrowing, the central bank can buy government bonds, bills, or other government-issued notes. This buying can, however, also lead to higher inflation. When it needs to absorb money to reduce inflation, the central bank will sell government bonds on the open market, which increases the interest rate and discourages borrowing. Open market operations are the key means by which a central bank controls inflation, money supply, and price stability:

Open market operation: The purchase or sale of government securities – treasury bills and bonds – in the open market by the central bank

Currently in many market economies the so-called *inflation targeting* is used as the basic element of the central bank's monetary policy. The inflation targeting helps make the central bank's monetary policy actions more readily understandable to financial markets and the public. The target also provides a clear measure of the effectiveness of monetary policy. One of the most important benefits of a clear inflation target is its role in anchoring expectations of future inflation. This, in turn, leads to the kind of economic decision making – by individuals, businesses, and governments – that brings about non-inflationary growth in the economy.

In order to study how the monetary system works, we need to understand balance sheet of a central bank as well as balance sheets of commercial banks. In general, balance sheet includes assets and liabilities broadly defined as:

Assets: Anything a bank owns

Liabilities: Anything a bank owes

The following are basic assets of the central bank:

- Cash reserves in the bank's vault
- Gold
- Reserves of foreign currency
- Government securities: Bonds, notes, and other debt instruments sold by a government to finance its borrowings

Liabilities of the central bank include:

- Currency in circulation
- Deposits of commercial banks
- Deposits of government

The central bank's balance sheet can be presented as follows:

ASSETS	LIABILITIES
Cash reserves in the bank's vault	Currency in circulation
Gold	Deposits of commercial banks
Reserves of foreign currency	Deposits of government s
Government securities	

Monetary policy conducted by the central bank is aimed at the change in the quantity of money in the economy or what is called money supply. Technically the change comes through the monetary base:

Monetary base: The sum of deposits of commercial banks at the central bank plus coins and bank notes in circulation

Sometimes economists refer to monetary base as high-powered money because it includes only highly liquid financial assets.

10.3 Monetary system: Commercial banks

Commercial banks are the largest and most popular representatives of the other financial (monetary) institutions in the structure of an economy's monetary system. That is why we study this branch of the monetary system via commercial banks.

Commercial bank: A privately owned, profit-seeking institution that provides a variety of financial services

Financial services of a typical commercial bank are:

- To take deposits
- To make loans
- To operate a payments system

A typical commercial bank accepts three types of deposits:

1. Chequing deposits
2. Savings deposits
3. Time or term deposits

A bank pays a low interest rate on chequing deposits (sometimes zero) and higher interest rates on other deposits. The goal of commercial banks is to maximize its stock-holders' long-run wealth. That is why a commercial bank makes loans at a higher interest rate than it takes deposits. Businesses and governments can also make deposits at commercial banks. All these deposits are reported as liabilities on the commercial banks' balance sheets.

A typical commercial bank has the following assets:

- Cash assets:

1. Reserves of the currency in the bank's vault
2. Deposits at the central bank
3. Funds that can be converted into reserves quickly such as overnight loans, Treasure bills, short-term loans (highly liquid assets)

- Bonds: securities issued by the government or other large governmental organizations and private corporations
- Loans: the provision of funds to businesses and individuals

There are four basic economic functions of commercial banks:

- Create liquidity
- Lower the cost of lending and borrowing
- Pool risks
- Make payments

A bank creates liquidity by borrowing short and lending long. Borrowing short means accepting deposits and standing ready to repay them whenever the depositor requires the funds. Lending long means making loan commitments for a long term.

People who want to borrow funds can do so using the facilities offered by banks, which lowers the cost of borrowing.

However, lending is a risky activity. Some loans are never repaid. By offering loans to diverse groups of borrowers, banks diversify the risk involved or they pool risks.

Banks provide a payments system that allows transferring funds from one person to another at low cost. Electronic cheque-clearing system is the main mechanism.

In general, commercial banks are financial intermediaries. They are called intermediaries because they stand between savers, from whom they accept deposits, and investors, to whom they make loans.

10.4 How money is created by the monetary system

Suppose that a new $100 bill is given to you. How this bill appeared will be addressed later. As a smart person, you deposit the money at Bank # 1. In such a case, the Bank's balance sheet looks like this:

ASSETS	LIABILITIES
Reserves = $100.00	Deposits = $100.00
Loans = $0.00	

Note that total assets equal total liabilities. Bank # 1 is not satisfied with the outcome because it does not earn a penny from such a transaction. Therefore, the bank decides to keep just 10% of the amount as reserves and lend out the rest in a form of a loan to a customer. The 10% ratio is called the desired reserve ratio:

Desired reserve ratio: The ratio of reserves to deposits that banks wish to hold so that they are sure to have enough cash to cover withdrawals by their customers

In general, such monetary system is called fractional-reserve banking:

Fractional-reserve banking: A banking system in which banks hold only a fraction of deposits as reserves

The bank's new balance sheet is

ASSETS	LIABILITIES
Reserves = $10.00	Deposits = $100.00
Loans = $90.00	

The person, who obtains the loan, deposits the money at Bank # 2 with the same desired reserve ratio of 10%. It means that Bank # 2 lends out $81 out of $90, received as a deposit. The Bank's balance sheet is

ASSETS	LIABILITIES
Reserves = $9.00	Liabilities = $90.00
Loans = $81.00	

The person, who obtains the loan of $81, deposits the money at Bank # 3 with the same desired reserve ratio of 10% and so on.

Now let us determine the total quantity of deposits all the banks created as a result of the extra $100 introduced into the monetary system. The extra deposits created are:

$$\$100 + \$90 + \$81 + ...$$

or

$$\$100 + 0.9 \times \$100 + 0.9^2 \times \$100 + ... = \$100 \times (1 + 0.9 + 0.9^2 + ...)$$

The last sequence has a finite sum which produces

$$\frac{\$100}{(1-0.9)} = \$1{,}000$$

Therefore, introduction of a new $100 bill eventually generates $1,000 of deposits under the desired reserve ratio of 10%. The process is captured by the deposit multiplier which numerically is defined as follows

deposit multiplier = 1/desired reserve ratio

Deposit multiplier: The number by which an increase in bank reserves is multiplied to find the resulting increase in bank deposits

In the above example

deposit multiplier = 1/0.1 = 10

In our previous analysis of the money multiplier, we assumed a new $100 bill. Now we are about to learn how the bill appeared in the first place. In order to do so we need to understand how open market operations work.

In order to inject new money into the system, central bank has to buy government bonds (securities) from the public. Let us trace all transactions involved via two balance sheets:

1. As central bank buys government bonds from the public, its assets increase (an increase in the stock of government securities). Physically the bank writes a cheque to an individual who sells the bonds.
2. The individual deposits the cheque in his account at a commercial bank (deposits increase).
3. The commercial bank sends the cheque back to the central bank as its new deposit. At the same time, the commercial bank's reserves, which are now held at the central bank, increase as well.

So, as a result, the commercial bank reserves have gone one-for-one with its deposit obligations. This transaction also increased the monetary base by increasing the bank's reserves at the central bank. The increase in the monetary base equals the amount of the open market purchase.

Increase in deposits of the commercial bank implies that there is now excess reserves in the amount of

excess reserves = amount of open market purchase×(1 – desired reserve ratio)

The commercial bank lends out the excess in a form of a loan, and the following sequence of events takes place:

- As a result of a new loan, total deposits go up
- The quantity of money in the economy increases
- Some of the new money is used to make payments
- Some of the new money remains in deposits in banks
- Some of the new money is held as currency – currency drain
- The bank's desired reserves increase again
- New excess reserves is created: Smaller than before but still positive

This is just the first round of the multiplier process. Each consecutive round starts with a smaller amount of money and repeats all the steps described above. In the end, the original increase in monetary base (equal to the size of the open market purchase) causes cumulative effect which results in the quantity of new money created:

$$\textit{quantity of new money created} = m \times \textit{open market purchase}$$

where m is the money multiplier. Let C be the fraction of money held in the form of currency (currency drain) and R be the desired reserve ratio. Then the money multiplier can be defined as follows:

$$m = \frac{1}{1-(1-C)(1-R)}$$

For example, if as a result of open market operation, the central bank bought $100,000 of government bonds, and the currency drain is 33% while the desired reserve ratio is 10%, then:

$$m = \frac{1}{1-(1-C)(1-R)} = \frac{1}{1-(1-0.33)(1-0.1)} = 2.52$$

Quantity of new money created is

$$2.52 \times \$100{,}000 = \$252{,}000$$

In general, it is possible to introduce the following relationship:

$$\Delta M^{s} = m \times \Delta MB$$

where ΔM^{S} is an increase in money supply, m is the money multiplier defined above, ΔMB is the change in monetary base in the first round. The latter is exactly equal to the value of an open market operation.

Another way to increase money supply in the economy through open market operations is for the central bank to purchase foreign currency. In such a case again, the central bank injects extra money into the system. It works the same way as previously discussed. Usually the central bank buys or sells foreign exchange in order to affect the international value of local currency. Hence it appears to be that exchange policy of this sort and monetary policy is the same thing.

11 Fundamentals of Fiscal and Monetary Policy

Key concepts discussed in this chapter: fiscal policy, budget, budget deficit, national debt, discretionary fiscal policy, automatic fiscal policy, government expenditure multiplier, tax multiplier, balanced budget multiplier, marginal propensity to consume, marginal propensity to save, expansionary fiscal policy, contractionary fiscal policy, automatic stabilizers, induced taxes, induced transfer payments, money market, interest rate, demand for money, contractionary monetary policy, expansionary monetary policy

11.1 Fundamentals of fiscal policy

In defining fiscal policy, it is useful to discuss the broader idea of government economic policy. Previously we discussed the role of the government in terms of delivering public goods and services such as law and order, national defense, social programs, environmental protection and others. This is mostly a microeconomic component of public policy. Another component of government activity is associated with macroeconomic policies: Government issues currency, enforces rules concerning the production and sale of goods and services, and even operates businesses themselves (for example, crown corporations in Canada or state owned enterprises in Europe).

Fiscal policy represents a particular area of macroeconomic policy. More specifically, it involves the use of government finances to influence the overall behavior of a national economy. Of particular importance to fiscal policy is a government's budget, or annual levels of spending, taxation, and borrowing. In general, fiscal policy can be defined as

Fiscal policy: The use of government revenue collection (taxation) and expenditure (spending) to influence the economy

Fiscal policy is based on the concept of government budget:

Government budget: Annual statement of the expenditures, tax receipts, and surplus or deficit of the government

In this context, the budget functions not just as a process by which the government collects revenues to pay for goods and services for its citizens, but as a mechanism for promoting economic stability over the short- and long-terms.

With respect to budget, government can run either budget surplus or budget deficit. Government's surplus or deficit is equal to its tax receipts minus its expenditures or

Budget surplus (+) or deficit (-) = Tax receipts – Expenditures

If receipts are exactly equal to the expenditures, it is said that the budget is balanced. So, if we define *G* as government spending and *T* as tax revenue, budget *B* in any year *t* can be define mathematically as

$$B_t = T_t - G_t$$

When the government has budget deficit over a period of time, it accumulates debt and the accumulated debt is referred to as national debt:

National debt: The total amount of debt outstanding that arises from past budget deficits

If we define budget deficit as *BD* then in any year *t* it is

$$BD_t = G_t - T_t$$

and, according to the above definition, national debt *ND* in any year *t* is

$$ND_t = \sum_{i=0}^{\infty} BD_{t-i}$$

Traditionally fiscal policy has been seen as an instrument of demand management. This means that changes in spending and taxation can be used "counter-cyclically" to help smooth out some of the volatility of real national output (real GDP) particularly when the economy has experienced an external shock. Therefore, the main objective of a viable fiscal policy is to manage the level of aggregate demand through the change in taxes and government expenditures. In general, fiscal policy can be either:

- Discretionary

or

- Automatic

Discretionary fiscal policy: A fiscal policy action that is initiated by an act of Parliament

Automatic fiscal policy: A fiscal policy action that is triggered by the state of the economy, such as an increase in payments to the unemployed and a decrease in tax receipts triggered by recession

When we talk about fiscal policy, most of the time we mean discretionary fiscal policy and therefore, let us start our discussion with this policy first.

