Competition and monopoly

ASSIGN BUSTER

Introduction

In this chapter we discuss the basic elements of the neoclassical theory of the firm and competition. We begin with the evolution of the notion of competition as a dynamic process of rivalry of firms in their struggle for dominance and continue with the neoclassical notion of competition as an " end state" and we discuss the different types of returns to scale. Sraffa demonstrated that neither the increasing returns to scale nor the decreasing returns to scale are consistent with the assumption of perfect competition in the determination of the supply curve in the industry. The only assumption which is consistent with perfect competition is the case of constant returns to scale, which however leads to implausible results. Pierro Sraffa in his articles (1925, 1926 and 1930) where he concluded that the way out of this conundrum is to side step perfect competition and adopt in its place the notion of monopolistic or imperfect competition. His suggestion was pursued by economists in Cambridge England (mainly J. Robinson and Richard Kahn) during the 1930s. In the same time period in Cambridge-Massachusetts we had the monopolistic competition revolution (mainly E. Chamberlin, J. Bain). These developments in both Cambridges faced the criticism from the economists of Chicago University. Thus, during the 1930s we had a revolution in microeconomic analysis known as " imperefect competition" which was taking place, at the same time, with the macroeconomic revolution of Keynesian economics.

In this microeconomic revolution economists were divided into two camps. The first comprised the proponents of monopolistic competition, who were arguing that the actual economy was characterized by monopolistic

elements that give rise to distortions and who tried to theorize these elements and also correct them by proposing specific antitrust and regulation policies. We shall call these economists, " imperfectionists". On the second camp there were economists mainly from the Chicago University, who claimed on both methodological and empirical grounds that there is no such a thing as " monopolistic" or " oligopoloistic" competition and that the actual economic life is not in any empirically significant deviation from the ideal model of perfect competition. Naturally, this camp of economists may be called " perfectionists".[1] In the ensuing debates, the " perfectionists'" view dominated over the " imperfectionist" one. Fierce as it may have been the debate between the economists in the two camps we recognize that, at the end, they both assumed the importance of perfect competition. The imperfectionists used the perfect competition concept as a yardstick to gauge the extent to which real economic life differs from the perfectly competitive state, while the perfectionists argued that there are no significant differences between the actual and the perfectly competitive economy.

It is ironic, that this process of return to perfect competition begun initially as an attempt to escape from perfect competition through the introduction of realistic elements in the economic analysis of the firm. These efforts led to the development of industrial organization, as an entirely new field of economic research, and to regulation policies that regarded the various market forms as deviations from an ideal model of the perfectly competitive economy, which should be the prototype of actual economic life.

Neoclassical Theory and Perfect Competition

The analysis of competition in the neoclassical theory is contained in the model of perfect competition, which describes the ideal conditions that must hold in the market so as to ensure the existence of perfectly competitive behavior from the typical firm and by extension the characterization of the market or industry as competitive or not. The model of perfect competition describes a market form which consists of a large number of small¾relative to the size of the market¾buyers and of a large number of small producers selling a homogeneous commodity. Both buyers and sellers have perfect information on the prices and the costs of each good. Moreover, there is perfect mobility of the factors of production. The result of the above conditions is that the producers and consumers¾because of their large number and small size¾ are incapable of influencing the price of the product. As a consequence, the price of the product becomes a datum, and the behaviour of the firms is completely passive, that is, firms display a price taking behaviour deciding only the optimal quantity that they will produce. The criterion is the maximization of profit, which is achieved, when the selling price of the good is equal to its marginal cost of production.

