1

Introduction to Micro-economics

This book is about the economics of electricity markets. It is therefore essential that the reader understands a number of basic concepts in economics. Much of the material here can be found in introductory textbooks in economics. However, we hope that setting out this material at the start of this textbook will assist readers who do not have a background in economics.

Readers who have a background in economics may choose to skip this part. However, this presentation probably contains some new material, even for readers familiar with economics. In addition, we introduce notation and a few key ideas which are used throughout the rest of the book. We recommend at least a review of this material.

1.1 Economic Objectives

Economics is the study of the production, consumption, and exchange of goods and services in an economy – including how production, consumption, and exchange are organised, how information flows and how participants are rewarded and incentivised for playing their part. Economics seeks to both create theories which explain the patterns of behaviour and organisation that we see in the real world (so-called positive theories), and to develop policies and proposals for changing the arrangements that exist in the real world (normative theories).

But, if we are to recommend changes to existing arrangements, we need a commonly agreed set of objectives that we are trying to achieve. In our view, this common set of objectives must relate, in some way, to a common vision of the overall *economic welfare* of the society or economy as a whole.

There may be many different ways of articulating the overall economic welfare of a society or economy, if such a thing exists at all. It may never be possible to get consensus over whether or not some alternative state of the world, B, is preferred to the status quo, A. Economists tend to focus on areas where, in principle, there could be consensus – that is, situations where, in principle, every member of society could agree that B is preferred over A. These tend to be situations where there is what might be described as waste or inefficiency – where we could reorganise things so that we could achieve the same outcomes with fewer resources, or achieve better outcomes with the same resources. If these situations exist we could, in principle, leave everyone better off.

Although there is some variation in economic theory, in practice most public policy economists make certain assumptions which simplify the task of determining the total

The Economics of Electricity Markets, First Edition. Darryl R. Biggar and Mohammad Reza Hesamzadeh.

economic welfare. Chief amongst these is the assumption that we can ignore income effects. In effect, this means that the benefit of an additional dollar to me is about the same as the benefit to any other member of the society. This assumption rules out the possibility of deriving any benefit from income redistribution alone. Alternatively, we can imagine that such income redistribution has already occurred through some other mechanism.

If we make this assumption, for any public policy change we can envisage, we can value the benefits and the costs imposed using a simple monetary metric. The change in economic welfare brought about by the public policy change is the simple sum of the monetised benefits and costs. An arrangement which maximises the economic benefits less the costs is said to be *efficient*. This is the usual meaning of the term economic efficiency in public policy analysis.

This notion of economic efficiency does not incorporate everything which the broader public might consider important. In particular, it does not usually directly deal with controversial questions about how income should be distributed in the economy. Neither does it normally directly address questions of fairness or equity.[1] Nevertheless, this notion of economic efficiency captures important, and broadly acceptable, notions of social welfare, and for most economists represents a legitimate goal for economic policymakers.

It is valuable to break down this notion of efficiency further. It is useful and helpful to distinguish between *short-run* and *long-run* concepts of efficiency. In the short-run we have an existing stock of assets in place. Short-run efficiency relates to getting the most out of the existing stock of assets: producing as much as possible and allocating those goods and services to those customers which value them most highly. In the longer run we can change the stock of assets, creating new assets or removing old assets. Longer-run efficiency includes the notion of efficiency in changing the stock of assets over time. We can distinguish between both *production assets* (used to make other goods or services, including electricity) and *consumption assets* (which are used to directly provide services to customers, such as electrical appliances or electrical machinery).

Specifically, we can distinguish between the following:

a. Efficiency in the *use* or *operation* of an existing stock of assets. This includes efficiency in the allocation of goods and services (ensuring that goods and services are consumed by those who value them most highly) and efficiency in the production of goods and services (ensuring that goods and services are produced at the lowest possible cost, given the existing stock of assets).
b. Efficiency in investment in the *creation* of new assets (or the disposal of old assets), including investment by seller(s) in new production assets (of the right size, in the right location, in the right amount, of the right type, and so on) and investment in developing new goods and services, and investment by buyer(s) in assets which increase their value for the goods or services produced (new consumption assets).

Many textbooks distinguish between allocative, productive and dynamic efficiency. These terms are defined in different ways by different economists. We prefer the following definition:

Allocative and productive efficiency are short-run efficiency concepts, relating to the efficient use of the existing stock of assets. *Allocative efficiency* refers to ensuring that the goods and

[1] This does not mean to imply that economic policies will not accord with principles of fairness or equity – but rather that those terms are interpreted in an economic way where they are addressed at all.

services produced are allocated to those who value them most highly. *Productive efficiency* refers to ensuring that goods and services are produced at the lowest possible cost. In contrast, dynamic efficiency is a longer-run concept, relating to changes in the existing stock of assets. *Dynamic efficiency* refers to efficient decisions regarding investment in new assets (what, where, when, and what type of investment), including investment in developing new products and services over time.

> *Result*: Economic efficiency has both a short-run and long-run dimension. In the short run, economic efficiency is about the efficient use of a given set of production and consumption assets (productive and allocative efficiency). In the longer run, economic efficiency is about efficient decisions in the creation of new assets or the disposal of old assets (dynamic efficiency).

This book is about the design of arrangements to achieve these economic efficiency objectives in the electricity industry. Any particular arrangement is only desirable to the extent it achieves these objectives. In particular, this book will explore the extent to which competitive markets in the electricity sector can achieve the objectives above. We will see that in many situations, competitive markets deliver economically efficient outcomes. In other situations, competitive markets will not achieve these outcomes and we must substitute alternative arrangements, such as direct price controls. Particular institutional arrangements, such as competitive markets, are not an end in themselves. They are only the *means to an end* – a means to the achievement of the objectives set out above.

In many parts of the text that follow we will first seek to determine the efficient outcome (that is, the efficient use/operation of existing assets and/or the efficient investment outcomes) and then seek to determine whether particular market arrangements can achieve those outcomes, and under what conditions the market arrangements might achieve those outcomes.

When it comes to achieving efficient outcomes using a given stock of assets, economists typically focus separately on the buying (or demand) side of the market and the selling (or supply) side of the market. The next two sections focus in turn on the buying side of a market and the supply side of a market. We will then bring these ideas together to look at what it means to achieve efficiency in the use of a given stock of assets. In the subsequent section we will explore whether or not short-run efficient outcome can be achieved using a competitive market. First, however, we review the principles of constrained optimisation.

1.2 Introduction to Constrained Optimisation

Optimisation lies at the heart of economics. Under conventional economic theory all economic actors are assumed to be maximisers of an objective function. This underlying assumption is made even more explicit in the smart markets introduced in Section 1.6. It is therefore essential for all students of electricity markets to have some understanding of the theory of constrained optimisation.

In this text we will often see a constrained optimisation problem expressed in the following form:

$$\max f(x)$$

Subject to the following conditions:

a. For $i = 1, \ldots, n, \quad g_i(x) = c_i \leftrightarrow \lambda_i$
b. For $j = 1, \ldots, m, \quad h_j(x) \leq d_j \leftrightarrow \mu_j$

Here x is a vector of k variables, and the functions $f(\cdot), g_i(\cdot), h_j(\cdot)$ are all functions from $\mathbb{R}^k$ to $\mathbb{R}$. The equations $g_i(x) = c_i$ and $h_j(x) = d_j$ are known as constraint equations.

The variables λ_i and μ_j are known as *Lagrange multipliers* and their value will become apparent shortly. Each constraint equation has its own Lagrange multiplier. Here we are following the convention which uses the symbol $\leftrightarrow$ to show the Lagrange multiplier which is associated with each constraint equation.

Let us suppose that we have a set of values x, λ, μ which satisfy the following conditions, known as the *Karush–Kuhn–Tucker* (or KKT) conditions. Then x is a solution to the constrained optimisation problem above.

The KKT conditions are as follows:

1. For $l = 1, \ldots, k$,

$$\frac{\partial f}{\partial x_l} - \sum_i \lambda_i \frac{\partial g_i}{\partial x_l} - \sum_j \mu_j \frac{\partial h_j}{\partial x_l} = 0$$

This condition is known as the First Order Condition

2. For $i = 1, \ldots, n, \quad g_i(x) = c_i$
3. For $j = 1, \ldots, m, \quad \mu_j \geq 0, \quad$ and $\quad h_j(x) \leq d_j$ and $\mu_j(h_j(x) - d_j) = 0$

In other words, the problem of finding a solution to the constrained optimisation problem above reduces to the problem of finding a solution to the KKT conditions (1)–(3).