11.2 Discretionary fiscal policy and the multiplier process

Discretionary fiscal policy works through budget or rather through its components. According to the definition of national budget presented above, it is possible to identify three tools of discretionary fiscal policy:

1. Government spending, G
2. Taxes, T
3. Government spending and taxes together, G and T

Although fiscal policy directly affects aggregate demand, eventually it influences both aggregate demand and aggregate supply. Changes in government expenditures and taxes separately or at the same time have a multiplier effect on aggregate demand. As a result of the three tools of discretionary fiscal policy, three types of multipliers can be defined:

1. Government expenditure multiplier
2. Tax multiplier
3. Balanced budget multiplier

Let us first discuss the government expenditure multiplier.

Government expenditure multiplier: The magnification effect of a change in government expenditure on goods and services on aggregate demand

Suppose the government increases its spending G by amount ΔG. Since G is a component of aggregate demand, aggregate demand AD increases as well. The increase requires an increase in production by the same amount to meet expansion of aggregate demand. Since *production = income*, the latter increases by ΔG. A fraction of extra income is consumed by households, but the rest is saved:

Marginal Propensity to Consume (MPC): The fraction of extra income that a household consumes rather than saves

Extra income generates extra consumption of $MPC \times \Delta G$ which, in turn, requires the corresponding increase in production by $MPC \times \Delta G$ which in turn increases income by $MPC \times \Delta G$. Extra income generates extra consumption $MPC \times (MPC \times \Delta G)$ and so on. It is possible to summarize the process by a means of the following table:

Round	Increase in AD	Increase in production	Cumulative increase in income, ΔY
1	ΔG	ΔG	ΔG
2	$MPC \times \Delta G$	$MPC \times \Delta G$	$\Delta G + MPC \times \Delta G$
3	$MPC \times (MPC \times \Delta G)$	$MPC \times (MPC \times \Delta G)$	$\Delta G + MPC \times \Delta G + MPC \times (MPC \times \Delta G$
n+1	$MPC^n \times \Delta G$	$MPC^n \times \Delta G$	$\Delta G + MPC \times \Delta G + MPC \times (MPC \times \Delta G) + \ldots + MPC^n \times \Delta G$

Therefore, it turns out that total increase in income ΔY is equal to

$$\Delta Y = (1 + MPC + MPC^2 + \ldots + MPC^n) \times \Delta G$$

In this regard, the government expenditure multiplier m is defined as follows

$$m = \frac{\Delta Y}{\Delta G} = (1 + MPC + MPC^2 + \ldots + MPC^n)$$

The infinite sum in brackets has a finite total equal to $\frac{1}{1 - MPC}$ and therefore, government expenditure multiplier can be written as

$$m = \frac{1}{1 - MPC}$$

For example, if $MPC = 0.6$ then

$$m = \frac{1}{1 - MPC} = \frac{1}{1 - 0.6} = 2.5$$

If we now introduce marginal propensity to save *MPS*, then according to definition of the marginal propensity to consume:

$$MPC + MPS = 1$$

and the formula for the expenditure multiplier can be written as

$$m = \frac{1}{MPS}$$

The tax multiplier works through the so-called disposable income.

Tax multiplier: The magnification effect of a change in taxes on aggregate demand

A decrease in taxes increases households' disposable income Y^d defined as

$$Y^d = Y - T$$

where Y is national income and T is tax revenue. An increase in disposable income increases consumption C because

$$C = MPC \times Y^d$$

and the multiplier process starts. Since a decrease (or increase) in taxes affects consumption indirectly through disposable income magnitude of the change is smaller compared to the government expenditure multiplier.

In addition, there is the so-called balanced budget multiplier

Balanced budget multiplier: The magnification effect on aggregate demand of simultaneous changes in government expenditures and taxes that leave the budget balance unchanged

The size of this multiplier is smaller than either the government expenditure multiplier or the tax multiplier: To achieve a balanced budget, both government expenditures and tax revenue must change in the same direction and by the same amount (recall definition of budget). However, they have opposing effects on aggregate demand: As a matter of fact, a $1 increase in government expenditures increases aggregate demand by more than a $1 increase in taxes decreases aggregate demand. Therefore, if both government expenditures and taxes increase by $1, aggregate demand still increases.

11.3 Discretionary fiscal stabilization

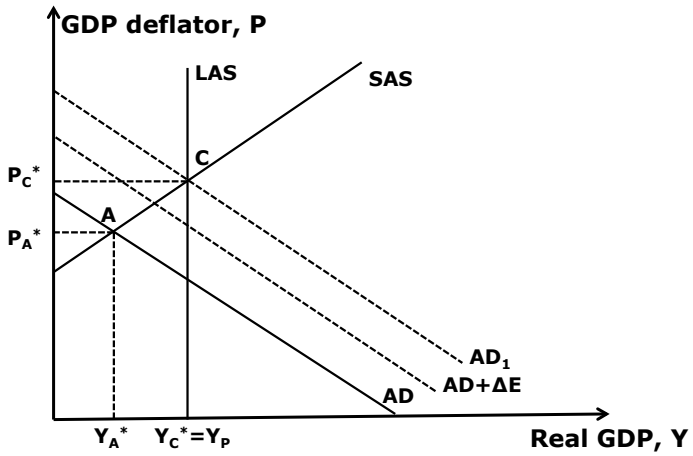
As we have already discussed, changes in government expenditures or in taxes or both at the same time affect aggregate demand. Previously we defined aggregate demand as follows:

$$AD = C + I + G + NX$$

The right hand side shows aggregate expenditures by four groups. In chapter 9, we have already identified two possible situations for economy in the short run as a result of various shocks:

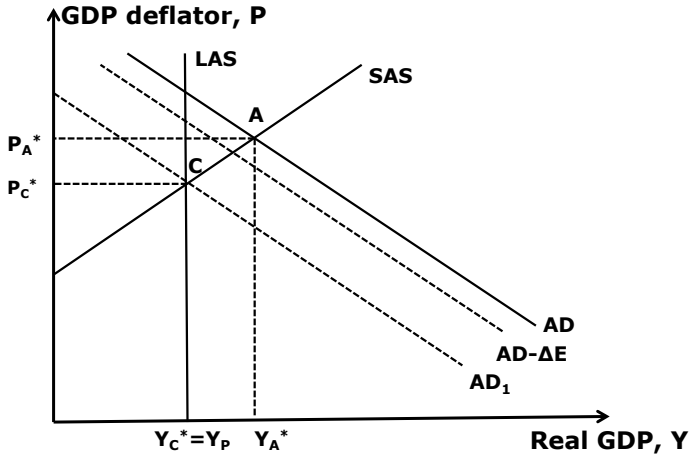
- Deflationary (recessionary) gap
- Inflationary gap

Let us analyze discretionary fiscal policy with respect to a deflationary (recessionary) gap. To eliminate the deflationary gap and restore full employment in the economy, the government takes a discretionary fiscal action described at the beginning of the previous section. The economy starts at point A or it starts with deflationary (recessionary) gap:



An increase in government expenditures or a cut in taxes increases aggregate expenditures by some amount *ΔE*. Aggregate demand increases originally, and the AD-line shifts to the right by the amount *ΔE*. New aggregate demand is AD +ΔE. This action sets off a multiplier process. As the multiplier process plays out, aggregate demand increases and the AD-line shifts rightward to AD_1. Point C is the new equilibrium with the economy at full employment level Y_P. This type of fiscal policy is called expansionary fiscal policy.

A contractionary fiscal policy can be used to eliminate an inflationary gap. Again, the economy starts at point A with inflationary gap:



A decrease in government expenditures or an increase in taxes or both at the same time shifts the AD-line to AD-ΔE line and initiates the multiplier process. As a result, eventually the AD-line shifts leftward to AD_1 and the economy returns to full employment at point C.

Frequently government expenditures are used to increase or decrease our production possibilities. For example, a highway is an input in production of transportation services. If government invests in a new highway, the volume of transportation increases and therefore, there is a supply side feedback. In general, whenever there is an increase in *G*, there is an increase in production possibilities which causes a rightward shift in aggregate supply AS. As a result of our previous discussion, both expansionary and contractionary fiscal policies resulted in changes in the price level. However, aggregate supply feedback can cause either a change or no change in the price level depending on the magnitudes of the changes in aggregate supply versus aggregate demand.

11.4 Automatic fiscal policy

Automatic fiscal policy can be defined as:

Automatic fiscal policy: Policies or institutions, built into an economic system that automatically tend to dampen business cycle fluctuations in GDP and employment without direct government intervention.

It is realized through the so-called automatic stabilizers:

Automatic stabilizers: Features of fiscal policy that stabilize real GDP without explicit action by the government

On the receipt side of national budget, usually taxes are introduced as induced taxes:

Induced taxes: Taxes that vary with real GDP

When real GDP increases during expansions (booms), wages and profits rise, but so do the taxes on these incomes which eventually shifts aggregate demand downwards back towards potential GDP level. When real GDP decreases in recessions, wages and profits fall, and so do taxes on these incomes which, in turn, shifts aggregate demand towards potential GDP as well.

On the spending side of national budget, the government can create social programs that pay unemployment benefits and welfare benefits to qualified people and businesses. Government spending on these programs is called induced transfer payments:

Induced transfer payments: Payments made under social program such as unemployment benefits and welfare benefits, the size of which depends on the state of economy

In a recession, the number of people experiencing economic hardship increases and so do induced transfer payments. In booms, the number of people experiencing economic hardship decreases and so do induced transfer payments. Both affect aggregate demand: As previously, it shifts towards potential GDP.

Overall induced taxes and induced transfer payments decrease the multiplier effect of a change in autonomous expenditures such as, for example, investment or net exports. They moderate both expansions and recessions and decrease magnitude of the short-run fluctuations in real GDP. At large, automatic fiscal policy cannot fine-tune an economy but it helps reduce short-run fluctuations in real GDP.

11.5 Fundamentals of monetary policy

Let us recall definition of monetary policy given in chapter 10:

Monetary policy: The process by which the monetary authority of a country controls the supply of money, often targeting a rate of interest for the purpose of promoting economic growth and stability

Similar to fiscal policy, monetary policy works through aggregate demand. Central bank of a country conducts monetary policy, and in a market economy it does so independently of fiscal policy. Of course, the two should be coordinated to achieve macroeconomic stability.

In practice monetary policy is based on the relationship between the rates of interest in an economy and total supply of money. Monetary policy uses a variety of tools to control one or both of these, to influence outcomes like economic growth, inflation, exchange rates and unemployment.

Therefore, monetary policy is associated with such macroeconomic aggregates as money supply, interest rate and exchange rate. The latter will be discussed separately. In this chapter we analyze money market with respect to two monetary aggregates – quantity of money and interest rate.