The intensity of competition is directly proportional to the number of producers and in general the structure of an industry. In this " quantitative notion of competition", the firm is conceived as the legal entity that hires the services of the factors of production and combines them in order to supply goods in the market. It is important to note that the firm does not own any factors of production; it merely hires the services of the factors of production which are offered by their owners, that is, the individuals. The larger the

number of firms that operate in an industry the more vigorous is their competitive behaviour and by extension we have the establishment of a uniform rate of profit across industries. By contrast, the smaller the number of firms the more oligopolistic and monopolistic is the behavior of the firm in the market and the higher the interindustry profit rate differentials.[2] In this non-competitive state of equilibrium, some prices are above the marginal cost and so society as a whole suffers losses from the underproduction and the underutilization of disposable productive resources. In the neoclassical microeconomic theory, if the firm or the industry displays profits above the normal, for a fairly long period of time, these are attributed to imperfections in the operation of the market and thus in the existence of some degree of monopoly.

We say that firms in perfect competition are price takers, but at the level of general equilibrium, we want to determine the prices which change as a result of the action of some firms. The question, however, is if each and every firm is a price taker, then how do prices change? The usual answer is that prices change exogenously; for example, consumers' preferences change which lead to the increase (or decrease) in demand. In other words, if there is a deficit (or surplus) of the output produced, which is equivalent to saying that all firms face a negatively sloped demand curve meaning that firms in and of themselves cannot increase their price without reducing their market share. In other words, firms in this case operate as if they were in conditions of monopolistic competition. As a consequence, perfect competition exists only in conditions of equilibrium. It is important to stress that perfect competition is a mathematical assumption imposed by

neoclassical economics in order to determine equilibrium and not as a market form that arises from historical observation of the way in which firms are organized and compete with each other.

Similar conclusions are drawn from Walras's conception of attainment of equilibrium through the mediation of the auctioneer. We know that the participants in this model act independently of each other and simply react to the prices announced by the auctioneer, who is supposed to know all the facts. Clearly, if the participants in the Walrasian model act differently then the attainment of equilibrium is problematic. As a consequence, perfect competition is a sine qua non assumption in both Marshallian and Walrasian models of equilibrium. One corollary of the above is that some theories of competition, that were developed in the past, were eventually rejected not for their lack of realism, but precisely because they were out of the analytical framework of neoclassical economics which is oriented towards equilibrium.

In neoclassical economics competition is defined from the way in which technology is being used. More specifically, competition secures that the agents of production (that is firms) will tend to choose the lowest unit cost and price in order to maximize their profits and reduce the market share of their competitors. Thus, competition will combine technology with the behavior of the firms in the market. Unlike classical, neoclassical economists view production not as a process but rather as a result derived from a functional relationship between inputs and outputs. The production functions are assumed to be continuous and differentiable up to the desired degree. The techniques that are used in production are usually assumed as continuous, nevertheless the neoclassical analysis is not affected, if we have

fixed input-output coefficients and L-shaped isoquant curves. Thus, the production functions in neoclassical analysis may take on various forms, such as fixed proportions or the direct opposite of it which is that of perfect substitutability between factors. The assumption of substitutability between inputs is represented with the aid of a concave production function. The proportions between inputs are convex for every single combination of inputs. Hence, we have the already known from the previous chapter isoquant curves, according to which a given level of output can be produced by a variety of input combinations. The curves that we derive are convex to the origin as shown in Figure 1. The negative slope of the isoquant curves represents the diminishing marginal rate of substitution of one factor of production from the other. The isoquants cover the positive quantrant, exactly as in the case of indifference curves, with the difference that the isoquants are measurable, that is, they are amenable of absolute, not only relative, measurement.

As in the case of consumer behaviour, where choices are made at the point of tangency of the highest attainable indifference curve to the income constraint, so in the case of production, the producer chooses the combinations of capital and labor to the point where the isoquant is tangent to the isocost curve, that is the curve C= rK+wL, where r and w are the rewards of the services of capital (K) and labor (L) respectively, and C is the total cost of production. By using the different isocost curves we can form the expansion path, that connects all the points of tangency of isoquants and isocost curves and, therefore, represents the optimal technique in use, that

is, the technique with the minimal cost of production in the case of the different proportions of inputs.