It is worth noting that the Lagrange multipliers have a particular interpretation. The Lagrange multipliers measure the extent to which the objective function can be improved following a small change in the constraints. For example, let us define $\mathcal{L}$ to be the value of the objective function at the solution of the constrained optimisation above. Then the Lagrange multiplier λ_i is the change in the objective function with respect to a small change in the parameter c_i. Similarly, the Lagrange multiplier μ_j is the change in the objective function with respect to a small change in the parameter d_j.

$$\lambda_i = \frac{\partial \mathcal{L}}{\partial c_i} \quad \text{and} \quad \mu_j = \frac{\partial \mathcal{L}}{\partial d_j}$$

1.3 Demand and Consumers' Surplus

Let us focus more closely on the buying (or demand) side of a market for a particular good or service, such as electricity. We will focus on an abstract buyer or customer of this service. Although we will use the word customer, we do not intend to limit ourselves to small customers or consumers. Rather this customer could be a large business, such as an aluminium smelter, a small business, such as an office or restaurant, or a residential household.

To model the behaviour of customers in a market, in principle we need to specify two things: (a) the range of actions or choices that the customer faces; and (b) some form of objective which the customer is seeking to pursue.

1.3.1 The Short-Run Decision of the Customer

In principle, customers can take a range of actions which affect the value they receive from a good or service. This is particularly true in the case of electricity. Customers do not consume electricity directly; instead they consume the services of a range of machinery, pumps, heaters, devices and appliances which consume electricity. The demand for electricity at any one point in time depends on the stock of past investments made by the customer. More generally, the demand for a particular good or service in the economy will depend on the past actions taken by customers.

For the moment we will focus on short-run decisions of the customer. Let us assume that the customer has made a set of decisions in the past regarding devices which consume electricity. The key remaining decision of the customer is how much electricity to consume at a given point in time – or more precisely, the rate at which electricity is consumed.[2]

1.3.2 The Value or Utility Function

In order to complete the model of customer behaviour, we need to specify the customer's objective. Many introductory textbooks in economics start by introducing the demand curve. However, we will follow a slightly unconventional path and start with the notion of a *value or utility function*. This approach is straightforward and allows us to draw simple parallels between the demand and the supply sides of each market.

Let us suppose we have a customer which is consuming a particular good or service at the rate Q (units per interval of time). Let us assume that we can express the utility or value (also known as surplus) that this customer receives from consuming this particular good or service in the form of a function, known as a utility function $U(Q)$ ($ per interval of time). (This customer could itself be a firm, in which case the utility is equal to the profit of the firm from the activity).

This utility will depend on a number of factors, such as the investments the customer has made in equipment which uses the good or service in question, or, if the customer is itself a firm, the demand for the final product produced by the firm and the substitutes available for the good or service in question. Typically, the customer is assumed to obtain higher utility from consuming at a higher rate. In other words, $U'(Q) > 0$ (here the prime symbol signifies the first derivative of the utility function with respect to the rate of consumption). Also, by assumption the rate at which value increases with consumption decreases the higher the rate of consumption (i.e. $U''(Q) < 0$).

In practice, a customer will typically not consume just a single good or service, but several different goods and services at the same time. The utility function can be a function of the rate of consumption of each of these goods and services. For example, if a customer consumes two goods, at the rates Q_1 and Q_2, the rate at which the customer receives value or utility might be denoted by $U(Q_1, Q_2)$ ($/interval of time).

1.3.3 The Demand Curve for a Price-Taking Customer Facing a Simple Price

Let us suppose that the customer obtains a particular good or service through arm's length transactions in a market. The simplest assumption we can make is that the customer pays a

[2] Strictly speaking, the customer does not directly choose the rate of consumption of electricity – instead he/she chooses the rate at which to enjoy the services for which electricity is used (such as the rate of manufacture of aluminium), and the rate of consumption of electricity follows.

simple constant price P ($ per unit) for this good or service independent of the rate that he/she consumes. Many goods and services have more complicated pricing schemes, but for the moment, it is convenient to assume that each customer pays a simple constant price. As a consequence, if the customer consumes the good or service at the rate Q (units/interval of time), the customer must make a payment equal to PQ ($/interval of time) each time period.

Let us assume that the customer is a *price-taker* – that is, the customer has no influence over the market price, regardless of how much he/she consumes. The customer is assumed to choose a rate of consumption which maximises his/her net surplus or net utility – that is the utility from consumption less the revenue paid to obtain the good or service. In other words, the customer is assumed to maximise the following expression:

$$\varphi(Q) = U(Q) - PQ$$

What rate of consumption maximises the net utility? The first-order condition for the maximum is as follows:

$$\frac{\mathrm{d}\varphi}{\mathrm{d}Q} = U'(Q) - P = 0$$

which implies that the optimal rate of consumption is where the marginal utility is equal to the price:

$$U'(Q) = P$$

The first derivative of the utility function $U'(Q)$ ($ per unit) is known as the *inverse demand curve* and will be denoted by $P^{\mathrm{D}}(Q)$. The inverse demand curve shows, for each rate of consumption, the corresponding price that a price-taking customer is prepared to pay to sustain that rate of consumption. The inverse demand curve is downward sloping: An increase in the market price corresponds to a lower rate of consumption, and vice versa.

The result above shows that a price-taking customer (that is a customer who cannot influence the market price) will choose to consume at a rate where the inverse demand curve is equal to the market price. This is illustrated in Figure 1.1.

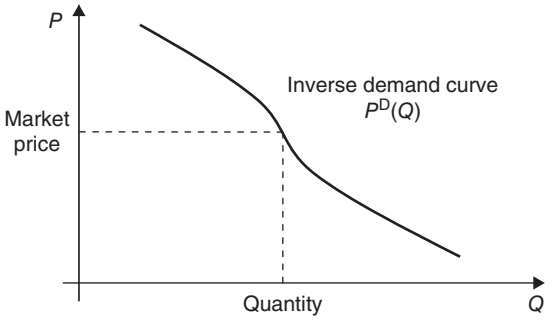

Figure 1.1 A typical inverse demand curve

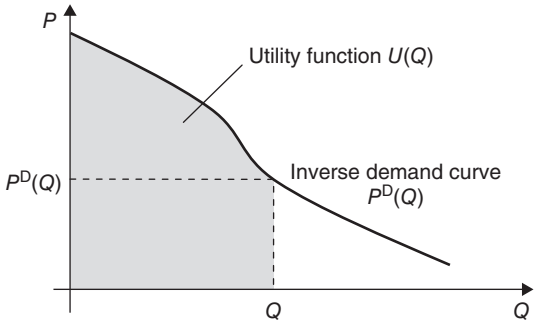

Figure 1.2 The utility function is the area under the inverse demand curve

Result: For a price-taking consumer facing a simple linear price, the consumer's demand function is given by (the downward sloping part of) the marginal utility curve.

Because the inverse demand function is downward sloping, it has an inverse, which is known as the demand curve and will be denoted here by $Q^{\mathrm{D}}(P)$. The demand curve shows, for a given level of the market price, the rate at which this customer is willing to consume.

Given the inverse demand curve for a good or service, we can work out the corresponding utility function, and vice versa. Since the inverse demand curve is the first derivative of the utility function, the utility function is found by integrating the inverse demand curve. This gives the level of the utility function up to some constant (here denoted by c):

$$U(Q) = \int^{Q} P^{\mathrm{D}}(Q)\mathrm{d}Q + c$$

This has been illustrated diagrammatically in Figure 1.2. The utility or value function (up to a constant) is equal to the area under the (inverse) demand curve:

Where there are many different customers, all consuming the same good or service at different rates $Q_1, Q_2, \ldots$, the total value function for the market as a whole from consuming this particular good or service is just the sum of the value functions of each customer. This is also known as the total or gross *consumers' surplus*.

$$CS(Q_1, Q_2, \ldots) = U(Q_1, Q_2, \ldots) = \sum_i U_i(Q_i)$$

If we have a fixed total rate of consumption of the good or service available, Q, say, how should that rate of consumption be allocated between customers to achieve the highest possible overall value or surplus? This problem can be written mathematically as follows:

$$\max_{Q_1, Q_2, \ldots} U(Q_1, Q_2, \ldots) \text{ subject to } \sum_i Q_i = Q$$

This is a constrained optimisation problem. We can solve this problem by setting out the Lagrangian and computing the KKT conditions as explained in the previous section. The first-order condition for this problem is

$$U_i'(Q_i) = P_i^{\mathrm{D}}(Q_i) = \lambda$$

where λ is a Lagrange multiplier.

In other words, the problem of efficiently allocating a fixed total rate of consumption of a given good or service between any number of customers can be solved by setting a market price λ for the good or service and allowing each customer to buy at whatever rate he/she desires at that market price. The market price should be chosen in such a way that the total demand (i.e. the total rate of consumption) at that price is equal to the total rate of consumption required:

$$\sum_i Q_i^{\mathrm{D}}(\lambda) = Q$$

We are starting to see how competitive markets can solve allocation problems. As we will discuss in more detail later, if all participants are price-takers, the market solves the problem of efficiently allocating a given rate of consumption of a good or service between customers with different needs and preferences.