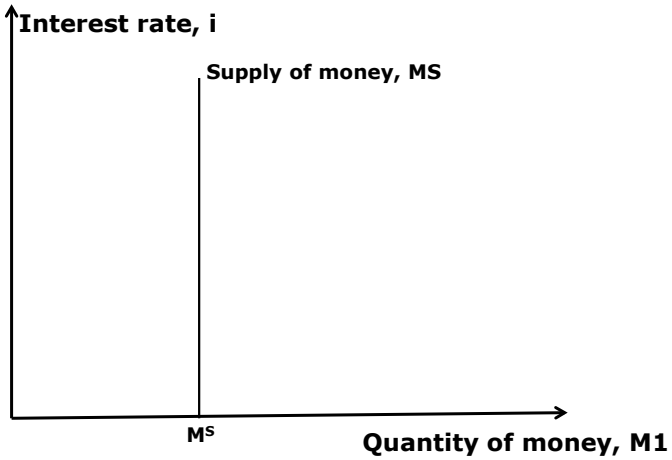
In order to understand determinants of these monetary aggregates we need to analyze the so-called money market:

Money market: A component of the financial markets for assets involved in short-term borrowing, lending, buying and selling with original maturities of one year or less

In this market, money is regarded as a commodity characterized by demand and supply. Interest rate is viewed as the price of the commodity called "money". It is so since interest is a fee paid by a borrower of assets to the owner as a form of compensation for the use of the assets. In other words, it is the price paid for the use of borrowed money or money earned by deposited funds.

Therefore, money market in economic theory is described by a two-dimensional diagram "interest rate – quantity of money". Quantity of money is measured in terms of M1, M2 or M3 as discussed in chapter 10. In their empirical work, economists prefer M1.

As we already found out, the central bank in a market economy is the principle supplier of money which implies vertical money supply line MS – not sensitive to the interest rate shown below:



Demand for money is associated with households and firms and is defined as follows:

Demand for money: The total amount of money balances that the public wishes to hold for all purposes

Households and firms demand money for various reasons. When economists say money, they mean cash in the first place. There are the following three reasons for households and firms to hold cash:

1. To carry out day-to-day transactions
2. As a precautionary measure
3. As a means for speculations

While the first two reasons are the main reasons for households and firms, the third one applies more to large businesses and to professional money managers. There are three macroeconomic variables that affect demand for money:

- The interest rate, i
- The level of real GDP, Y
- The price level, P

If an individual holds money in the form of cash, the individual forgoes opportunity to earn interest if he/she invested the money into interest-earning securities. An increase in the interest rate leads firms and households to reduce their desired money holdings because of higher opportunity costs. In turn, a reduction in the interest rate means that holding money is less costly, and hence households and firms increase their desired money holdings. Based on this logic, we can conclude:

Other things being equal, the demand for money is negatively related to the interest rate

The amount of transactions that firms and households wish to make positively depends on the level of production in the economy, which in macroeconomics is given by real GDP. Therefore:

An increase in real GDP increases the volume of transactions in the economy and therefore, leads to an increase in desired money holdings.

An increase in the price level leads to an increase in the dollar value of transaction even if there is no change in the real value of transactions (real GDP). Therefore:

An increase in the price level causes an increase in desired money holdings

Our discussion leads to the following conclusion: The demand for money as a function is:

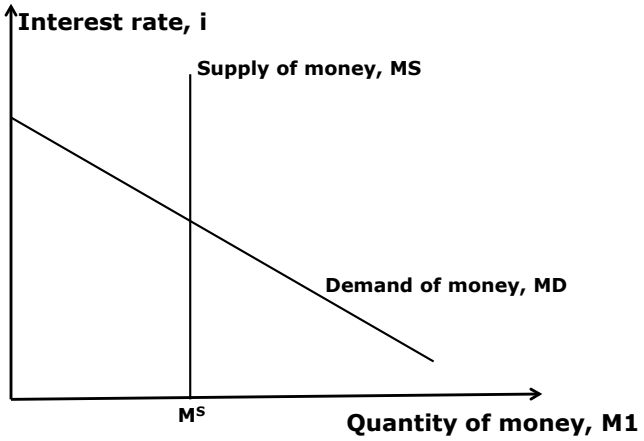
$$MD = f(i, Y, P)$$

If we now recall that nominal GDP is defined as $P \times Y$ then the demand for money becomes

$$MD = f(i, \overline{P \times Y})$$

or the demand for money is a function of the interest rate and nominal GDP.

It is necessary to emphasize that the demand for money is a function of two variables. In a two dimensional space, such a function is drawn for a fixed second variable, which is in our case nominal GDP equal to $P \times Y$. That is why in the above mathematical specification there is "bar" on top of the expression $P \times Y$. It means that MD-line is down-sloping in the "interest rate-quantity of money" space because of the negative relationship between quantity of money demanded and the interest rate i as discussed above, and it is drawn for a fixed value of nominal GDP which is $P \times Y$. Any change in nominal GDP (change in P or in Y or both) shifts the entire MD-line. Hence, the money market is represented by the following diagram:



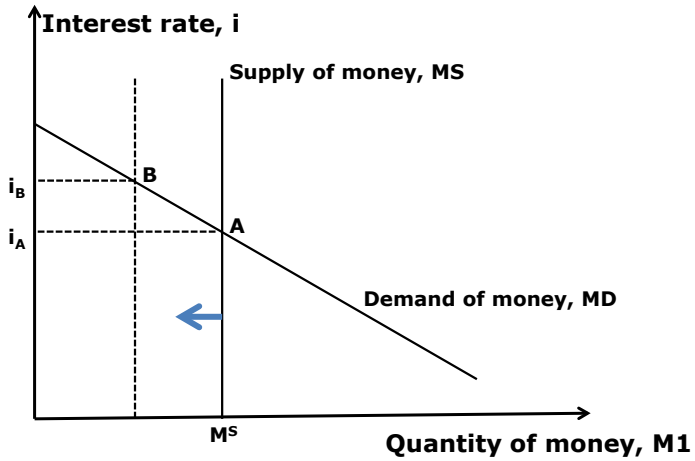
11.6 Monetary stabilization

As already discussed in chapter 10, the central bank conducts monetary policy through open market operations. Let us first analyze this process in the case of inflationary gap.

Inflationary gap is associated with high price level (inflation) and real GDP being higher than potential GDP. In such a case economists say: "Economy is overheated". It is possible to stabilize the economy by means of a contractionary monetary policy:

Contractionary monetary policy: Monetary policy that seeks to reduce the size of the money supply

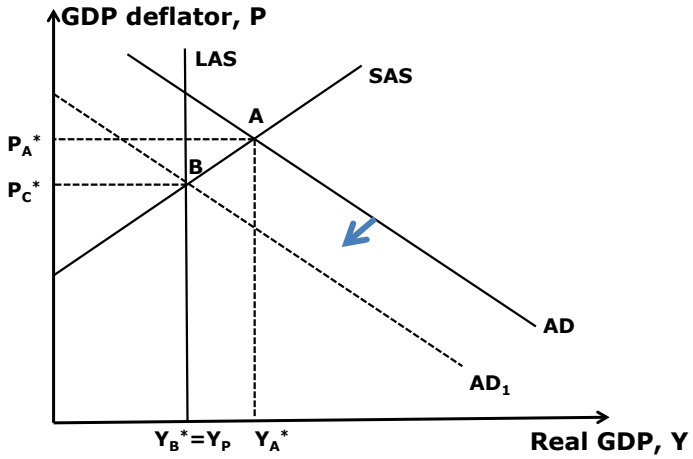
Technically the central bank sells government securities to the public. If the central bank decreases money supply through open market operations, the MS-line shifts to the left.



New equilibrium level of interest rate increases (point B in the above diagram). Three main events follow:

- Investment and consumption expenditures decrease: An increase in the interest rate means an increase in the costs of borrowing for households and firms and that is why they decrease consumption and investment
- Local currency increases in value (is explained in chapter 12)
- The multiplier process induces further decrease in consumption

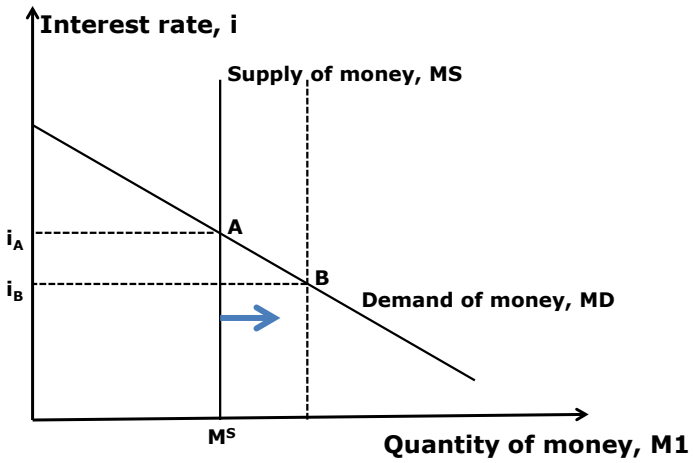
Since all of the above are expenditures or components of aggregate demand, as a result the AD-line shifts leftward, and the new equilibrium price level decreases as it is shown below (new equilibrium is point B):



Expansionary monetary policy works in opposite direction:

Expansionary monetary policy: Policy that seeks to increase the size of the money supply

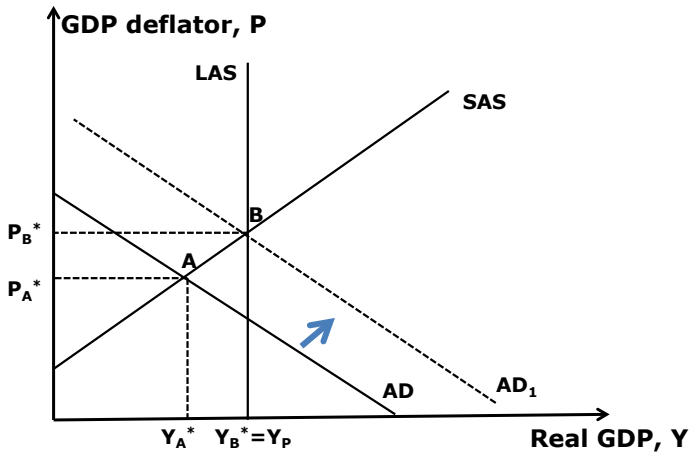
Technically the central bank buys government securities from the public. By doing that, the bank injects extra money into the system which means an increase in supply of money – rightward shift in MS:



As a result, interest rate decreases ($i_B < i_A$) and the following impacts follow:

- Consumption and investment increase because of a decrease in the costs of borrowing
- Local currency depreciates (decreases in its value)
- The multiplier effect increases consumption further

All of the above effects lead to the increase in aggregate demand – upward shift in AD as shown below:



As can be seen from the diagram, expansionary monetary policy is needed in the cases of recessions (deflationary gaps).

12 International Finance and Open Economy

Key concepts discussed in this chapter: balance of payments, current account, capital account, official reserves, net lender, net borrower, debtor nation, creditor nation, flow variable, stock variable, private sector balance, public sector balance, twin deficit, forex, foreign exchange rate, currency appreciation, currency depreciation, export effect, expected profit effect, flexible exchange rate, fixed exchange rate regime, in-between exchange rate regimes, small open economy, large open economy, purchasing power parity, international arbitrage, interest rate parity

12.1 Balance of payments

A country's balance of payments (BOP) is a record of its international activity:

BOP: A summary record of a country's transactions with the rest of the world, including the buying and selling of goods, services and assets

There are two main categories in the balance of payments:

- The current account, *CA*
- The capital account, *KA*

Current account: Record of international receipts and payments arising from trade in goods and services as well as from interest and dividends that are earned by capital owned in one country and invested in another, plus net transfers received from abroad

Technically current account includes:

- Net exports, *NX*
- Net interest on investment
- Net transfers from abroad

Net exports *NX*, also known as trade balance, is the difference between exports and imports of goods and services. It is the largest and dominating component of the current account.

Capital account: Record of receipts and payments arising from the import and export of long-term and short-term financial capital

Technically capital account includes:

- Direct investment abroad
- Portfolio investment abroad
- Other investment abroad

Direct investment refers to capital participation in a company in which the investor has a lasting or permanent interest. By convention, this is manifested by ownership of at least ten percent of the company's equity.