From the above it becomes clear that the givens of the neoclassical theory, that is, the preferences of individuals, the endowments as well as the technology, when combined, impose a type of competition which cannot be different from perfect competition. Firms, that is, the carriers of choice of technique maximize their profits at the point where the value of the marginal product of each and every factor of production is equal to its price. The issue that we will deal with is the level and the composition of output of a firm as well as the method of production. The analysis of the firm bears many similarities with that of the consumer. For example, the isocost curves correspond to the income constraint and the isoquants to the indifference curves.

There are two major differences, between the pure exchange model and that of production. The first is that individuals and not firms own the available resources (endowment). Firms simply hire the services of the factors of production owned by the individuals and through the production process transform them into commodities. The second difference is that the isoquants, unlike isoutility (indifference) curves, are objective, that is, isoquants depend on the level of technology. And technology is not about a free choice (as in the case of individuals) but rather is imposed upon the firms through competition.

Economies of Scale

The role of the firm in the neoclassical theory of production is that of the organization of production process through the hiring of the services of the means of production (which are owned by individuals) and transform them into goods and services and subsequently sell them in the market. In other words, firms organize a process according to which the demands of individuals for goods and services are transformed to respective supplies of goods and services. Firms are viewed as price takers and do not know a priori the price at which they are going to sell their products. The size of the firm is directly proportional to its market share, and therefore, returns to scale are particularly important in determining the level of production of a firm.

It is worth mentioning that the concept of economies of scale as it develops within the neoclassical theory and especially in Marshall (1890, chs. 9-13) is static, that is, it does not arise over time, but rather at a particular moment in time. More specifically, one estimates the level of output in each increase in inputs and according to the answer, the economies of scale are distinguished to the following three categories:

- Increasing returns to scale arise, when inputs are doubled and output increases by more than double.
- Decreasing returns to scale arise, when inputs are doubled and output increases by less than double.
- Constant returns to scale arise when inputs are doubled and output doubles as well.

It is important to stress that the returns to scale imply a change in inputs and a subsequent change in output. In this sense, in the neoclassical analysis the returns to scale are derived from a unified analysis of cost. This is a quite different derivation of the returns to scale of the classical economists, whose analysis is dynamic, and therefore the variables involved are dated and evolve during time. Thus, the case of increasing returns to scale is described in Smith's famous exemplar of a pin factory. The difference from the Marshallian and by extension neoclassical analysis is found in that Smith's economies of scale have a dynamic dimension resulting from the division of labor, which in turn depends on the growth process of the total economy and not on the individual initiatives that are assumed at the level of production units or even at the level of industry. As a result, for the classical economists, economies of scale can only be dynamic and particularly in Smith economies of scale in industry are only increasing.

Decreasing returns to scale in the classical analysis are associated with the theory of rent. For example, Ricardo refers to the law of diminishing productivity of land, a law which is the result of the rising population and the subsequent rising demand for food that forces the cultivation of less productive parcels of land leading to a rising average cost of production. Diminishing returns to scale according to Ricardo are counteracted in part by the technological progress; nevertheless, in the long run the rise in population offsets the technological progress with the net result of the diminishing returns on land. If, however, one does not account for the technological progress and accounts only for the increase in population then we end up with diminishing returns in production, but this result is in

deviation to Ricardo's dynamic analysis. Furthermore, within the static analysis the assumption of diminishing returns to scale is questionable for it presupposes that one of the factors of production is fixed. In fact, when we double the inputs, it is always possible to repeat the production process with the optimal use of resources without reducing the output produced. Consequently, when we refer to diminishing returns to scale, we essentially presuppose that one of the factors of production remains fixed, and therefore as the other factors increase the proportions of inputs that are used differ from the optimal. The question that comes to the fore is; why should firms produce at a range of output associated with diminishing returns when the can produce at the optimal level of output associated with constant returns to scale. In other words, there is no motive what so ever for a firm to move away from the minimum cost of production associated with constant returns to scale and produce at a range of output associated with a higher cost of production and decreasing returns to scale.