This analysis has focused on the case where the customer consumes a single good or service. In practice, of course, a customer will typically consume several goods or services at one time. The value that a customer places on any one good will typically depend on the rate at which he/she is consuming another good. This complicates the analysis above a little bit. We can distinguish two extreme cases, where the utility function depends only on the total rate of consumption of the two goods:

$$U(Q_{\mathrm{A}}, Q_{\mathrm{B}}) = U(Q_{\mathrm{A}} + Q_{\mathrm{B}})$$

In this case the two goods are said to be perfect substitutes and can be treated as though they are really one good in practice. The other extreme case is where the utility function can be separated into two separate functions:

$$U(Q_{\mathrm{A}}, Q_{\mathrm{B}}) = U_{\mathrm{A}}(Q_{\mathrm{A}}) + U_{\mathrm{B}}(Q_{\mathrm{B}})$$

In this case, the two goods are independent of each other.

1.4 Supply and Producers' Surplus

Now let us focus on the selling or supply side of a market for a particular good or service. Without loss of generality we can assume that goods and services are produced by an economic entity which we will refer to as a *firm*. A firm purchases certain goods and services, known as *inputs*, and converts them into different goods and services, known as *outputs*.

As with the customer side of the market, in order to model the behaviour of a firm we need to know something about its possible range of actions, and something about its objective. In the longer run, firms can take a wide range of actions, such as investing in new production capacity, marketing their goods and services, and investing in research and development for new

products. But, for the moment, let us follow the approach set out in the previous section: let us take the stock of existing investments of a firm as given, and focus on the short-run decisions of the firm – primarily the decision as to the rate of production.

In the course of production a firm will incur expenditure on inputs. Some of the expenditure of a firm will take the form of investments to increase the productive capacity, or the demand for the services of the firm. Such expenditures, once made, do not vary with the rate of production of the firm at any given point in time. Let us focus for the moment on those expenditures which are directly related to the rate of production at a point in time, known as the *variable costs*. This might include the labour costs of staff involved in production, or the cost of purchasing inputs which vary with production.

More generally, when discussing the costs or expenditure of a firm we must be clear about the time frame we have in mind. In the short run, when a firm has already made sunk investments in buildings, equipment and so on, the managers of the firm need only be concerned with the expenditure which they can alter in the short run by altering the rate of production of the firm. In the longer run, the firm may be able to alter its size or scale, or change its location.

1.4.1 The Cost Function

Let us assume that the firm produces just a single good or service. The rate of production of this good or service is denoted by Q (units/interval of time). The rate at which expenditure is incurred to produce at a given rate is known as the *cost function* and will be denoted by $C(Q)$ (\$/interval of time). Typically we will assume that higher rates of production correspond to higher costs (i.e. $C'(Q) > 0$). In addition, for most firms the rate at which cost increases with the rate of production itself increases as the rate of production increases (i.e. $C''(Q) > 0$).

We will often be interested in how costs change with a small increase in the rate at which output is produced, which is known as the *marginal cost*. The marginal cost function is the slope of the cost function and will be denoted by $MC(Q)$.

$$MC(Q) = \frac{\mathrm{d}C}{\mathrm{d}Q} = C'(Q)$$

By the assumptions above, the marginal cost function is positive and upward sloping.

Most firms produce not just a single good or service, but many hundreds or thousands of different goods and services. The costs of the firm can be a function of the rate of production of each of these goods and services. For example, if the firm produces two goods, at the rates Q_1 and Q_2, the rate at which costs are incurred can be denoted by $C(Q_1, Q_2)$ (\$/interval of time). There is a separate marginal cost for each good or service, which is the partial derivative of the cost function with respect to the corresponding rate of production:

$$MC_1(Q_1, Q_2) = \frac{\partial C}{\partial Q_1}$$

1.4.2 The Supply Curve for a Price-Taking Firm Facing a Simple Price

Let us suppose the firm sells its output in a conventional market. Let us take the simplest case and suppose that the firm obtains the same price P (\$ per unit) for this good or service for each

unit, independent of the rate at which it produces. As a consequence, if the firm produces the good or service at the rate Q(units/interval of time), the producer receives a flow of funds at the rate PQ (\$/interval of time).

Let us consider the case where the firm is a price-taker – that is where the firm cannot influence the market price by varying his/her output. In economics it is conventional to assume that the short-run objective of a firm is to maximise its *profits*. Profits are conventionally defined as the revenue the firm receives from sales of outputs in a period less the expenditure incurred by the firm in the purchase of inputs. In other words, the firm is assumed to maximise:

$$\pi(Q) = PQ - C(Q)$$

This is an adequate statement of the short-run objective of most firms. In the case of longer-run decisions (such as decisions regarding investment in assets which last multiple periods), the investment decision will change the cash flow of the firm not just in a single period but over many periods into the future. In this case it makes more sense to assume that the firm maximises the *present value* of the stream of profits (the determination of the value of a stream of cash flows, especially under conditions of uncertainty, takes us beyond the scope of this text).

Even if we limit ourselves to short-run decisions, not all economic firms will choose to maximise profits. In particular, some government-owned firms pursue a range of broader objectives. However, for many purposes, especially for short-run decisions, it is reasonable to make the assumption that firms seek to maximise profits.

The rate of production which maximises the rate at which the firm receives profits satisfies the following first-order condition:

$$\frac{\mathrm{d}\pi}{\mathrm{d}Q} = P - C'(Q) = 0$$

We find that, for each price, a profit-maximising price-taking firm chooses the rate of production where the marginal cost curve is equal to that price.

$$C'(Q) = P$$

The *supply curve* for a firm (which we will denote as $P^{\mathrm{S}}(Q)$) shows, for each level of market price, the rate of production which the firm will choose. We have shown that a profit-maximising price-taking firm will choose to produce where its marginal cost is just equal to the market price. In other words, the (upward sloping) part of the marginal cost curve is the supply curve for the firm.

> *Result*: For a price-taking producer facing a simple linear price the supply function is given by (the upward sloping part of) the marginal cost curve.

Under the assumptions set out above, the marginal cost curve (and therefore the supply curve for a firm) is upward sloping (Figure 1.3).

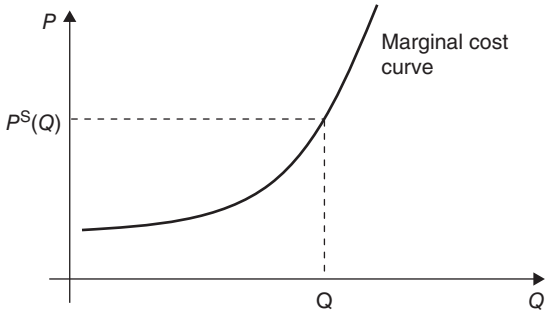

Figure 1.3 The marginal cost curve is the supply curve for a competitive firm

Given the supply curve for a price-taking firm, we can work out the corresponding cost function, and vice versa (in exactly the same way we showed earlier for the value function and the demand curve). Since the supply curve is the first derivative of the marginal cost curve, the cost function is found by integrating the marginal cost curve. This gives the level of the value function up to some constant (here denoted by c):

$$C(Q) = \int^{Q} P^{S}(Q)\mathrm{d}Q + c$$

This has been diagrammatically shown in Figure 1.4. The cost function (up to a constant) is equal to the area under the supply curve:

Where there are many different firms, all producing the same good or service at different rates $Q_1, Q_2, \ldots$, the total cost of production (\$/interval of time) for the market as a whole is just the sum of the cost functions of each producer:

$$C(Q_1, Q_2, \ldots) = \sum_i C_i(Q_i)$$

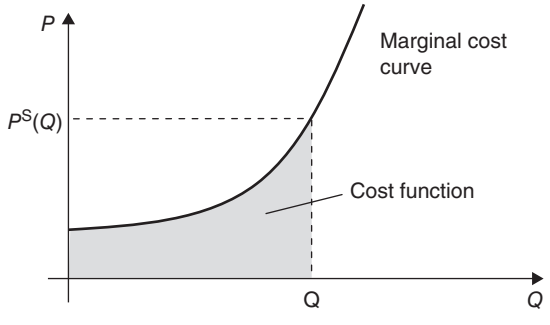

Figure 1.4 The cost function is the area under the marginal cost curve

Let us define the gross producers' surplus to be the negative of the total cost of production for the market as a whole:

$$\mathrm{PS}(Q_1, Q_2, \ldots) = -C(Q_1, Q_2, \ldots)$$

If we want to produce a given good or service at a given total rate, Q, say, how should we allocate this total rate of production across different producers to end up with the overall lowest-cost rate of production? This problem can be written mathematically as follows:

$$\max_{Q_1,Q_2,\ldots} \mathrm{PS}(Q_1, Q_2, \ldots) = \min_{Q_1,Q_2,\ldots} C(Q_1, Q_2, \ldots) \text{ subject to } \sum_i Q_i = Q$$

The first-order condition for this problem is

$$C_i'(Q_i) = P_i^{\mathrm{S}}(Q_i) = \lambda$$

where λ is the Lagrange multiplier. In other words, the problem of efficiently allocating a given rate of production between a number of different producers with different costs can be solved by setting a market price λ for the goods or services and allowing each producer to sell as much as he/she desires at that market price. The market price should be chosen in such a way that the total supply (i.e. the total rate of production) at that price is equal to the desired rate of production.