Portfolio investment is usually referred to as "hot money" since the investor's motive is short-term as opposed to a direct investor's lasting interest in a company. A portfolio investor will buy or sell a financial instrument at anytime there is an indication of immediate gain or loss. Portfolio holdings may be in the form of stocks, bonds and notes, and money market instruments, which are tradable in the market, and therefore can easily be acquired or disposed of. In case of stockholdings, a portfolio investor is differentiated from a direct investor if he owns less than ten percent of a company's total equity.

Other forms of investments are financial derivatives, loans (trade and non-trade), holdings of currency and deposits, and other investments.

In addition, the capital account includes the so-called official financing account. This account shows the government's transactions in its official foreign-exchange reserves:

Official reserves: The government's holdings of foreign currency

If, for example, a country's official reserves increase, it means that purchases of foreign currency increased which is recorded in the official financing account with negative sign. The reason for this is: Holding foreign money is like investing abroad which is capital outflow.

We call it balance of payments because

$$BOP = CA + KA = 0$$

which means that the sum of the balances on the two accounts always equals zero. Hence, a country with a current account deficit must either borrow more from abroad than it lends abroad or use its official reserves to cover the deficit.

In general, any surplus on the current account must be matched by an equal deficit on the capital account. A current account surplus thus implies a capital outflow. Any deficit in the current account must be matched by an equal surplus in the capital account. A current account deficit thus implies a capital inflow.

If the current account is in surplus, the country is a *net lender* to the rest of the world in the amount of the surplus or the excess in the current account transactions. Net lending occurs when the national saving is more than the country's investment in real assets:

Net lender: A country that is lending more to the rest of the world than it is borrowing from the rest of the world

If the current account is in deficit, the country is said to be a *net borrow* or a *user of funds*. In this case, the country invested more than what its national saving can finance:

Net borrower: A country that is borrowing more from the rest of the world than it is lending to the rest of the world

Identification of a country as a net lender or net borrower is mostly based on capital flows. Over a period of time, these capital flows accumulate and become stocks. With respect to stocks, it is possible to distinguish between a *debtor nation* and *creditor nation*:

Debtor nation: A country that during its entire history has borrowed more from the rest of the world than it has lent to it

Creditor nation: A country that during its entire history has invested more in the rest of the world than other countries have invested in it

The difference between *stock* and *flow* variables is in the heart of the above distinction. Borrowing and lending are flows – amounts per unit of time. Debts are stocks – total amounts owed at a point in time. Normally flows change the stocks.

However, the question is not whether it is bad or not to borrow money. The question is how the money is spent. If borrowing finances investment then it is not a problem. In such a case a higher economic growth as a result of investment financing generates higher income, and the debt will be repaid in the future. On the other hand, if borrowing finances consumption, then it becomes a problem. Interest payments accumulate which will eventually decrease consumption in the future and affect an economic performance of the borrower.

12.2 Current account balance and the twin deficits hypothesis

As already noted, net exports *NX* is the main item in the current account. According to the earlier presented composition of the current account, it can be presented as follows:

$$CA = NX + \textit{Net interest on investment} + \textit{Transfers from abroad}$$

Fluctuations in net exports are the main sources of fluctuations in the current account balance. The government budget and private savings and investment determine net exports.

It is possible to identify the following relevant macroeconomic concepts:

- Net exports: $NX = X - M$, where X is exports and M is imports
- Private sector balance = $S - I$ where S is savings and I is investment
- Public sector balance = $T - G$ where T is tax revenue and G is government purchases of goods and services; public sector balance is national budget

GDP identity states:

$$Y = C + I + G + NX$$

Let us subtract tax revenues from both sides:

$$Y - T = C + I + G - T + NX$$

Left-hand side shows disposable income available to households. Fraction of this income is saved as savings S and the rest is spent on consumption C. Therefore:

$$S + C = C + I + G - T + NX$$

Rearranging the last equation:

$$NX = (S - I) + (T - G)$$

or

$$\textit{Net exports} = \textit{Private sector balance} + \textit{Public sector balance}$$

If savings exceed investment, the surplus is lent to other private sectors of an economy. If investment exceeds savings, deficit is financed by other sectors of economy. The same is true regarding public sector: If tax revenue exceeds government spending, budget surplus is lent to the other sectors. If there is a budget deficit, other sectors of economy finance the deficit. Therefore, fluctuations in private and public sectors determine fluctuations in net exports and consequently in current account balance.

The *twin deficits hypothesis*, also called the double deficit hypothesis or twin deficits anomaly, is a concept in macroeconomics that contends that there is a strong link between a national economy's current account balance and its government budget balance as shown above. Assume that an economy is at potential output, meaning Y is fixed. In this case, if the budget deficit increases or $(T - G) < 0$, and savings remains the same, then our last equation implies that either investment I must fall or net exports NX must fall, causing a trade deficit; hence, the twin deficits arises.

12.3 Foreign exchange market

Foreign exchange market or FOREX is an economic way of describing the world's financial markets. In order to understand how this market works, we have to introduce some concepts:

The foreign exchange market: The market in which the currency of one country is exchanged for the currency of another

Foreign exchange rate: The price at which one currency exchanges for another

Currency appreciation: The rise in the value of one currency in terms of another currency

Currency depreciation: The fall in the value of one currency in terms of another currency

Given the above definition, the exchange rate between euro € and US dollar $ can be presented as follows:

$$E = US\$/€$$

For example, on May 28, 2012 the exchange rate between euro and US dollar was 1.2538. It means that €1 = US$1.25. If it increases to 1.28, it is said that the euro has appreciated.

In general, the foreign exchange market assists international trade and investment by enabling currency conversion. The foreign exchange market is the most liquid financial market in the world. Traders include large banks, central banks, institutional investors, currency speculators, corporations, governments, other financial institutions, and retail investors.

Like in any market, demand and supply determine the price of local currency in terms of foreign currency – the exchange rate. However, supply and demand for any given currency, and thus its value, are not influenced by any single element, but rather by several. These elements generally fall into three categories: economic factors, political conditions and market psychology. In its simplest form, it is possible to identify the following economic determinants behind demand for and supply of local currency.

The demand for local currency arises from all international transactions that represent payments in the country's balance of payments. The quantity of local currency demanded in the FOREX depends on:

- The exchange rate
- The interest rate in the country vis-à-vis other countries or the interest rate differential
- The expected exchange rate

It turns out that the demand for local currency in a country is a derived demand. It is so because people who demand local currency actually want to buy that country's goods and services or/and financial and capital assets. The law of demand applies to this market as well: Other things remaining the same, the quantity of the local currency demanded decreases if the exchange rate rises.

There are two separate reasons behind the derived demand for local currency:

- Exports effect
- Expected profit effect

Exports effect implies: The lower the exchange rate, the cheaper local goods and services to people from the rest of the world. As a result, the rest of the world demands more local currency to buy local goods and services – the country's exports go up.

Expected profit effect implies: The higher expected profit from holding local currency, the higher demand for local currency. Such an expectation arises if one believes that the exchange rate is going to increase. Why? If you keep in mind that the exchange rate is the price of local currency in terms of the currency of another country then you may think of a purchase of local currency as a cost to the buyer. When the current exchange rate falls (the cost is lower), for people who believe in appreciation of local currency the expected profit is higher and these people demand more local currency.

The above discussed two effects stand behind the negative (inverse) relationship between the exchange rate and the quantity of local currency demanded. Therefore, graphically demand for local currency is a down-sloping line as shown later.

Change in the value of the exchange rate leads to movements along the demand line. Change in other things remaining the same shifts the demand. What are those other things? They are:

- The interest rate differential
- The expected exchange rate

People buy financial assets based on the rate of return. If the interest rate in a country rises above the rest of the world's interest rate, then the return on investment in that country becomes higher. However, in order to invest in that country one has to buy this country's currency. It means that the interest rate differential – the difference between the country's interest rate and the world interest rate – positively affects the quantity of local currency demanded. If we define the country's interest rate as i and the world interest rate as i_W, then positive interest rate differential implies $i > i_W$

The higher expected exchange rate also positively affects the demand for local currency. If you believe in appreciation of local currency in the long-run, you will buy this currency now to make profit later.

Therefore, changes in the exchange rate differential and the expected exchange rate shift the demand for local currency.

Like demand for local currency, supply of local currency also depends on three things:

- The exchange rate
- The interest rate differential
- The expected exchange rate

However, the relationship is positive. Who are suppliers of local currency of a country in FOREX? The answer is: Importers of local goods and services as well as buyers of foreign financial assets. If the exchange rate rises, it means that foreign goods and services as well as foreign financial assets become cheaper relative to local situation. Hence the importers and buyers of foreign financial assets will buy foreign currency supplying local currency at the same time. Again there are two effects behind this process:

- Imports effect
- Expected profit effect

We have already explained the imports effect. Expected profit effect works the same way as in the case of demand for local currency. If you believe that local currency is going to depreciate (decrease in value), you are going to get rid of this currency now. Hence you will sell (supply) local currency and buy foreign currency because you expect higher profit from holding foreign currency. And the higher the current exchange rate, the higher the amount of local currency you will supply now, looking for higher profit in the future.

The described relationship between the exchange rate and the quantity of local currency supplied is positive, other things remaining the same. Again by other things we mean:

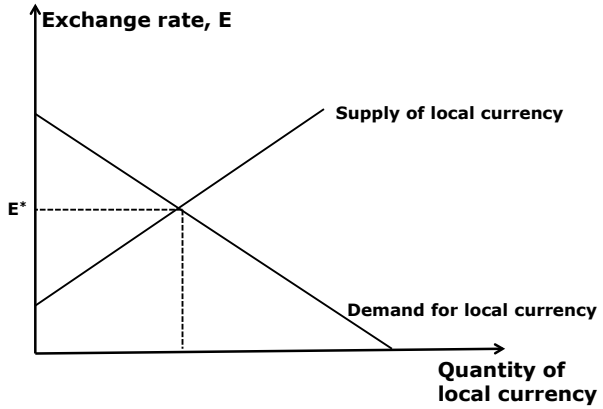
- The interest rate differential
- Expected exchange rate

With respect to the supply, the two work in opposite direction compared to the demand for local currency:

- Lower interest rate differential ($i < i_W$) causes people to demand more foreign currency and supply more local currency;
- Lower expected exchange rate causes people to sell local currency (increase in supply) and buy foreign currency.

Similar to demand, both shift supply of local currency.

Market equilibrium occurs when the quantity of local currency supplied exactly equals the quantity of local currency demanded. Exchange rate E^* associated with the equilibrium is called equilibrium or market exchange rate:



12.4 Small open economy

It turns out that in current global world any country is highly integrated into the world economy. However, economic developments in a country can have different impacts on the world economy and financial markets. Economic theory distinguishes between small open economy and large open economy. A small open economy is defined as follows:

A small open economy (SOE): An economy that participates in international trade, but is small enough compared to its trading partners that its policies do not alter world prices, interest rates, or incomes

Thus, the countries with small open economies are price takers. This is unlike a *large open economy*, the actions of which do affect world prices and income.

Given the above, basic assumption in this framework is that interest rate is given *exogenously* as the world interest rate i_w. Exogenously means that it is defined outside our model. This assumption implies that in the long-run equilibrium the interest rate in a SOE is equal to the world interest rate or

$$i = i_W$$

However, interest rate differential which the difference between the two interest rates is a driving force in a SOE.

Positive interest rate differential $i - i_W > 0$ implies capital inflow: Financial capital flows into the SOE because the rate of return in this economy is higher, creating higher demand for domestic currency. Negative interest rate differential $i - i_W < 0$ implies capital outflow: Financial capital flows from the SOE because the rate of return in the rest of the world is higher, creating higher supply of domestic currency.