Sraffa (1925) pointed out that increasing or decreasing returns to scale in the classical analysis are derived from quite different economic phenomena. Increasing returns, for example, are derived from the process of accumulation and technological change, associated with the division of labour and the extension of the market. Decreasing returns were derived from the limited availability of land, and were an important component of the theory of income distribution, being the foundation of the theory of rent.

The case of constant returns to scale is quite reasonable and is found quite frequently in economic analysis; for example, it is adopted by classical economists and Marx. Marshall on the other hand while he accepts whenever

there is pressure on the raw materials that are being used in industry there is a tendency for rising prices, nevertheless he observes that because the cost of raw materials is only a small fraction of total cost it then follows that they cannot in and of themselves affect the scale of production. Walras in the first edition of his book (1874) also assumed fixed input coefficients and constant returns to scale. In the second edition of his book (1877) he allowed for more substitutability between inputs. Finally, the empirical research has shown that at least in manufacturing the average cost curves have a wide range of output associated with constant returns to scale.

Clearly, Marshall was worried about the case of increasing returns to scale as an assumption that does not fit to the neoclassical static paradigm and this is the main reason that he distinguishes between the economies of scale that are internal to the firm and to those internal to the industry and external to the firm.

Cost Curves

We know from introductory microeconomics that the cost curves of a firm are derived from the production function and the expansion curve (Figure 1b). In the beginning the firm is producing at the falling cost part of the usual U-shaped average cost curve. The shape of these cost curves has to do with the average fixed cost which is supposed to follow a rectangular hyperbola shape which when added to the average variable cost gives rise to the typical U-shaped average cost curves. If we furthermore suppose perfect competition the profit maximizing firm for the particularly given price selects the output at the point where P = MC and in the long-run at the point where

P = d= AR = MR = MC = minAC (see Figure 2), where d is the demand curve faced by the firm, and the other notation is usual.

In the short run we may have P > P*, which means that firms in the industry make excess profits. The result is that firms from other industries are attracted and as the number of firms increases the supply increases and the price of the product falls. If, on the other hand, P < P*, the firms realize losses and so we expect an exit of firms from the industry, a reduction in supply and an increase in price. Finally, we have the case where P= P*, which gives equilibrium, given that the firms that operate simply make normal profits and there are no motives neither for entry of firms from other industries nor for exit of firms that already operate in the industry.

It is important to note that the AC curve has the same shape in both the short run and the long run (Figure 3).[3] In the short run, the average cost curve of the firm is drawn under the assumption of a fixed production capacity. In the long runthe firm has the capacity to change the initial proportions between the factors of production in an effort to achieve their optimal combination. We define the long run average cost of a firm from the points of equilibrium achieved by the firm for different levels of output. We realize that the points of tangency are not the minimum points of the short run average cost (SAC) curves and this can be contemplated theoretically by recalling that the SAC are constructed under the assumption of no optimum use of the available inputs at each output level. In the long run, however, this optimal combination is achieved for the given output. Point E is the minimum cost, which nevertheless is the highest from this which is achieved in the long run if all the productive factors are used optimally. Hence, we

have the well known envelope curve which is attributed to Viner (1931), that is, the long run average cost curve (LAC) is a frontier or an envelope for the short run cost curves. The LAC curve owes its shape to the succession of increasing returns to scale, to the point of constant returns to scale, (corresponding to the optimal firm size) and past this point, to diminishing returns to scale. The plausible question is why this optimal size is not reproduced as the scale of production increases, given that in the long run there is no fixed cost to prevent this from happening. The usual answer is that there are diminishing returns to the entrepreneurship, each firm is run by a president and as the size of the firm increases it becomes more and more difficult for the same person to run effectively the firm.