Here we see how competitive markets, in which all the participants are price-takers, help solve the problem of producing a given total rate of production at the lowest possible ongoing costs.

The analysis above focuses on the case of a firm which produces a single output. Most real-world firms produce hundreds or thousands of different goods or services. As in the previous section, we can extend the analysis above to a firm which produces two or more goods or services at different rates. The cost function of the firm may now depend on the rate at which the firm produces two or more outputs.

1.5 Achieving Optimal Short-Run Outcomes Using Competitive Markets

Let us suppose we have a market with a large number of buyers (customers) and a large number of sellers (firms or producers). Let us suppose that both the sellers and the buyers have made past sunk investment in assets for producing or consuming the good or service in question. Given these investments, there is a remaining question as to the rate at which each seller or buyer should produce or consume so as to maximise overall economic welfare in the short run.

We are interested here, and throughout this book, in two questions: (i) What is the optimal (welfare-maximising) outcome? (ii) Can this optimal outcome be achieved using competitive markets?

1.5.1 The Short-Run Welfare Maximum

Let us focus on the first question: What is the welfare-maximising outcome?

Let us suppose that we have a set of producers. Producer i produces at the rate Q_i^{S} and has the cost function $C_i\left(Q_i^{\mathrm{S}}\right)$. In addition, let us suppose that producer i faces some generic constraints on his/her rate of production which we will denote by $g_i\left(Q_i^{\mathrm{S}}\right) \leq 0$.

Similarly, let us suppose we have a set of consumers. Consumer i produces at the rate Q_i^{B} and has the utility function $U\left(Q_i^{\mathrm{B}}\right)$. In addition, consumer i faces some generic constraints on his/her rate of consumption which we will denote by $h_i\left(Q_i^{\mathrm{B}}\right) \leq 0$.

The welfare-maximisation problem is to find a rate at which each producer should produce, Q_i^{S} and a rate at which each customer should consume, Q_i^{B} which maximises the total economic surplus, subject to the condition that the total rate of production must equal the total rate of consumption:

$$\begin{aligned}
&\max \mathrm{TS} \\
&\quad = \max \mathrm{CS}\left(Q_1^{\mathrm{B}}, Q_2^{\mathrm{B}}, \ldots\right) + \mathrm{PS}\left(Q_1^{\mathrm{S}}, Q_2^{\mathrm{S}}, \ldots\right) \\
&\quad = \max \textstyle\sum_i U_i\left(Q_i^{\mathrm{B}}\right) - \sum_i C_i\left(Q_i^{\mathrm{S}}\right)
\end{aligned}$$

$$\begin{aligned}
\text{Subject to} \quad &\text{(a)} \ \textstyle\sum_i Q_i^{\mathrm{B}} = \sum_i Q_i^{\mathrm{S}} \leftrightarrow \lambda \\
&\text{(b)} \ g_i\left(Q_i^{\mathrm{S}}\right) \leq 0 \leftrightarrow \alpha_i \\
&\text{(c)} \ h_i\left(Q_i^{\mathrm{B}}\right) \leq 0 \leftrightarrow \beta_i
\end{aligned}$$

From the KKT conditions for this problem, at the optimum, the following conditions must hold:

i. For each consumer: $U_i'\left(Q_i^{\mathrm{B}}\right) - \lambda - \beta_i \dfrac{\partial h_i}{\partial Q_i^{\mathrm{B}}} = 0$

ii. For each producer: $C_i'\left(Q_i^{\mathrm{S}}\right) - \lambda + \alpha_i \dfrac{\partial g_i}{\partial Q_i^{\mathrm{S}}} = 0$

In the case where there are no other constraints on the rate of consumption of producers and consumers, we have the result that, at the optimum, the marginal utility of all customers and the marginal cost of all producers are the same: $U_i'\left(Q_i^{\mathrm{B}}\right) = \lambda = C_i'\left(Q_i^{\mathrm{S}}\right)$.

1.5.2 An Autonomous Market Process

Although a few markets in a modern economy are organised around a central market operator or auctioneer, most markets in an economy operate autonomously without a central auctioneer or market-maker. There are various ways in which this market process could operate. The key question we would like to consider is whether or not a competitive market process can lead to the efficient outcome described above.

Let us suppose that, by some mechanism a single market price is determined. Let us assume that there are a large number of buyers and a large number of sellers, so that no buyer or seller has any influence over that market price. Each buyer or seller is assumed to be free to choose its own rate of production or consumption given the market price.

Let us focus first on the task of a producer. As before we will assume the producer seeks to maximise his/her profit, given the market price and any constraints on the rate of production. The task of the producer is, therefore, to solve the following problem:

$$\begin{aligned}
&\max \pi_i\left(Q_i^{\mathrm{S}}\right) = PQ_i^{\mathrm{S}} - C_i\left(Q_i^{\mathrm{S}}\right) \\
&\text{Subject to :} \ g_i\left(Q_i^{\mathrm{S}}\right) \leq 0 \leftrightarrow \alpha_i
\end{aligned}$$

Similarly, the task of a consumer is to solve the following problem:

$$\min U_i\left(Q_i^{\mathrm{B}}\right) - PQ_i^{\mathrm{B}}$$

$$\text{Subject to : } h_i\left(Q_i^{\mathrm{B}}\right) \leq 0 \leftrightarrow \beta_i$$

The KKT conditions for these problems are as follows:

i. For each consumer, $U_i'\left(Q_i^{\mathrm{B}}\right) - P - \beta_i \dfrac{\partial h_i}{\partial Q_i^{\mathrm{B}}} = 0$

ii. For each producer, $P - C_i'\left(Q_i^{\mathrm{S}}\right) - \alpha_i \dfrac{\partial g_i}{\partial Q_i^{\mathrm{S}}} = 0$

Comparing these conditions with the conditions above, we can draw the following conclusion: Provided the market price P is equal to the constraint marginal value of the overall supply–demand balance constraint (which we have labelled λ) the decentralised profit-maximising decisions of producers and the utility-maximising decisions of consumers will achieve the welfare-maximising combination of production and consumption.

All that remains to show is that the market price can (and will) be set equal to the constraint marginal value on the overall supply–demand balance constraint.

In a typical market, this equilibration of supply and demand occurs through an adjustment process, through adjustments to the market price. If the market price is too high, the rate of production of the good or service will exceed the rate of consumption – inventories or stocks of the good or service will build up, or some producers will not be able to sell all they want at the price. In either case, some producers will cut their price, resulting in a fall in the market price. Conversely, if the market price is too low, the rate of production of the good or service will fall short of the rate of consumption. Inventories will decline and/or some customers will not be able to purchase at the rate they desire. In either case, customers will bid up the price. The equilibrium market price is where the total rate of production is equal to the total rate of consumption.

In other words, in a conventional, competitive economic market, where buyers and sellers are price-takers, actions taken by individual buyers and sellers, choosing their own short-term rate of production to maximise their own objectives, achieve an overall short-run welfare maximum. This is one of the key reasons why economists like competitive markets – they are a mechanism by which individuals, pursuing their own ends, can be coordinated 'as if by an invisible hand' to achieve outcomes which are efficient for the economy as a whole.

Result: Economic welfare-maximising outcomes can be achieved with a decentralised market process. Provided the market price is equal to the constraint marginal value of the overall supply–demand balance constraint in the welfare-maximisation problem, and provided all producers and consumers are price-takers, the decentralised profit-maximising decisions of producers and the utility-maximising decisions of consumers will achieve the welfare-maximising combination of production and consumption.

1.6 Smart Markets

Most markets in a modern economy operate entirely autonomously, without any form of central role for a market operator or auctioneer. However, for reasons which we will see later, wholesale electricity markets are not like this. Due to the complexities of the way that electric power flows out of electricity networks, it is not possible to separate the market for the transportation of electric power from the market for the production or consumption of power – these two markets must be integrated through the central role of the market operator.