As already discussed, there are two types of macroeconomic policies – fiscal policy and monetary policy. In addition, there are two regimes with respect to the exchange rate – flexible exchange rate and fixed exchange rate.

Flexible Exchange Rate: A currency exchange rate determined by the market forces of supply and demand and not interfered with by government action

Fixed Exchange Rate: A currency exchange rate pegged by government and therefore prevented from rising or falling

Those are two extreme regimes with respect to the exchange rate and in academic discussions, the decision is often posed as a choice between a fixed or a flexible exchange rate. In reality however, there are different varieties of fixed and flexible arrangements, providing a range of alternatives. The different alternatives have different implications for the extent to which national authorities participate in the foreign exchange markets. According to their degree of flexibility, exchange rate regimes are arranged into three categories (Edwards, Sebastian, and Miguel Savastano, 1999, "Exchange Rates in Emerging Economies: What do we Know? What do we Need to Know?" NBER Working Paper No. 7228, Cambridge, Massachusetts):

1. Currency unions, dollarized regimes, currency boards and conventional fixed pegs are defined as "fixed-rate regimes";
2. Horizontal bands, crawling pegs and crawling bands are grouped into "intermediate regimes";
3. Managed and independent floats are defined as flexible regimes

Monetary Union is a zone where a single monetary policy prevails and inside which a single currency or currencies, which are perfect substitutes, circulate freely. A Monetary Union has common monetary and fiscal policy to ensure control over the creation of high-powered money and the expansion of government debts; it has a central management of the common pool of foreign exchange reserves, external debts and exchange rate policies. The Monetary Union has common regional monetary authority i.e. common regional central bank, which is the sole issuer of economy wide currency, in the case of a full currency union.

Dollarization/Euroization: A foreign currency acts as legal tender. Dollarization is a summary measure of the use of foreign currency in its capacity to produce all types of money services in the domestic economy. Monetary policy is delegated to the anchor country.

Currency Board is monetary regime adopted by countries that intend to discipline their Central Banks, as well as solve their external credibility problems by "tying their hands" with institutionally binding arrangements. A currency board combines three elements: an exchange rate that is fixed to an "anchor currency"; automatic convertibility or the right to exchange domestic currency at this fixed rate whenever desired; and a long-term commitment to the system.

Fixed peg means fixed rate against a single currency or a currency basket.

Crawling peg: A rule based system for altering the par value, typically at a predetermined rate or as a function of inflation differentials. It is an attempt to combine flexibility and stability. Often used by (initially) high inflation countries pegging to low inflation countries in attempt to avoid trend real appreciation. At the margins a crawling peg provides a target for speculative attacks. Among variants of fixed exchange rates, it imposes the least restrictions, and may hence yield the smallest credibility benefits. The credibility effect depends on accompanying institutional measures and record of accomplishment.

Bands exchange rate is flexible within a present band; endpoints are defended through intervention, typically with some intra-band intervention. An attempt to mix market-determined rates with exchange rate stabilizing intervention in a rule based system.

Managed float exchange rates are determined in the foreign exchange market. Authorities can and do intervene, but are not bound by any intervention rule; often accompanied by a separate nominal anchor, such as inflation target. The arrangement provides a way to mix market-determined rates with stabilizing intervention in a non-rule-based system.

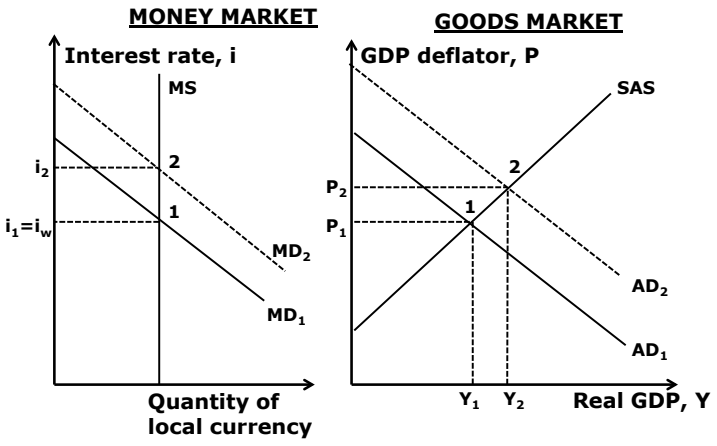
Pure float: The exchange rate is determined in the market without public sector intervention.

We will follow academically accepted framework with just two regimes – flexible (float) and fixed since those two are fundamentals of the above presented real-world regimes. We are going to analyze the following four frameworks:

1. Fiscal policy with a flexible exchange rate
2. Fiscal policy with a fixed exchange rate
3. Monetary policy with a fixed exchange rate
4. Monetary policy with a flexible exchange rate

12.5 Fiscal Policy in a small open economy with flexible exchange rate

We begin our discussion by assuming that both, the goods market and money market in a SOE are in equilibrium at point 1. In equilibrium, domestic interest rate is equal to the world interest rate which is reflected in the following diagram:



Suppose that the government increases its spending *G* in an attempt to stimulate aggregate demand AD. In the short run, AD-line shifts rightward form AD_1 to AD_2, and a new equilibrium in the goods market is point 2. At this point both general price level *P* and total output *Y* increase, which increases nominal GDP = $P \times Y$. Nominal GDP is a shift parameter of the demand for money function MD, and that is why the MD-line shifts from MD_1 to MD_2. Point 2 is the new short run equilibrium in the money market as well. So, the following chain of events takes place in the short run:

$$G\uparrow \rightarrow \text{rightward shift in AD} \rightarrow P \text{ and } Y\uparrow \rightarrow \text{rightward shift in MD} \rightarrow i\uparrow$$

As a result, domestic interest rate i_2 becomes higher than the world interest rate i_W or SOE generates positive interest rate differential. Positive differential implies that savings from the rest of the world flow into SOE since the return on these savings is higher in the SOE than in the rest of the world. The rest of the world residents supply their currencies to buy local currency or demand for local currency shifts to the right which pushes local currency to appreciate.

When local currency appreciates, it becomes more expensive for foreigners to buy local goods and services, and therefore, exports in the SOE fall. For the same reason, it becomes cheaper for locals to buy foreign goods and services which increases imports. Since net exports *NX* is the difference between exports *X* and imports *M*, *NX* decrease. Net exports are a component of aggregate demand, and its decrease shifts the AD-line from AD_2 to AD_1. Nominal GDP decreases, resulting in a backward shift of the MD-line. The AD-line and MD-line shift backwards until old equilibrium at point 1 is restored in both markets.

Therefore, the following adjustment process between short run and long-run is observed:

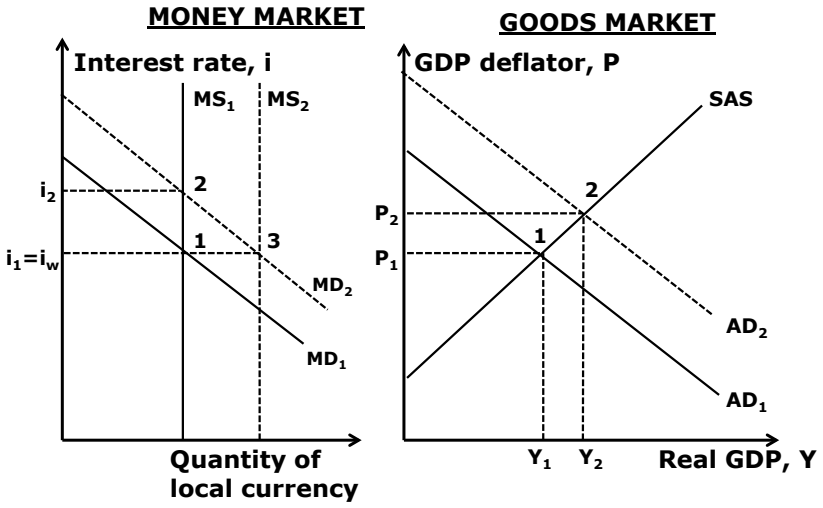
$i_2 > i_W \rightarrow E^{\uparrow} \rightarrow X\downarrow$ while $M\uparrow \rightarrow NX\downarrow \rightarrow$ leftward shift in AD $\rightarrow$ P and $Y\downarrow \rightarrow$ nominal $GDP\downarrow \rightarrow$ leftward shift in MD $\rightarrow$ old equilibrium is restored

So, in a small open economy with flexible exchange rate, there is no lasting long-run effect on domestic interest rate as a result of increase in government spending or in general fiscal policy. Jobs were created in the government sector due to increase in the government spending, but they were destroyed in the exports sector. Hence, we ended up with the crowding-out effect that worked its way through the exchange rate and net exports.

Conclusion: Fiscal policy in a small open economy under flexible exchange rate has no lasting effect. The only result is appreciation of the domestic currency

12.6 Fiscal Policy in a small open economy with fixed exchange rate

Again we start in equilibrium in both markets and increase government spending *G*:



In the short run, the developments in the economy are the same as under flexible exchange rate or:

G↑ → rightward shift in AD → P and Y↑ → rightward shift in MD → i↑

Short run equilibrium is point 2 in both markets. Positive interest rate differential and the resulting inflow of foreign currency put pressure on local currency. However, in order to keep it from appreciating, the central bank in SOE buys some of the foreign currency injecting extra local currency. This action results in an increase in money supply in the SOE and the MS-line shifts from MS_1 to MS_2 until the positive interest rate differential is eliminated or $i_1 = i_W$ in the long run. New equilibrium in the money market is point 3. So, the following adjustment from the short run to the long run is observed:

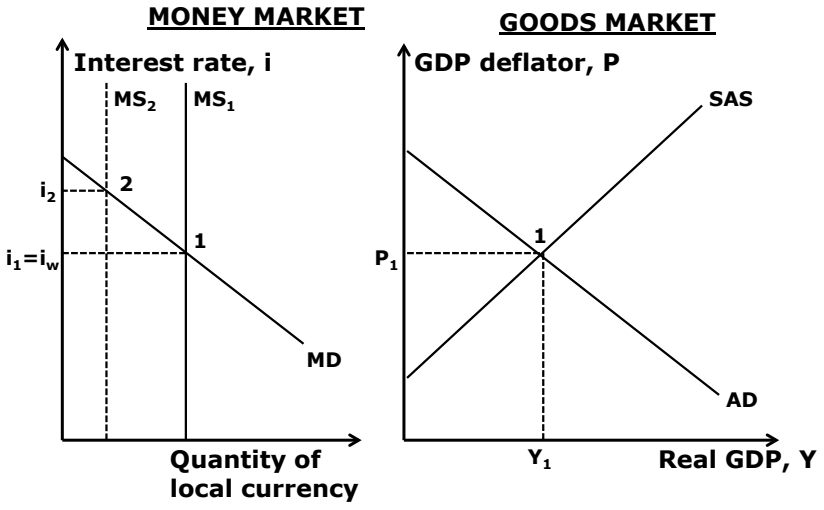
$i_2 > i_W \rightarrow$ inflow of foreign currency $\rightarrow$ central bank buys foreign currency $\rightarrow$ MS↑ $\rightarrow$ rightward shift in MS $\rightarrow i_1 = i_W$

Since exchange rate remains the same, aggregate demand stays at AD_2, and point 2 is the new long-run equilibrium in the goods market. As a result of the fiscal policy, the economy experiences more jobs but higher inflation.