Let us refer to the long run position of the economy where point ? indicates the optimal combination of all inputs. The size of the firm is determined from the minimum point of the average cost curve which is associated with a given level of production. We claim that the supply curve of the industry is the sum of the supply curves of the firms that form the industry. In other words, the supply curve of the industry is equal to the sum of the marginal cost curves of the firms for levels of output past the minimum point of the average cost curve. A precondition of the above is that we know the exact position of equilibrium of each and every firm, which is characterized as a relation between increasing and decreasing returns to scale.

John Clapham, an economic historian at Cambridge, found the discussion on economies of scale less than satisfactory for he thought there is distance between the theoretical discussion and the economic reality. His article of " empty economic boxes" impressed the economists of the time, because he

pointed out the distance that separates Marshall's theoretical discussion on the economies of scale and the well known shape of the average cost curve and the difficulties of economists in using these ideas in empirical research. More specifically, he argued that we cannot know what percentage of the performance of a firm is attributed to the economies of scale and what percentage to innovations (Clapham, 1922, p. 129). Simply put, Clapham essentially claimed that economists could not ascertain the type of economies to scale. For this reason he characterized the economic theories that could not be demonstrated empirically as " empty economic boxes". Since we cannot discern the type of economies of scale and thus their characterization is an extremely difficult or even an impossible task, then, following this theoretical deficiency, some plausible questions follow; as for example, what kind of measures should governments follow in designing their policies with respect to taxation or the provision of subsidies and incentives in general as components of an economic policy.

In the ensuing debates, it was argued that the incongruence between Marshall's theory of variable returns to scale and empirical observation is solely attributable to the undeveloped nature of statistical analysis and not to any weakness of the theory. We could say that this is the usual response that one gets by applying an empirical critique, which in and of itself could not overturn or create a significant theory. Empirical critique, as it repeatedly has been pointed out, can, at best, ascertain correlations between the variables and not verify causal relations, that is, it cannot derive theoretical relationships between the variables at hand. This does not mean that the empirical critique is redundant. On the contrary, the empirical critique may

enhance our understanding of the underlying relationships between the variables and to reveal relationships hitherto unknown.

Sraffa's Critique of the Marshallian Theory of the Firm

Sraffa's criticism focused on Marshall's hypothesis of returns to scale in production and the assumptions of the competitive firm. The assumption of increasing returns to scale for a large range of output implies that the average cost curve of the firm displays negative slope over a large part of its range and that the marginal cost curve lies always beneath it. Two are the reasons for the decreasing average cost; the first is related to the average fixed cost of the firm which, naturally, as the output expands decreases asymptotically, and thereby, since average fixed cost is a part of average total cost, the total average cost curve tend towards a negative shape. The second reason has to do with the more efficient use of the resources. Between the two reasons only the second is associated with a diminishing marginal cost, whereas the first reason leaves the marginal cost unaffected. With this description of the cost structure, if we assume the case of increasing returns to scale, which are internal for the perfectly competitive firm, then there will be a continuous pressure on the (perfectly) competitive firm to expand its size until its absolute dominance in the market.[4]

In particular, Sraffa argued that in the case of increasing returns to scale, which are internal to the firm, there would be a continuous motive by the firm to expand its production until it can supply the whole market. Clearly, such a hypothesis of returns to scale prima facie contradicts the notion of perfect competition for it leads to monopoly. Marshall had also noticed this inconsistency, for example, the case of increasing returns internal to the firm

that lead to monopoly was detailed by Marshall (1920, p. 666, n. 3) who credited this idea to Cournot and as an act of intellectually honesty, Marshall characterized the increasing returns case as " Cournot's dilemma" (Marshall, 1920, p. 380, n. 1). This is the reason why Sraffa pointed out that the case of increasing returns to scale " was entirely abandoned, as it was seen to be incompatible with competitive conditions" (Sraffa, 1926, pp. 537-8).[5] The only case of increasing economies of scale which is consistent with the requirements of perfect competition is when these economies of scale are external to the firm and internal to the industry, a case, however, which is rarely met in real economies (Sraffa, 1926, p. 540). Furthermore, this type of returns to scale cannot be limited to a single industry, and sooner or later its effects are diffused throughout the economy. The problem in this case is that the Marshallian partial equilibrium framework is inadequate to deal with the complexities emanating from the subsequent development of strong interactions between industries (Sraffa, 1926, pp. 538-9).