So-called *smart markets* were developed to incorporate more sophisticated physical constraints into market processes. Smart markets are used in the allocation of take-off and landing slots at airports, in the allocation of radio spectrum and in natural gas markets. Most importantly, smart markets are used in the electricity industry to integrate physical network constraints with the trading of electric power.

1.6.1 Smart Markets and Generic Constraints

Mathematically, a smart market involves an extension to the problem discussed in the previous section. Let us suppose that, as before, we seek a combination of production and consumption which maximises economic welfare, subject to (a) the overall supply–demand balance and (b) constraints on individual producers and consumers. In addition, now let us introduce some generic constraints on the rate of production and consumption, which we will denote: $k_j\left(Q^{\mathrm{S}}, Q^{\mathrm{B}}\right) \leq 0$. The overall welfare-maximisation problem is now as follows:

$$\max \mathrm{TS} = \max \sum_i U_i\left(Q_i^{\mathrm{B}}\right) - \sum_i C_i\left(Q_i^{\mathrm{S}}\right)$$

$$\text{Subject to :} \quad \text{(a)} \sum_i Q_i^{\mathrm{B}} = \sum_i Q_i^{\mathrm{S}} \;\leftrightarrow\; \lambda$$

$$\text{(b)}\; g_i\left(Q_i^{\mathrm{S}}\right) \leq 0 \leftrightarrow \alpha_i; \quad h_i\left(Q_i^{\mathrm{B}}\right) \leq 0 \leftrightarrow \beta_i$$

$$\text{(c)}\; k_j\left(Q^{\mathrm{S}}, Q^{\mathrm{B}}\right) \leq 0 \leftrightarrow \gamma_j$$

The KKT conditions for this problem are as follows:

$$U_i'\left(Q_i^{\mathrm{B}}\right) - \lambda - \beta_i \frac{\partial h_i}{\partial Q_i^{\mathrm{B}}} - \sum_j \gamma_j \frac{\partial k_j}{\partial Q_i^{\mathrm{B}}} = 0$$

$$C_i'\left(Q_i^{\mathrm{S}}\right) - \lambda + \alpha_i \frac{\partial g_i}{\partial Q_i^{\mathrm{S}}} + \sum_j \gamma_j \frac{\partial k_j}{\partial Q_i^{\mathrm{S}}} = 0$$

Let us proceed as before and ask whether or not this outcome can be achieved through a decentralised competitive market process. Comparing with the conditions for profit maximisation and utility maximisation above, we see that these outcomes can be achieved in a competitive market process provided that every producer and consumer is a price-taker and, in addition, the *i*th consumer faces the price:

$$P_i^{\mathrm{B}} = \lambda + \sum_j \gamma_j \frac{\partial k_j}{\partial Q_i^{\mathrm{B}}}$$

Similarly, the *i*th producer must face the price:

$$P_i^{\mathrm{S}} = \lambda - \sum_j \gamma_j \frac{\partial k_j}{\partial Q_i^{\mathrm{S}}}$$

The parameters λ and γ_j must be chosen in such a way that the overall supply–demand balance constraint is satisfied, and each of the generic constraints $k_j\left(Q^{\mathrm{S}}, Q^{\mathrm{B}}\right) \leq 0$ is satisfied.

1.6.2 A Smart Market Process

We have demonstrated that we can, in principle, decentralise the welfare-maximisation task using a competitive market process, provided we can set the prices correctly. But how might we achieve this objective?

Let us therefore consider a slightly different market process – one in which there is a central market operator which operates a constrained-optimisation process. Specifically, let us assume that the centralised market process operates through a series of steps as set out below:

a. Time is divided into a series of time intervals. Each interval of time, every seller submits to the market operator an offer function showing, at each price, the rate at which that producer is willing to produce. Similarly, every buyer submits to the market operator a bid function which shows, for every price, the rate at which he/she is willing to consume.
b. The market operator then carries out mathematical optimisation to find the combination of rate of production and consumption which maximises the total surplus (the sum of the utility functions of the customers and the cost functions of the buyers) under the assumption that the offer curve of each producer accurately reflects its true supply curve, and the bid curve of each buyer accurately reflects its true demand curve.
c. The market operator then sends out instructions to each buyer as to the rate at which that buyer should consume and, to each seller, the rate at which that producer should produce. The market operator also announces a market price (which may differ across producers and consumers).
d. Each producer and each customer, given the market price they face, choose a rate to produce or consume, respectively. Each producer is then paid the corresponding price for its production during an interval. Similarly, each customer must pay the corresponding price for its consumption during the interval.

Does this market process achieve an overall efficient allocation?

Let us suppose that each producer announces a cost function $\hat{C}_i\left(Q_i^{\mathrm{S}}\right)$ and each consumer announces a utility function $\widehat{U}_i\left(Q_i^{\mathrm{B}}\right)$. These announced functions may differ from the underlying 'true' cost and utility functions.

Given these announced functions, the market operator then carries out an optimisation task as follows:

$$\max \widehat{\mathrm{TS}} = \max \sum_i \widehat{U}_i(Q_i^{\mathrm{B}}) - \sum_i \hat{C}_i(Q_i^{\mathrm{S}})$$

$$\text{Subject to} \quad \text{(a)} \sum_i Q_i^{\mathrm{B}} = \sum_i Q_i^{\mathrm{S}} \leftrightarrow \lambda$$

$$\text{(b)}\, k_j\left(Q^{\mathrm{S}}, Q^{\mathrm{B}}\right) \leq 0 \leftrightarrow \gamma_j$$

Now, if the announced cost and utility functions accurately reflect the true cost and utility functions of the producers and consumers (as modified by any production or consumption constraints), then this maximisation yields the overall welfare maximising outcome. Furthermore, this optimisation yields the constraint marginal values λ and γ_j from which the prices can be calculated.

All that remains to show is that the producers and consumers will announce to the market operator cost and utility functions which reflect their true cost and utility functions – allowing for any production or consumption constraints.

Intuitively, this is straightforward to verify when there are no producer or consumer production or consumption constraints. Let us suppose a customer is a price-taker in the sense that he/she cannot, by changing his/her bid function, influence the market price. For each potential market price P, the customer must announce the rate at which he/she is willing to consume. But the amount that the customer is willing to consume at that price is just the amount which maximises the net value $U(Q) - PQ$, which is just the amount given by the customer's demand curve $Q^{\mathrm{B}}(P)$. In other words, if the customer is a price-taker in the market, he/she can do no better than submit a bid curve which perfectly reflects his/her demand curve. The same is true, of course, for the producers in the market. If each producer is a price-taker, it can do no better than submitting an offer curve which perfectly reflects its supply curve $Q^{\mathrm{S}}(P)$.

The argument is very similar when we take into account production or consumption constraints. We have seen that a welfare-maximising price-taking consumer which faces a price P will choose a quantity that satisfies

$$U_i'\left(Q_i^{\mathrm{B}}\right) - \beta_i \frac{\partial h_i}{\partial Q_i^{\mathrm{B}}} = P$$

In other words, the consumer has an incentive to offer to the market operator a demand curve $\widehat{U}_i'\left(Q_i^{\mathrm{B}}\right)$ which satisfies

$$\widehat{U}_i'\left(Q_i^{\mathrm{B}}\right) = U_i'\left(Q_i^{\mathrm{B}}\right) - \beta_i \frac{\partial h_i}{\partial Q_i^{\mathrm{B}}}$$

Similarly, each producer has an incentive to offer to the market operator a bid curve $\widehat{C}_i'\left(Q_i^{\mathrm{S}}\right)$ which satisfies

$$\hat{C}_i'\left(Q_i^{\mathrm{S}}\right) = C_i'\left(Q_i^{\mathrm{S}}\right) + \alpha_i \frac{\partial g_i}{\partial Q_i^{\mathrm{S}}}$$

In summary we have proven the following result:

Result: Even when there are generic constraints on production and consumption, the welfare-maximising outcome can be achieved using a generalised market process known as a smart market, involving a central market operator. In this market process:

a. Each producer and each consumer submits a bid or offer curve which reflects his/her true supply and demand function (after taking into account private production or consumption constraints).

b. The market operator carries out a constrained optimisation task, seeking to maximise economic welfare based on the announced bid and offer curves subject to (i) an overall supply–demand balance constraint and (ii) generic production-consumption constraints. This process determines a set of prices (which may vary between producers and consumers) and a rate of production and consumption for each producer and consumer.
c. Given the announced prices, each producer and each consumer chooses to voluntarily comply with the assigned rate of production or consumption.

The market process described here appears needlessly clumsy. It involves the communication of large amounts of information (bid and offer curves and despatch instructions) to and from a centralised market operator. And it requires the central market operator to carry out a large constrained-optimisation problem. In most markets, this degree of information sharing and centralised control is unnecessary and burdensome. However, we will see later that this is not the case in wholesale electricity markets. Wholesale electricity markets require a centralised process of the kind described here.