Conclusion: Fiscal policy in a small open economy under fixed interest rate does have lasting effect, increasing general price level (inflation) and total output if it is supported by the central bank

12.7 Monetary policy in a small open economy with fixed exchange rate

Here we are going to analyze the consequences of a contractionary monetary policy. Suppose that the central bank in a SOE decreases money supply by selling government bonds to the public:



Decrease in money supply shifts MS-line to the left from MS_1 to MS_2, and positive interest rate differential arises in the short run at point 2 as the short-run equilibrium in the money market. The following chain of events takes place:

$$MS\downarrow \rightarrow \text{leftward shift in MS} \rightarrow i_2 > i_W$$

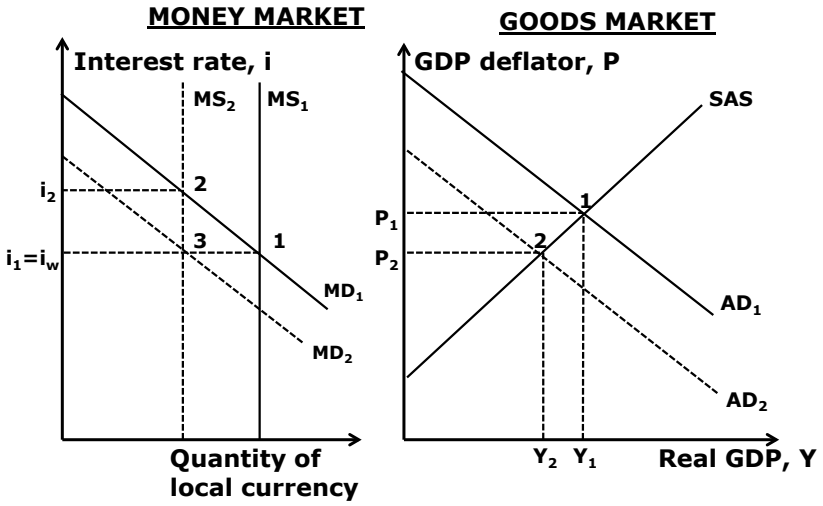
Positive interest rate differential implies inflow of foreign currency into the SOE. With a fixed exchange rate policy, the central bank is obliged to buy the increased supply of the foreign currency. In doing so, the bank injects more local currency, pushing the MS-line all the way back to its original position where $i_1 = i_W$. Therefore, point 1 is the final long-run equilibrium in both markets.

Aggregate demand AD is not affected because the exchange rate is unchanged. Hence, it turns out that the central bank in the SOE has merely performed two open market operations that cancelled each other: The bank started with open market sale of government bonds and ended up with buying foreign currency.

Conclusion: Monetary policy in a small open economy with a fixed exchange rate has no lasting effect

12.8 Monetary policy in a small open economy with flexible exchange rate

Once again we analyze the proceeds of the contractionary monetary policy:



In the short run, the decreased money supply MS shifts the MS-line to the left from MS_1 to MS_2, and positive interest rate differential results. With flexible exchange rate, the central bank in the SOE does not intervene, and inflow of the foreign currency causes an increase in demand for local currency. As a result, local currency appreciates. Hence, it becomes more expensive for foreigners to buy local goods and services, which decreases exports *X*. At the same time, it is now cheaper for locals to buy foreign goods and services, which increases imports *M*. Net exports *NX* decreases, shifting aggregate demand from AD_1 to AD_2. Point 2 is the short-run equilibrium in both markets. The following chain of events takes place:

$MS\downarrow \rightarrow$ leftward shift in MS $\rightarrow i_2 > i_W \rightarrow E\uparrow \rightarrow X\downarrow$ while $M\uparrow \rightarrow NX\downarrow \rightarrow$ leftward shift in AD $\rightarrow$ P and $Y\downarrow$

However, with both general price level *P* and total output *Y* decreased, nominal GDP = $P \times Y$ decreases as well. Since nominal GDP is a shift parameter in the demand for money function, the MD-line shifts to the left from MD_1 to MD_2 until positive interest rate differential is eliminated. Point 3 is the new long-run equilibrium in the money market. The adjustment process between the short run and long run is as follows:

P and Y$\downarrow \rightarrow$ nominal GDP$\downarrow \rightarrow$ leftward shift in MD $\rightarrow i_1 = i_W$

In the end, contractionary monetary policy under flexible exchange rate results in two very undesirable effects – deflation (decrease in general price level) and unemployment.

Conclusion: Monetary policy in a small open economy under flexible exchange rate has a long lasting effect

The impacts of different policy regimes are summarized in a table below, answering the following question: Does the policy have a long-lasting impact on economy?

Type of policy	Fixed exchange rate	Flexible exchange rate
Fiscal policy	Yes	No
Monetary policy	No	Yes

It is possible to perform the same analysis for expansionary monetary policy and contractionary fiscal policy. The results are opposite of the ones discussed above.

It appears to be that the exchange rate plays the role of a shock absorber: When aggregate demand falls, a depreciating local currency provides one way for demand to recover through increase in net exports and vice versa. However, this degree of freedom is lost if the exchange rate is fixed.

12.9 Large open economy

Formal definition of a large open economy is

Large open economy: An economy that participates in international trade and its policies have significant impact on world prices, interest rates, or incomes

In theoretical sense, a large open economy, like the U.S. or European Union (EU), is in between two extremes – small open economy and closed economy. For example, if government increases its spending *G* then:

Closed economy: G$\uparrow \rightarrow$ AD shifts to the right $\rightarrow$ nominal GDP$\uparrow \rightarrow$ MD shifts upwards $\rightarrow i\uparrow \rightarrow I\downarrow \rightarrow$ crowding out of investment:

$Y = C + I\downarrow + G\uparrow$

An increase in G is bigger than a decrease in I and overall impact is positive: goods market – P and Y ↑; money market – i↑

SOE: G↑ → AD shifts to the right → nominal GDP↑ → MD shifts upwards → i↑ → capital inflow → appreciation of local currency → NX↓

Y = C + I + G↑ + NX↓

A decrease in NX completely offsets an increase in G: goods market is unchanged; money market is unchanged

Large open economy: G↑ → AD shifts to the right → nominal GDP↑ → MD shifts upwards → i↑ → I↓ → capital inflow → appreciation of local currency →NX↓

Y = C + I↓ + G↑ + NX↓

A decrease in NX and I is usually smaller than an increase in G: goods market – P and Y↑; money market – i↑ but the impact is smaller than in closed economy.

Note: SOE described in this chapter assumes perfect capital mobility. If it is not the case, outcomes of fiscal and monetary policies are less predictable and usually tend to be somewhat between small open economy outcome and closed economy outcome depending on degree of capital mobility. It means that in this case outcomes in a SOE are similar to a large open economy but with smaller magnitudes.

12.10 Exchange rate expectations

As we have already seen, both demand for and supply of currency in any open economy have the same common influences. Therefore, when these influences change, both demand and supply shift which results in fluctuations of the equilibrium exchange rate E^*. In addition, exchange rate expectations cause fluctuations as well.

There are several theories that describe the exchange rate expectations. Amongst them there are two the most popular views on formation of the exchange rate expectations:

- Purchasing power parity (PPP)
- Interest rate parity (IRP)

Purchasing power parity (PPP): A situation in which money buys the same amount of goods and services in different currencies

Expectations about the exchange rate are usually formed with respect to the long-run (equilibrium) value. In this regard, the PPP theory states that in the long run the nominal exchange rate E moves primarily as a result of the difference in the price level between two countries or the long-run nominal exchange rate is

$$E^* = \frac{P_F}{P}$$

where P_F is foreign price level (cost of foreign basket of goods in foreign currency) and P is domestic price level (cost of domestic basket of goods in domestic currency).

Therefore, if foreign prices are higher than domestic (which actually means if foreign inflation is higher than domestic), then according to the PPP, the expected exchange rate is going to increase or economic agents expect domestic currency to appreciate in the long-run. In general, under PPP, a currency with a higher inflation rate is expected to depreciate vis-à-vis a currency with a lower inflation rate.

Therefore, if inflation in a SOE is 10% while inflation in the US is 3%, then local currency in the SOE is expected to depreciate approximately by 10% – 3% = 7%.

This is often called relative PPP, in contrast to a more stringent absolute PPP under which the level of exchange rate will be determined to equalize levels of prices across countries.

PPP is subject to the following two assumptions that do not hold in real world:

- Goods in different countries are perfect substitutes
- International arbitrage is possible

International arbitrage: The practice of taking advantage of a price difference between two or more markets in different countries

Of course, goods in different countries are not perfect substitutes and international arbitrage is not possible because of transportation costs and non-traded goods such as output of construction industry or various services.

Interest rate parity: A situation in which the interest rate in one currency equals the interest rate in another currency when exchange rates are taken into account

Assumption of $i = i_W$ in the long-run we used to analyze a SOE is known as perfect capital mobility: Whenever it is not satisfied, there is either immediate capital inflow or capital outflow to eliminate the interest rate differential.

However, in real life we see that the interest rate in a SOE and the world interest rate i_W (usually approximated by the US interest rate in empirical work) follow closely each other, but there is a recognizable difference between the two. Economists assign it to the so-called exchange rate expectations.

Under interest rate parity, a currency with a higher interest rate is expected to depreciate by the amount of interest rate differential, thereby equalizing expected returns from investments in two currencies. The argument goes as follows:

If you invest in financial assets in a SOE, then you will earn the rate of return i. If instead you invest in US financial assets, then you will earn the rate of return i_{US} (remember, we use the US interest rate as a proxy to the world interest rate) plus premium on appreciation of US$ with respect to local currency, which is depreciation of local currency. Interest rate parity is achieved through

$$i = i_{US} + \textit{depreciation of local currency}$$

According to the above discussion, if $i = 5\%$ and $i_{US} = 3\%$ it means that local currency is going to depreciate approximately by 5% – 3 % = 2% vis-à-vis US$.

There are two basic assumptions associated with the interest rate parity that do not hold in real world:

- Perfect capital mobility
- Perfect substitutability of domestic and foreign financial assets

13 Economic Growth and Development

Key concepts discussed in this chapter: growth rate, real GDP per person, labor productivity, capital productivity, multi-factor productivity, total productive hours, human capital, productivity curve, diminishing marginal returns, classical growth theory, subsistence level, neoclassical growth theory, exogenous growth model, new growth theory, endogenous growth model, endogenous technological change, increasing marginal returns, economic freedom, property rights, sustainable development

13.1 Economic growth, growth rate and economic development

Economic growth is best defined as a long-term expansion of the productive potential of the economy. Formal definition of economic growth is:

Economic growth: An increase in the capacity of an economy to produce goods and services, compared from one period of time to another

Therefore, economic growth is a sustained expansion of production possibilities measured as the increase in real GDP over a period of time. In economics, economic growth is expressed in terms of economic growth rate:

Economic growth rate: The rate of change of real GDP expressed as a percentage per year

Mathematically:

$$g_Y = \frac{(Y_t - Y_{t-1})}{Y_{t-1}} \times 100\%$$

where g_Y is the growth rate of real GDP, Y_t is real GDP in current year and Y_{t-1} is real GDP in previous year. The growth rate tells us how rapidly the total economy is expanding.

On the other hand, the standard of living depends on real GDP per person defined as follows:

Real GDP per person: Real GDP divided by the population

Mathematically in any given year t

$$y_t = \frac{Y_t}{N_t}$$

where N_t is population in year t.

Hence contribution of real GDP growth to the change in the standard of living depends on the growth rate of real GDP per person. Let us denote the growth rate of real GDP per person as g_P and the growth rate of population as g_N. Differentiating the above expression with respect to time and using definitions of growth rates, we arrive at

$$g_p = g_Y - g_N$$

The last expression makes it clear that real GDP per person grows only if real GDP grows faster than population. That is how economic growth and economic development are linked: If economy grows at a rate lower than population growth, there is no increase in standards of living which implies bad economic development.