The same is true a fortiori with the economies of scale which are external to the firm and to the industry, since the interactions across industries are expected to be much stronger and, therefore, reinforcing the case for abandoning the analysis of partial equilibrium. Turning to the diminishing returns to scale and perfect competition, it follows that since firms buy their inputs in competitive markets they face no restrictions whatsoever in the quantities that they buy and, therefore, there is no reason for the increasing part of the usual U-shaped average cost curves. Hence, the structure of the theory of perfect competition does not allow for the case of increasing cost, as the scale of production increases, simply because there is no mechanism

to force firms to abandon the minimum cost of production and move to higher cost of production.

Consequently, the only assumption that remains is that of constant returns to scale, which give rise to the constant part of the average cost curves (Sraffa, 1926, p. 540). Thus, Sraffa through a critique of the Marshallian theory of the firm was led to a description of the average cost (graphically presented as a line parallel to the horizontal axis) similar to that of the classical economists. This is the reason why he notes:

In normal cases the cost of production of commodities produced competitively [...] must be regarded as constant in respect of small variations in the quantity produced. And so, as a simple way of approaching the problem of competitive value, the old and now obsolete theory which makes it dependent on the cost of production alone appears to hold its ground as the best available (Sraffa, 1926, pp. 540-1).

Hence, Sraffa endorses the theory of value of classical economists, where the price is determined by the cost of production, and not by the intersection of demand and supply curves. More specifically, in the case of perfect competition since the average and marginal cost curves will be identical to each other and since, in equilibrium, the given price (the demand curve) will coincide with the marginal cost (or supply) curve, it follows that equilibrium is not determined uniquely and so the size of the firm is indeterminate.

There are two alternatives out of this conundrum; first, abandon partial equilibrium analysis and adopt the general equilibrium; second, abandon the perfect competition model and adopt monopolistic competition. The first

alternative is the best but it is extremely difficult to pursue in any satisfactory way

[T]he conditions of simultaneous equilibrium in numerous industries: a well-known conception, whose complexity, however, prevents it from bearing fruit, at least in the present state of our knowledge, which does not permit of even much simpler schemata being applied to the study of real conditions. (Sraffa, 1926, p. 542)

Sraffa concluded that the second alternative that is the imperfectly (or monopolistic) competition model might offer a simple and, at the same time, viable solution. In this second one while maintains the partial equilibrium framework and the large number of participants with the difference that their product is differentiated, at least, in the eyes of consumers (Sraffa, 1926, p. 542).

Consumers' preferences do not easily change, because they are determined by factors, such as the marketing of the product, the personal acquaintance and the loyalty of customers to a specific firm that last for long. Thus, he proposed the replacement of the assumption of perfect competition by that of monopoly:

It is necessary, therefore, to abandon the path of free competition and turn in the opposite direction, namely, towards monopoly (Sraffa, 1926, p. 542).

In short, the theory of firm cannot be built on the assumption of perfect competition, because in actual competition firms cannot sell any quantity

they produce at a given price. The production is not limited by cost, but rather by demand.

The initial reaction of neoclassical economists was to assume certain fixed characteristics in the operation of the firm that give rise to diminishing returns to scale. Thus, they argued that entrepreneurship is a characteristic which does not increase with the size of the firm and so there will be diminishing returns to this factor of production.[6] The logical consequence of this argument according to Kaldor is that we are led to the idea that the optimal size of the firm is determined by the working time of the entrepreneur, in other words we have one entrepreneur firms. Another way to address Sraffa's critique was to assume general equilibrium where entrepreneurial talents not only are