1.7 Longer-Run Decisions by Producers and Consumers

The previous section focused on short-run actions by producers and consumers. In practice, however, producers and consumers make a number of different forms of investments which affect either the cost of production or the value in consumption of the good or service.

For example, a producer might make a decision to expand its production capacity or to lower its production cost. A customer might make a decision to install new equipment which increases its demand for a particular good or service.

The difference between a short-run decision and a longer-run decision in practice depends on the extent to which the particular action can be changed in response to a change in the market price. If the action of the producer or consumer can be changed in response to each and every movement in the market price, the action is a short-run action. If, however, the action cannot be reversed as the market price changes, the decision is a longer-run (medium-term or long-term) decision.

The range of possible longer-run investments by producers and consumers is very large. For example, producers might face a decision as to what amount of productive capacity to install, at what location, at what time and of what type. Electricity consumers might face a decision as to what type of electrical devices to install, with what capabilities and of what size. Each of these decisions will require a slightly different economic model. We cannot cover all of these decisions here in this section. Instead, let us focus on a couple of simple questions.

1.7.1 Investment in Productive Capacity

To begin with, let us consider the decision to add more productive capacity. Let us suppose that there is a single producer and a single customer (which may represent the aggregate of many smaller producers and customers). To keep things as simple as possible, let us suppose that the cost of production of the producer varies linearly with the rate of production, up to the physical

maximum rate of production, which we will call the 'capacity' of the firm, and which we will label K (units/interval of time). We will suppose that each additional unit of output costs c (\$/unit) and an additional unit of capacity costs f (\$/unit/interval of time). In other words, let us assume that the cost function of the producer takes the form

$$C(Q,K) = cQ + fK$$

The rate of production of the firm is assumed to be able to be varied much more quickly than the capacity of the firm. In order to justify a change in the rate of production, we need to introduce some variability in the supply or demand conditions. Let us assume that the supply conditions of the firm are fixed over time, but that demand is varying. Let us suppose that there are several different states of the world, labelled s. In state of the world s, which occurs with probability p_s, the utility of customers for the good or service is given by $U_s(Q)$.

Since there is some uncertainty in this world, there is some risk. To keep things simple let us make the simplest possible assumption, which is that overall social welfare is neutral to risk. In practice this means that the overall social welfare is indifferent between receiving a fixed, certain pay-off and any uncertain pay-off which has the same expected or average value.

Let us assume that there is always sufficient demand relative to costs that it is always efficient to produce something.

To find the socially-efficient level of capacity K we must solve the following optimisation problem:

$$\max \sum_s \left[p_s(U_s(Q_s) - C(Q_s,K)\right]$$

$$\text{Subject to } Q_s \leq K \leftrightarrow \mu_s$$

The KKT conditions for a maximum of this problem are

$$\mu_s = U_s'(Q_s) - c \quad \text{and} \quad \mu_s(K - Q_s) = 0 \quad \text{and} \quad \sum\nolimits_s p_s\mu_s = f$$

In other words, the optimal level of capacity is the level where the average distance between the marginal value of consumption and the marginal cost of production at times when the capacity of the firm is exhausted is equal to the cost of additional capacity (Figure 1.5). This is a simple example of a classic *peak load pricing* problem.

We will see later that this conclusion generalises to the case where there are different types of producers, with different costs of production and different costs of expanding capacity. As in the example above, we find that for each different type of producer, the capacity of that type of producer is at the optimal level when the average gap between the marginal value of the customer and the marginal cost of production at times when the capacity of that type of producer is exhausted is equal to the cost of adding capacity of that type.

Now let us ask the question: will this level of capacity be chosen in a competitive market? Let us assume that we have a competitive market. In the short run, as we have seen above, the market price will be equal to the common marginal value of all the customers and the marginal cost of production. Let us suppose the market price in state of the world s is P_s. Let us consider

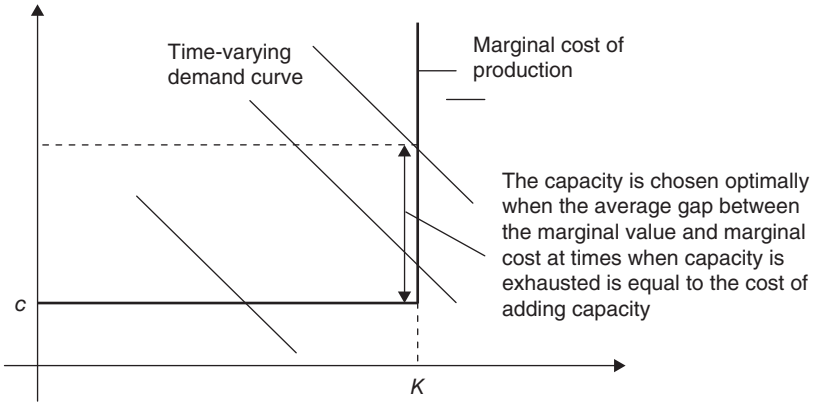

Figure 1.5 Determining the optimal level of capacity

the level of capacity chosen by a producer which seeks to maximise its expected profit. In other words, the producer solves the following problem:

$$\max \sum_{s} p_s[P_s Q_s - C(Q_s, K)]$$

$$\text{Subject to } Q_s \leq K$$

The first-order conditions for this problem show that (a) the firm produces at capacity when the market price exceeds the marginal cost of production and (b) the level of capacity is optimal if the average gap between the market price and the marginal cost of production at times when the market price exceeds the cost of production is equal to the cost of adding capacity.

But from the earlier analysis we know that the market price is equal to the marginal value of consumption. It therefore follows that in a competitive market a firm which can adjust capacity continuously will choose a level of capacity which is socially optimal.

It is straightforward to check that this result generalises to the case where there are several different types of production technologies. We carry out this analysis in Part IV. In other words, under the simple assumptions set out in this example, competitive markets will deliver socially efficient outcomes in both the short run and the long run.

There are many extensions to these simple results which we will explore further throughout this text.

1.8 Monopoly

The previous sections used the simplifying assumption that producers and customers could not influence the market price – in other words, that they were price-takers. This is a reasonable assumption where there are a large number of producers and customers, none of which is very large relative to the size of the market. However, this is not always the case. In some markets there is a single dominant firm, or a small group of firms. In either case it is likely that these firms will have some impact on the price or prices at which the good or service is transacted in the market.

In fact, a firm which faces little or no competition not only will be able to influence the price or prices at which the good or service is transacted in the market but will often also have some control over the way that goods and services are charged for in the market. For example, a firm with little or no competition may be able to charge different prices for different types of customers. Alternatively, such a firm may be able to charge its customers a separate fee for the right to consume at all, and then another fee per unit consumed known as a two-part tariff. These different pricing practices are known generically as *price discrimination*. Some forms of price discrimination arise in competitive markets. As a general rule, however, price discrimination is much more common in markets which lack competition.

Let us put aside this question of how the prices are structured, and simply focus on one component – the price for additional units of the good or service per interval of time.

As discussed in more detail in Part VII, a firm which is not a price-taker has some influence over this price and is said to have *market power*. In particular, if the firm produces at a very slow rate, the market price, which allocates this slow rate of production amongst customers, is likely to be very high. Conversely, if the firm produces at a very high rate, customers are likely to require a low price to absorb that high rate of production.

Let us suppose that for some group of customers, other things equal (in particular, holding constant the rate of production of all other firms), for each rate of production Q of the firm, there is some corresponding price P. The fact that the firm can influence the market price implies that the market price is a function of the rate of production of the firm. We can represent this by saying that the firm faces a downward sloping *residual demand curve*, which we will label $P^{\mathrm{RD}}(Q)$.

As before we will consider first the short-run decisions of such a firm – in particular, the decision as to the rate of production. Let us suppose that a firm facing a downward-sloping residual demand curve seeks to choose a level of output which maximises its short-run profits. Will it choose a level of output which maximises total economic surplus? The problem this firm faces can be expressed as follows:

$$\max P^{\mathrm{RD}}(Q)Q - C(Q)$$

The first-order condition for this problem is

$$P^{\mathrm{RD}}(Q) + Q\frac{\mathrm{d}P^{\mathrm{RD}}}{\mathrm{d}Q} = C'(Q)$$

The left-hand side of this expression is the *marginal revenue curve*. This expression says that a firm with market power will produce at a quantity where the marginal revenue (using the residual demand curve) is equal to the marginal cost of production.

Since the residual demand curve is downward sloping, the second term on the left-hand side is negative, so a profit-maximising firm with market power, selling to a group of customers at a fixed price independent of the rate of consumption, chooses to produce at a rate where the marginal cost of producing at that rate is less than the price at which it sells the output.