13.2 Sources of economic growth

One of the most important concepts associated with economic growth is productivity. There are many different productivity measures. The choice between them depends on the purpose of productivity measurement and on the availability of data. From a macroeconomic standpoint, productivity measures can be classified as single factor productivity measures – relating a measure of output to a single measure of input – or multifactor productivity measures – relating a measure of output to a bundle of inputs. Therefore, usually the following measures are most frequently used:

- Labor productivity
- Capital productivity
- Multifactor productivity or total factor productivity

The following are definitions of these measures:

Labor productivity: The quantity of real GDP per hour of labor

Capital productivity: The quantity of real GDP per unit of capital

Multifactor productivity: The ratio of real GDP to the combined input of labor and capital

Labor productivity is the most popular measure because of its interpretation and easiness of measurement. Productivity of individual labor shows how human factor is used according to the specific conditions, organization, qualification and intensity of work. It is measured by the quantity of the goods and/or services produced per unit of time as defined above.

In this regard, the growth of labor productivity means the process through which the same volume of work results in a bigger quantity of goods and services produced or the same quantity of goods and services can be produced with a smaller volume of work. This process assumes important changes in the entire production process. Therefore, labor productivity captures a special relationship between the results of the production and the consumption of work used in this production.

Based on definition of labor productivity, mathematically it can be presented as

$$Labor\ productivity = Real\ GDP/Total\ productive\ hours$$

We can re-write the latter as

$$Real\ GDP = Labor\ productivity \times Total\ productive\ hours$$

According to this simple interpretation, all influences on real GDP growth can be divided into those that increase

- Total productive hours
- Labor productivity

Over time in general total productive hours tend to increase. This growth mostly comes from the growth of labor force. In turn, labor force depends on population and the labor force participation rate. Therefore, it appears to be that total productive hours mostly depend on demographic factors and long-term preferences of a society. It is almost impossible to affect these factors in the short-run. This means that in order to achieve economic growth in the short-run labor productivity should be enhanced. In other words, labor productivity is the driving force of the economic growth.

Most discussions of economic growth are based on the following four determinants:

1. Growth in labor force
2. Growth in human capital
3. Growth in physical capital
4. Technological improvement

Based on those determinants it is possible to identify three fundamental sources behind enhancement in labor productivity and consequently economic growth:

- Investments in physical capital
- Expansion of labor and human capital
- Discovery of new technology

Investments in physical capital increase the amount of capital per workers or, in other words, our workers have more tools and machines to work with which increases their productivity.

Human capital is defined as follows:

Human capital: The accumulated skills and knowledge of human beings

Expansion of human capital comes from the following three sources:

- Education;
- Training;
- Job experience.

Expansion of human capital is the most fundamental source of economic growth because it directly increases labor productivity and is the source of the discovery of new technologies. Of course, the discovery of new technologies itself has significantly contributed to economic growth, and there are many examples in our history to support this statement.

In order to understand different theories of economic growth, it is necessary to introduce some analytical tools.

13.3 The productivity curve

Formal definition is:

The productivity curve: The relationship between real GDP per hour of labor and the quantity of capital per hour of labor with a given state of technology and human capital stock

Mathematically:

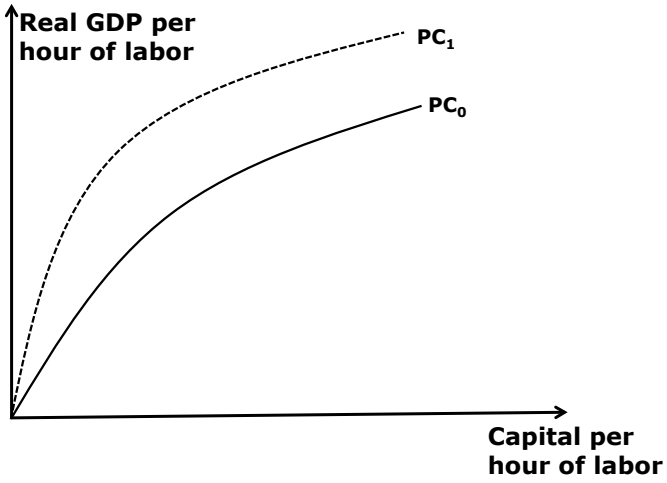
$$y = f(k)$$

where y is output per hour of labor (labor productivity) and k is capital per hour of labor. The two are defined as follows:

$$y = \frac{Y}{L} \quad \text{and} \quad k = \frac{K}{L}$$

where Y is real GDP, K is capital stock and L is quantity of labor.

Graphically:



An increase in the quantity of capital per hour of labor is associated with the movement along the productivity curve PC_0. However, an increase in human capital or/and improvement in technology leads to a higher productivity curve PC_1.

With constant average hours per worker and labor force participation rate, total productive hours grow at the same rate as population. Capital stock grows at the rate determined by savings and investment.

If the capital stock grows faster than labor, capital per hour of labor increases and vice versa. However, the growth from capital alone is limited by diminishing marginal returns:

The Law of Diminishing Marginal Returns: As the quantity of one input increases with the quantities of other inputs remaining the same, output increases but by ever smaller increments

Applied to capital, the law states that if a given number of hours of labor use more capital with the same technology, the additional output that results from the additional capital gets smaller as the amount of capital increases. Robert Solow – the Nobel Prize Winner – discovered that on average with no change in human capital and technology, a 1% increase in capital per hour of labor brings a 1/3% increase in labor productivity.

13.4 Theories of economic growth

We are going to discuss three growth theories:

- Classical growth theory
- Neoclassical growth theory
- New growth theory

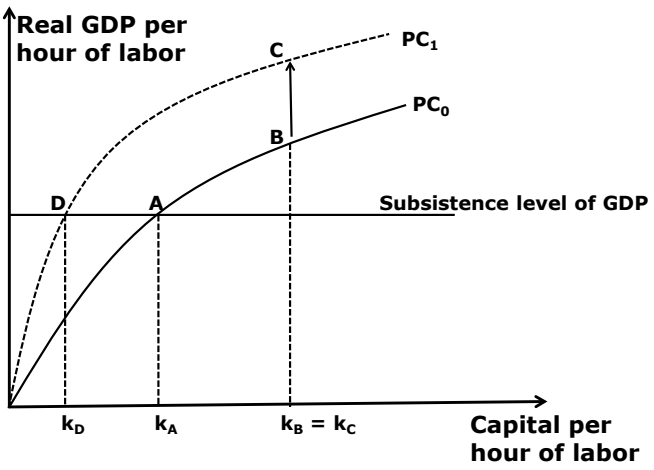
The *classical growth theory* consists primarily of the eighteenth-century pioneering work of Adam Smith, David Ricardo and Thomas Robert Malthus. Together, these three provided the foundations of modern growth theory. For the classical economists, interest in growth developed out of the philosophical question of progress – a basic tenet of Enlightenment thought that applied equally to ideas, scientific innovations, social norms, and more generally the material bases of society. On this basis, they sought a general account of the forces and mechanisms that influenced economic growth.

They recognized productive investment and capital accumulation as the principal impetuses to growth – processes that, under capitalism, involved the reinvestment of profits. They also emphasized the contributing effects of technological change, visible in the growing division of labor and in changes in methods of production. Malthus, in particular, also took into account population growth as providing both an increase in the supply of labor and an escalating risk of profit erosion, as population outstripped economic growth. This was one version of their general view that growth produces counter-veiling tendencies that undermine it and that ultimately yield a condition of stagnation.

Let us take a closer look at Malthusian point of view. This is a very pessimistic viewpoint, which states that growth of labor productivity is temporary. Eventually the higher output results in larger population and output per person returns to its subsistence level.

Subsistence level: The level associated with the basic human needs such as food, water, shelter, and others

To illustrate the theory, let us apply our productivity curve framework:



The economy starts out at point A on productivity curve PC_0 with real GDP at the subsistence level and some specific level of population. The economy moves to point B as capital per hour of labor increases from k_A to k_B and real GDP per person increases above the subsistence level. The economy moves to point C as technological advance and/or the accumulation of human capital increase productivity, and the productivity curve shifts upward to PC_1. With real GDP above the subsistence level, households have higher incomes which lead them to increase their families' size. As a result, the economy moves towards point D as the population grows and capital per hour of labor decreases. At point D, the economy is back to the subsistence level of real GDP per hour of labor. Therefore, according to this theory economic growth is temporary.

Neoclassical growth theory was developed independently by Robert Solow and Trevor Swan in 1956 and is now known as Solow-Swan growth model or *exogenous growth model*. Mathematically it can be represented by the so-called growth accounting equation:

$$g_Y = g_A + \alpha g_K + (1-\alpha) g_L$$

where g_Y is economic growth in the way we defined it previously, g_A is technological progress, g_K is growth in physical capital, g_L is growth in labor, and α is the share of capital in GDP.

In this model of economic growth, capital accumulation represented by the growth rate g_K is the driving force of the economic growth in the short-run. Capital accumulation leads to improvements in living standards but because of the law of diminishing marginal returns, these improvements become smaller with each additional increment of capital. With respect to our productivity curve PC, capital accumulation leads to a movement along the PC curve but vertical increase in the PC gets smaller and smaller.

In order to shift the PC upwards and speed up economic growth, changes in human capital and/or technological innovations are needed. In general, this theory states that real GDP per person will increase as long as technology keeps advancing. Real GDP will grow at a rate equal to the population growth plus the rate of productivity growth induced by technological change and the accumulation of human capital. So according to the neoclassical growth theory, economic growth will persist.

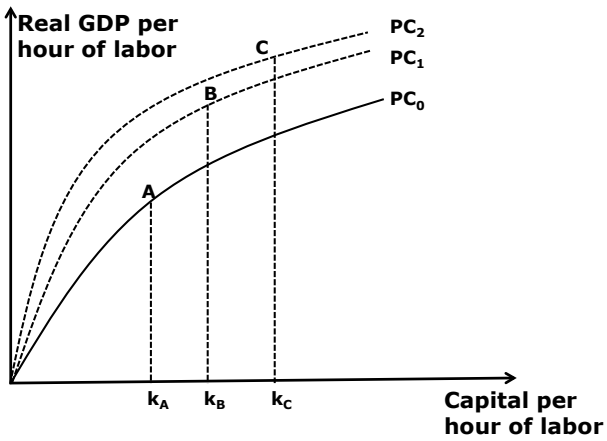
In general, neoclassical growth theory asserts that population growth and the pace of technological change determine, but are not themselves influenced by, the growth rate of real GDP. However, the theory does not explain what determines technological change. It is assumed that technological change results from chance or it enters the earlier developed framework exogenously. That is why this growth model is called the *exogenous* growth model.

So, diminishing marginal returns and exogenous technological change are the major features of the neoclassical growth theory. *The new growth theory* emphasizes *endogenous technological change* and *increasing marginal returns*.

In neoclassical growth theory, innovation increases the amount of output produced from a given level of factor inputs. However, this innovation is itself unexplained. In the new growth theory technological change stems from research and development and from innovating activities. When people discover a new product or technique they consider themselves lucky. But the pace at which new discoveries are made is not determined by chance. It depends on how many people are looking for a new technology and how intensively they are looking. Profit is a driving motive of the technological change. On the other hand, the forces of competition decrease profits. Thus to increase profits, people constantly seek either lower-cost methods of production or new and better products for which people are willing to pay higher prices.

In addition, new growth theory emphasizes the possibility of historical increasing returns: As investment in some new area, product or production technology proceeds through time, each increment of investment is more productive than previous increments. There are different sources of increasing marginal returns. For example, human capital is subject to increasing marginal returns since next generation is smarter than the previous one. Therefore, investments in human capital can be more productive than investments in physical capital. According to the new growth theory, people decide how long to stay in school, what to study and how hard to study. When they graduate, people make choices about job training and on-the-job learning. All these choices govern the speed at which human capital grows.