Earlier we saw that if customers are small, they will behave as price-takers and will consume at a rate where their marginal utility is equal to the price they are charged. But, we have just seen above: this price is higher than the marginal cost of production. So, a profit-maximising firm with market power, selling at a fixed price per rate of consumption, chooses a volume which is

too low relative to the efficient outcome. This results in a reduction in economic welfare relative to a theoretical ideal, which is known as the *deadweight loss*.

We can re-write the expression above as follows:

$$\frac{P^{\mathrm{RD}}(Q) - C'(Q)}{P^{\mathrm{RD}}} = -\frac{Q}{P^{\mathrm{RD}}}\frac{\mathrm{d}P^{\mathrm{RD}}}{\mathrm{d}Q} = \frac{1}{\varepsilon^{\mathrm{RD}}}$$

Here $\varepsilon^{\mathrm{RD}}$ is the elasticity of the residual demand curve, which is defined as follows:

$$\varepsilon^{\mathrm{RD}} = -\frac{P^{\mathrm{RD}}}{Q}\frac{\mathrm{d}Q}{\mathrm{d}P^{\mathrm{RD}}}$$

The expression on the left-hand side above is known as the *Lerner index* or the price-marginal-cost margin. This expression then says that a profit-maximising firm facing a downward-sloping residual demand curve $P^{\mathrm{RD}}(Q)$ will choose a rate of production where the Lerner index is equal to the inverse of the elasticity of the residual demand curve.

1.8.1 The Dominant Firm – Competitive Fringe Structure

As an example, let us consider the case of a large firm which faces competition from a number of very small rivals. This market structure is known as a *dominant firm* with a *competitive fringe*. The good or service produced by the dominant firm is assumed to be a perfect substitute in the eyes of the customers for the good or service produced by the rivals.

First let us ask what is the short-run social welfare maximising rate of production and consumption. Since the product of the dominant firm and that of the rivals are perfect substitutes in the eyes of the customers, the total rate at which the value is received by the customers depends only on the total rate of production: $U\left(Q^{\mathrm{DF}} + Q^{\mathrm{CF}}\right)$. The total cost of production is the sum of the cost of production of the dominant firm and the competitive fringe: $C^{\mathrm{DF}}\left(Q^{\mathrm{DF}}\right) + C^{\mathrm{CF}}\left(Q^{\mathrm{CF}}\right)$. The social welfare maximising rate of production and consumption is the solution to the following problem:

$$\max U\left(Q^{\mathrm{DF}} + Q^{\mathrm{CF}}\right) - C^{\mathrm{DF}}\left(Q^{\mathrm{DF}}\right) - C^{\mathrm{CF}}\left(Q^{\mathrm{CF}}\right)$$

Here the maximum is taken over both the production of the dominant firm Q^{DF} and the production of the competitive fringe Q^{CF}. Solving this problem we find that (as before) social welfare is maximised when the marginal cost of increasing the rate of output by the dominant firm and by the competitive fringe is the same and is equal to the marginal value for the good or service.

Let us suppose that all the rivals in the competitive fringe are sufficiently small, so that they are price-takers. As a consequence, we can construct a supply curve for the rivals which we will denote by $Q^{\mathrm{SCF}}(P)$. Since the rivals are price-takers, this supply curve reflects the marginal cost curve of the rivals:

$$C^{\mathrm{CF}'}\left(Q^{\mathrm{SCF}}(P)\right) = P$$

Similarly, we will suppose that there are a large number of small customers who are price-takers. We can therefore aggregate the demand curves of each customer to form the market

demand curve, which we will denote by $Q^{\mathrm{B}}(P)$. For any price chosen by the dominant firm, it will sell a product at a rate equal to the demand of customers less the supply rate of the competitive fringe. This is the residual demand curve:

$$Q^{\mathrm{RD}}(P) = Q^{\mathrm{B}}(P) - Q^{\mathrm{SCF}}(P)$$

Now suppose that this firm seeks to choose a price which maximises its profits. As noted above, the dominant firm chooses the price which maximises:

$$\max Q^{\mathrm{RD}}(P)P - C\left(Q^{\mathrm{RD}}(P)\right)$$

This is maximised by a price P which satisfies

$$P = C'\left(Q^{\mathrm{RD}}\right) - Q^{\mathrm{RD}}\frac{\mathrm{d}P}{\mathrm{d}Q^{\mathrm{RD}}}$$

In other words, the dominant firm will choose to sell at a price which is above its own marginal cost of production. The extent of the gap depends on the slope of the residual demand curve. Since the customers choose their rate of consumption, and the competitive firms choose their rate of production, based on the price they face, it follows that a short-run economic inefficiency will arise: Customers will consume at a rate which is too low relative to the efficient level, and the competitive fringe will produce at a rate which is too high.

From this analysis we can identify the two components of the deadweight loss: Customers consume at a rate which is inefficiently low relative to the efficient level, and the competitive fringe produces at a rate which is too high. Overall short-term economic efficiency could be improved by increasing the rate at which the monopoly supplier produces, reducing the rate at which the competitive fringe produces and increasing the rate at which customers consume.

We can re-write the expression above in the form of the Lerner index:

$$\frac{P - C'\left(Q^{\mathrm{RD}}\right)}{P} = -\frac{Q^{\mathrm{RD}}}{P}\frac{\mathrm{d}P}{\mathrm{d}Q^{\mathrm{RD}}} = \frac{1}{\varepsilon^{\mathrm{RD}}}$$

where $\varepsilon^{\mathrm{RD}} = s\varepsilon + (s/1-s)\varepsilon^{\mathrm{SCF}}$ and s is the market share of the dominant firm, ε is the elasticity of the market demand and $\varepsilon^{\mathrm{SCF}}$ is the elasticity of supply for the competitive fringe.

1.8.2 Monopoly and Price Regulation

The analysis above has highlighted the short-run economic harm from deadweight loss. There are other important economic harms from market power. Specifically, firms with market power can often price discriminate between customers. At the same time, customers will often be required to make material sunk investments in reliance on the monopoly service. This is particularly the case in, say, the electricity industry where customers must invest in significant consumption assets to draw economic value from the electricity service. The customer may be concerned that the monopolist will raise its prices after the customer has made a sunk investment. Facing the threat of hold-up, the customer may be reluctant to make investments in the first place.

In the longer run, the primary economic harm from market power appears to be the threat that market power poses to investment – especially investment by customers in reliance on the monopoly service.

There are two potential solutions to this hold-up problem: vertical integration and long-term contracts. Both are common in monopoly industries. Historically, in many countries, electricity network businesses were (and still are) government owned – a form of vertical integration. In other countries, electricity network businesses are privately owned but are subject to a form of long-term contract, known as public utility regulation.

1.9 Oligopoly

In some markets it is possible to accurately model the decisions of one producer or consumer in isolation from the decisions of all the others. For example, in the case of a competitive market we made the assumption that each firm was a price-taker – that is, each firm was too small to influence the market price. This allowed us to consider the actions of each firm separately from every other firm.

In the case of the dominant firm with a competitive fringe, we could not completely ignore the impact of the competitive fringe when considering the output decisions of the dominant firm. But in that case the reaction of the competitive fringe was particularly simple and could be summarised in the residual demand curve.

In certain other markets it is not possible to consider the decisions of one firm in isolation from the decisions of the others. This occurs in markets where there are several firms each large enough to have some impact on the market price. This market structure is known as an *oligopoly*.

A situation where the actions of one player affect the pay-off or earnings of another player is said to be a situation of *strategic interaction* and is conventionally modelled using the tools of *game theory*.

Previously we noted that, in order to model the behaviour of a producer or consumer, we need to specify (a) the actions they can take and (b) their objectives. In the case of a game of strategic interaction we also need to specify an *equilibrium concept* – that is a means of selecting predicted or equilibrium outcomes.

The simplest equilibrium concept is the *Nash equilibrium*. A Nash equilibrium is a set of actions, one for each player, which has the property that for each player the action specified for that player is the best (highest pay-off) for that player given the actions specified for the other players.

In addition, in a game of strategic interaction it turns out that it is important to distinguish between a game that is played once (referred to as a *one-shot game*) and a game that is played many times (referred to as a *repeated game*). For the moment let us focus on one-shot games.

The two simplest economic models of oligopoly are the games between firms known as the Cournot and Bertrand games. These games differ in the assumptions that each firm makes about the actions of the other firms in response to its own actions. In the *Cournot game*, each firm assumes that the other firms will hold their output fixed when the firm in question adjusts its price or output. In the *Bertrand game*, in contrast, each firm assumes that the other firms hold their prices fixed when the firm in question adjusts its price or output.

1.9.1 Cournot Oligopoly

To make this clear, let us suppose that we have two firms, each producing a single good or service. Let us suppose that the two goods are perfect substitutes in the eyes of consumers, so that consumers care only about the total rate of production of the goods.

Let us suppose that firm 1 chooses a rate of production Q_1 (units/interval) and firm 2 chooses a rate of production Q_2. Let us assume that we can represent the objectives of consumers in a single utility function $U(Q_1, Q_2) = U(Q_1 + Q_2)$. Then, if we assume that the good is sold at a simple linear price and if all consumers are price-takers, we can derive a single demand curve $P(Q_1, Q_2) = P(Q_1 + Q_2)$.

Let us suppose that the cost function of firm i is given as $C_i(Q_i)$. As we saw earlier, the efficient outcome is one in which the marginal cost of production is the same for the two firms (this may imply that one firm produces nothing if the other firm has a lower marginal cost of production).

Let us suppose that each firm chooses the rate of production which maximises the rate at which it earns profits under the assumption that the other firm holds its rate of production fixed. Each firm therefore solves the following problem:

$$\max Q_i P(Q_1 + Q_2) - C_i(Q_i)$$

The first-order condition for this maximisation yields

$$C_i'(Q_i) = P(Q_1 + Q_2) - Q_i \frac{\partial P}{\partial Q_i}$$

This yields two equations in two unknowns which can be solved to find the equilibrium rate of production of the two firms. This is left as a problem in Section 1.10. It is easy to check that this outcome is inefficient: The total rate of production is less than the efficient level. Furthermore, there is some production inefficiency in producing that less-than-efficient amount – the gap between the price and the marginal cost is larger for the larger firm.

It turns out that if we extend this model by allowing for more firms, we find that the Nash equilibrium approaches the competitive level as the number of firms in the industry increases.

In this simple model each firm chose its own rate of production assuming that other firms hold their own rate of production fixed. But, in order to hold their own rate of production fixed, they will typically have to adjust their prices. It might be thought that this seems unnatural. After all, isn't it the case that firms compete by adjusting their prices rather than their quantities? As noted above, there is a related game in which firms choose their own prices assuming that other firms hold their own prices fixed. This is known as a Bertrand game.

1.9.2 Repeated Games

The analysis here uses the notion of a Nash equilibrium. A Nash equilibrium has the desirable characteristic that if the other players happen to play the Nash equilibrium actions then no player will unilaterally want to change his/her action. The game the firms were playing above was a *one-shot game*: The firms come into existence, choose their actions, receive the pay-off, and then go out of existence. But this raises deep questions. If the firms did not exist in the past how do they reach the Nash equilibrium in the first place? What if

there is more than one Nash equilibrium? In that case how do the firms know which equilibrium to coordinate on?

In almost all practical circumstances, firms do not just interact once. Instead, they interact repeatedly in the market. This ongoing interaction allows for a much richer set of strategies and actions than can be represented in a one-shot game. In particular, firms can react to the choices made by other firms in the past, rewarding or punishing them. When players interact repeatedly over time, the appropriate modelling tool is a *repeated game*. The range of equilibrium possible outcomes is much larger in a repeated game than in the one-shot games discussed above.

The reason for this is that in a repeated game it is typically possible to use much more severe 'punishment strategies'. In a one-shot game, if the other player does not play the best co-operative outcome there is nothing that either side can do. But in a repeated game, if one player does not co-operate the other player can switch to playing a strategy which the first player really does not like.

If the game is repeated for long enough, the parties may be able to sustain an equilibrium which yields a much higher pay-off than the one-shot game. In fact, in some circumstances it may be possible for the firms to collude to sustain an outcome which is as though the firms are operated as a single firm – a monopoly. This is likely to involve a lower rate of production and a higher price than the Cournot game discussed above. The easiest way to achieve this collusion is simply for the two firms to merge (also known as *horizontal integration*). But even if merger is not feasible, they may be able to come to an explicit or implicit agreement as to how much to produce and how to share the proceeds.

Most countries have a competition law which places strict limits on mergers between rivals, and which prohibits anticompetitive agreements between firms. Nevertheless, in some oligopoly markets, firms may be able to sustain arrangements which are less-than-fully competitive without an explicit agreement. This is known as *tacit collusion*.

As we will see, wholesale electricity markets are prone to the exercise of market power. In some markets there may be a single dominant firm, facing a downward sloping residual demand curve, as discussed above. In other cases, however, there will be several firms each of which has a degree of market power, and the market is better modelled as an oligopoly. Since electricity generators repeatedly interact with each other in the wholesale market, they may be able to sustain various forms of collusion or co-operative outcomes. Market power in the electricity market is discussed further in Part VII.

1.10 Summary

Economists model the short-term behaviour of consumers using a utility function (also known as the consumers' surplus). Consumers are assumed to maximise their net utility (utility from consumption less the amount they must pay to consume). Faced with a simple linear price a price-taking consumer will consume to the point where the marginal utility is equal to the price. This yields a downward sloping demand curve for each consumer which (if all consumers face the same price) can be aggregated to form the market demand curve.

Similarly, the short-run behaviour of producers can be modelled using the concept of the cost function. Producers are assumed to maximise their profit (their revenue from production less the expenditure they incur). Faced with a simple linear price, a price-taking producer will produce to the point where the marginal cost is equal to the price. This yields an upward-sloping supply curve which (if all producers face the same price) can be aggregated to form the market supply curve.

The efficient level of production and consumption in a market occurs where the marginal utility of each customer is the same and is equal to the marginal cost of each producer. In a competitive market this can be achieved through a normal market process.

In a smart market the normal market process is replaced with a centralised mechanism in which producers and consumers submit bids and offer functions to a central market operator. The central market operator carries out a constrained optimisation problem which may include physical constraints. This market process also yields an efficient outcome under the assumption that all players are price-takers. Electricity markets are a form of smart market.

In the longer run producers and consumers make investments – for producers, investments in productive capacity; for consumers, investments in assets which increase the value of the good or service traded in the market. The short-run variation in the spot price in a competitive market can provide efficient signals for expansion of productive capacity.

Some firms are not price-takers, but have the ability to influence the wholesale market price. Such firms are said to have market power. When selling at a simple linear price a monopoly will not choose the efficient level of output, but will choose to produce too little. This results in a loss of welfare known as deadweight loss. By adopting more complicated price structures (known as price discrimination) the monopolist will often be able to reduce the deadweight loss.

In some markets the action of one participant affects the pay-off of another. These situations of strategic interaction are best modelled using the tools of game theory. The most common equilibrium concept is the Nash equilibrium. In economics, game theory is used to model the short-term strategic interaction of a small number of firms, also known as oligopoly. One common approach is to model a game in which each firm chooses its level of output assuming that the other firms hold their output fixed. This is known as a Cournot game. The resulting outcome lies between the monopoly and the competitive outcomes.

Questions

1.1 Show that the (inverse) demand curve is downward sloping.

1.2 Let us suppose there are two products in a market. The rate of consumption of these products is denoted by Q_A and Q_B. The utility function for consumers of this product is

$$U(Q_A, Q_B) = 100A_0 + 10Q_A + 20Q_B - 2Q_AQ_B - 3/2Q_B^2$$

Find the demand curves for products A and B. How does the demand curve for product A change in response to a change in the rate of consumption of product B?

1.3 Suppose we have a monopoly provider of a particular good or service which charges a simple linear price. Suppose that the demand curve for this service is linear and takes the form:

$$P(Q) = A - BQ$$

Suppose that the cost of production takes the form: $C(Q) = cQ$. Show that the monopoly profit-maximising rate of production is equal to half of the socially efficient rate of production. Show that deadweight loss is equal to 25% of the total potential

economic welfare in this market and that the deadweight loss decreases as P decreases provided $P > c$

1.4 Two firms produce two products, A and B, which are partial substitutes for each other. Given the price for A and B, P_A and P_B respectively, the rate at which customers choose to consume in the market is $Q_A(P_A, P_B)$ and $Q_B(P_A, P_B)$ respectively. Suppose that the cost of production is given by $C_A(Q_A) = c_A Q_A$ and $C_B(Q_B) = c_B Q_B$ respectively. Write down the profit of each firm as a function of the two prices in the market. Assuming that each firm chooses its price to maximise its profit while assuming that the other firm holds its price fixed, find an expression for the Bertrand–Nash equilibrium.

Further Reading

The material in this chapter is a modified form of the material which can be found in numerous introductory economics textbooks. For even more on these subjects you might be interested to explore the subject of Industrial Organisation which deals with the behaviour of firms in markets. One useful reference is Church and Ware (2000). See also Luenberger (1995). The concept of smart markets was introduced by Rassenti, Smith and Bluffing (1982).