So it turns out that people are driven by the profit motive which makes them aggressive to develop new products and technologies as old products and technologies become available to everyone. It is possible to illustrate the fundamentals of this theory graphically:



Economy starts out at point A on productive curve PC_o. An increase in capital per hour of labor brings a movement along PC_0 and the expansion of human capital and technological change increase labor productivity and shift the production curve upward to PC_1. The economy moves to point B. The process repeats. The economy moves to point C and then to points of yet greater capital per hour of labor and labor productivity.

13.5 Preconditions and policies for economic growth

The key reason why economic growth is either absent or slow is that some societies lack the incentives that encourage growth-producing activities. In general, there are economic preconditions as well as political preconditions for persistent economic growth. Here we discuss only economic preconditions.

It is necessary to mention economic freedom as the most important economic precondition:

Economic freedom: A condition in which people are able to make personal choices, their property is protected, and they are free to buy and sell in markets

This definition emphasizes the following three points:

1. People are able to make personal choices
2. Their property is protected
3. Markets are free and competitive

In a democratic society these points are present and satisfied all the time. However, in authoritarian societies these points are not satisfied. The best example is former socialist countries. Production and consumption of almost all goods and services in those countries were organized and managed by the central government. Theoretically government can achieve efficiency in an economy but only under very restrictive conditions which under normal circumstances are not satisfied. In turn, competitive markets achieve efficiency automatically. In addition, private property was prohibited in the socialist economies by law. Moreover, profit motive was also punishable.

In a democratic society, the rule of law, an efficient legal system and the ability to enforce contracts are essential foundations for creating economic freedom. Economic freedom requires protection of private property – the factors of production and goods and services that people own. This is known as property rights:

Property rights: The social arrangements that govern the protection of private property

Property rights include the rights to physical property as well as to intellectual property, transferability of property rights and protection from violation of property rights.

Economic freedom also requires free markets. However, markets cannot operate without property rights. Both property rights and markets create incentives for people to specialize and trade, to save and invest, to expand their human capital and to discover and apply new technologies. All of the above discussed features are preconditions for persistent economic growth.

On the other hand, for economic growth to be persistent, people must face incentives that encourage them to pursue the following three activities:

- Saving and investment in physical capital
- Expansion of labor force and human capital
- Discovery and application of new technologies

According to our previous discussion, in order to achieve faster economic growth, we must either increase the growth rate of capital per hour of labor or increase the growth rate of human capital or the pace of technological advance. Some possible actions that governments can take to achieve these objectives are:

- Create the incentive mechanisms
- Encourage savings
- Encourage research and development
- Encourage international trade
- Improve the quality of education

As we have already seen, there are many advantages of high economic growth. Let us present the most important ones:

- Higher living standards
- Employment effects: Growth stimulates higher employment
- Fiscal dividend: Growth has a positive effect on government finances – boosting tax revenues and providing the government with extra money to finance public projects
- The investment accelerator effect: Rising demand and output encourages investment in new capital goods – this helps to sustain the growth in the economy by increasing long run aggregate supply
- Growth and business confidence: Economic growth normally has a positive impact on company profits and business confidence – good news for the stock market and also for the growth of small and large businesses alike

However, there are some economic costs of a fast-growing economy. The two main concerns are high inflation and damaging effects on our environment with potentially long-lasting consequences for future generations.

- Inflation risk: If the economy grows too quickly there is a danger of inflation as demand races ahead of aggregate supply. Producer then take advantage of this by raising prices for consumers

- Environmental concerns: Growth cannot be separated from its environmental impact. Fast growth of production and consumption can create negative externalities: increased noise and lower air quality arising from air pollution and road congestion, increased consumption of de-merit goods, rapid growth of household and industrial waste and pollution that comes from increased output in the energy sector. These externalities reduce social welfare and can lead to market failures. Growth that leads to environmental damage can have a negative effect on people's quality of life and may also impede a country's sustainable rate of growth.

13.6 Sustainable Development

It has been only a few decades when there has been much public recognition of the implications of continued population growth and economic growth on the environment.

Environment needs to be regarded as providing not only our economic base in its natural resources but even our very life support system. Now we talk about the problems of pollution and depletion of non-renewable resources such as oil, coal, natural gas and renewable resources such as rain forests as problems that must be faced, since the human race has expanded to the point where it has very significant environmental impacts at a macro level and not only in some local areas.

Pollution becomes a problem when the absorptive capacity of the environment or the so-called environmental carrying capacity is exceeded by pollutants that threaten to alter the environmental balance in ways harmful to humans. As for depletion of natural resources, an economy that is geared to extensive and growing use of virgin raw materials is not indefinitely sustainable on that basis. Actually the prolonged slump and eventual break-up of the Soviet Union is the best example of this since the Soviet economy has heavily relied on extensive use of natural resources.

Therefore, it should be obvious that eventually recycling materials will necessarily become the rule wherever possible as well as energy conservation. Even though our current world is petroleum driven, the oil age will end which calls for utilization of renewable resources like solar energy, wind energy, biological energy, geothermal energy and others.

So, economic growth in its current form puts pressure on our planet if the latter is viewed in terms of not just economic system but also socio-ecological system. This pressure is being recognized by current economists through the concept of sustainable development.

The concept of *sustainable development* has been known to economists for many years. Originally the economic approach to sustainability was based on the Fisher (1906), Hicks (1939) and Lindahl (1933) concept of income as the maximum that we can consume without reducing our wealth. This principle was further generalized and developed in works of Hartwick (1977) and Solow (1986, 1992). In the 1990s, the concept of sustainable development was formalized as the maximization of net benefits of economic and social development subject to maintaining the stock of natural resources over time (Munasinghe, 1993). Although the framework of analysis has changed over the years, its general structure has been preserved by economists as the one that involved an economic criterion reflecting net social benefits plus sustainability constraints.

In its most general form, the concept of sustainable development is about balanced economic growth. Balanced means that economic, social and environmental dimensions of growth should be taken into account simultaneously.

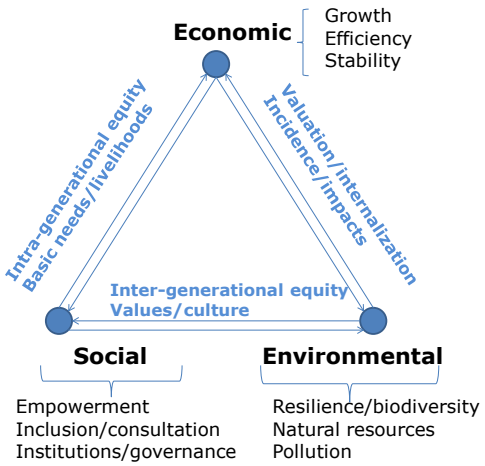
There are many definitions of the sustainable development. Actually John Pezzy in his paper (Pezzy, 1992) published by the World Bank presented more than 30 definitions of the concept. Nonetheless, there exists one definition that is commonly accepted and used. This is the definition presented in the so-called Brundtland Report by the World Commission on Environment and Development in 1987:

Sustainable development: Development that meets the needs of present generation without compromising the ability of future generations to meet their own needs

From this definition we can see that unlike the concept of economic growth discussed earlier, sustainable development has three dimensions:

- Economic
- Social
- Environmental

Graphically this concept can be presented as follows:



Economic dimension is associated with economic stability, growth and efficiency. Social dimension includes problems of poverty, education, unemployment, medical care as intra-generational problems or problems that should be taken into account by current and future generations. Finally, environmental dimension is associated with biodiversity, environmental protection and wise management of renewable and non-renewable resources again given in an intra-generational context.

Economic growth is said to be sustainable if an economy takes all three dimensions into account and achieves a non-decreasing growth of the so-called social welfare function, which is the society's utility function that reflects the society's preferences.

Of course if, for example, population growth exceeds economic growth then such a growth is not sustainable. However, a practice shows that higher educated countries can achieve both, higher growth rates and lower birth rates. Hence it is necessary to help developing countries to achieve higher level of education. Higher level of education leads to a higher technological progress which results in higher productivity. Therefore, the concept of sustainable development justifies financial aid to the developing countries by the World Bank and IMF.

So, the concept of sustainable development is a broader concept than economic growth. Let us summarize the features of sustainable development briefly discussed in this section:

1. Sustainable development is about balanced growth.
2. Balanced growth is viewed in terms of three dimensions – economic, social and environmental.
3. Inter and intra-generational equity is one of the most important goals.
4. In general, three dimensions of sustainable development are its three goals.
5. Sustainable development is a system concept. System approach implies that an object under study is a combination of elements whose collective behavior is essential rather that individual behavior of the system's elements.
6. Since sustainable development is a long-run macroeconomic concept, it is associated with structural changes rather than marginal changes.

As a matter of fact, the world economy in its present state is not sustainable. However, such a situation cannot last forever, and that is why the concept of sustainable development has to replace the concept of pure economic growth.

References

Baumol, W.J. 1977, "On the Proper Cost Tests for Natural Monopoly in a Multiproduct Industry", *American Economic Review*, Vol. 67, No. 5, pp. 809–822

Fisher, I. 1906, *The Nature of Capital and Income*, Macmillan, New York

Fisher, I. 1930, *The Theory of interest*. Philadelphia: Porcupine Press

Hartwick, J.M. 1977, "Intergenerational Equity and Investing the Rents from Exhaustible Resources", *American Economic Review 66*, pp. 9072–9074

Herfindahl, O.C. 1955, "Comment on Rosenbluth's Measures of Concentration" in Stigler, G. (ed.). *Business Concentration and Price Policy*. Princeton, USA: Princeton University Press

Hicks, J.R. 1939, *Value and Capital,* 2nd edn. Oxford University Press, New York

Hirschman, A.O. 1964, "The Paternity of an Index", *American Economic Review*, **54**, pp. 761–762.

Keynes, J.M. 1936, *General Theory of Employment, Interest and Money*, Macmillan Cambridge University Press, for Royal Economic Society

Lee, Dwight R. 1999, "Comparative Advantage Continued", The Freeman: Ideas on Liberty

Lerner, A.P. 1934, "The Concept of Monopoly and the Measurement of Monopoly Power", *The Review of Economic Studies* **1** (3), pp. 157–175

Lindahl, E. 1933, "The Concept of Income", in G. Bagge, *Economic Essays in Honour of Gustav Cassel*, Allen and Unwin

Malthus, T.R. 1826, *An Essay on the Principle of Population: A View of its Past and Present Effects on Human Happiness*; London: John Murray

Munasinghe, M. 1993, *Environmental Economics and Sustainable Development*, World Bank Environmental Paper # 3, Washington, D.C.

Marshall A., *Principles of Economics*, at the Library of Economics and Liberty

Smith, A. 1976, *The Wealth of Nations* edited by R.H. Campbell and A.S. Skinner, The Glasgow edition of the Works and Correspondence of Adam Smith, vol. 2b

Solow, R.M. 1956, "A Contribution to the Theory of Economic Growth". *Quarterly Journal of Economics,* The MIT Press **70** (1), pp. 65–94

Solow, R.M. 1986, "On the Intergenerational Allocation of Natural Resources", *Scandinavian Journal of Economics,* Vol. 88(1), pp. 141–149

Solow, R.M. 1992, An Almost Practical Step Towards Sustainability, Invited lecture, fortieth anniversary of Resources for the Future, Resources and Conservation Center, Washington, DC

Spence, A.M. 1974, *Market Signaling: Informational Transfer in Hiring and Related Screening Processes,* Cambridge: Harvard University Press

Stiglitz, J. 1975, "The Theory of Screening, Education, and the Distribution of Income", *American Economic Review*, Vol. 65, No. 3, pp. 283–300

Swan, T.W. 1956, "Economic Growth and Capital Accumulation". *Economic Record,* John Wiley & Sons **32** (2), pp. 334–361