CHAPTER 22

NONLINEAR PRICING

SHMUEL S. OREN

22.1 Introduction

While basic economic theory characterizes products as homogeneous commodities that are traded at uniform unit prices so that purchase price is proportional to purchase quantity, real products and services are more complex. Quantity metrics, to the extent they are meaningful, represent only one dimension upon which purchase prices are based. Nonlinear pricing is a generic characterization of any tariff structure where the purchase price is not strictly proportional to some measure of purchase quantity but also reflects other characteristics of the product, the purchaser, the purchase as a whole, its timing and any contractual terms imposing restriction on the purchase and its subsequent use. A fundamental aspect of nonlinear pricing methodology is the systematic exploitation of heterogeneity in customer preferences with respect to purchase characteristics and the explicit modeling targeting the preference structures underlying such heterogeneity. In that respect, nonlinear pricing theory differs from revenue management, which recognizes customer heterogeneity but typically models it as a random phenomenon. A key assumption of nonlinear pricing is the existence of identifiable differences among customers that affect their choices in a systematic way. Furthermore, it is assumed that these differences among customers are either directly observable or that customers can be sorted by observing measurable characteristics of the customer or her purchases.

Nonlinear pricing is motivated by several goals such as: efficient use of resources, cost recovery by a regulated utility, exercise of monopoly power, obtaining competitive advantage, rewarding customer loyalty as well as social goals such as subsidies to the poor and discounts to service persons in uniform. Being able to sell identical or similar products or services at different prices to different customers has powerful ramifications and can lead to win–win outcomes from the customers' and the sellers' perspectives.

To illustrate such potential benefits let us consider the classic case of a homogeneous commodity sold by a monopoly supplier at a uniform unit price. The demand for the commodity is characterized by a simple downward sloping demand function. Such a demand function does not distinguish between multi-unit purchases by a single customer

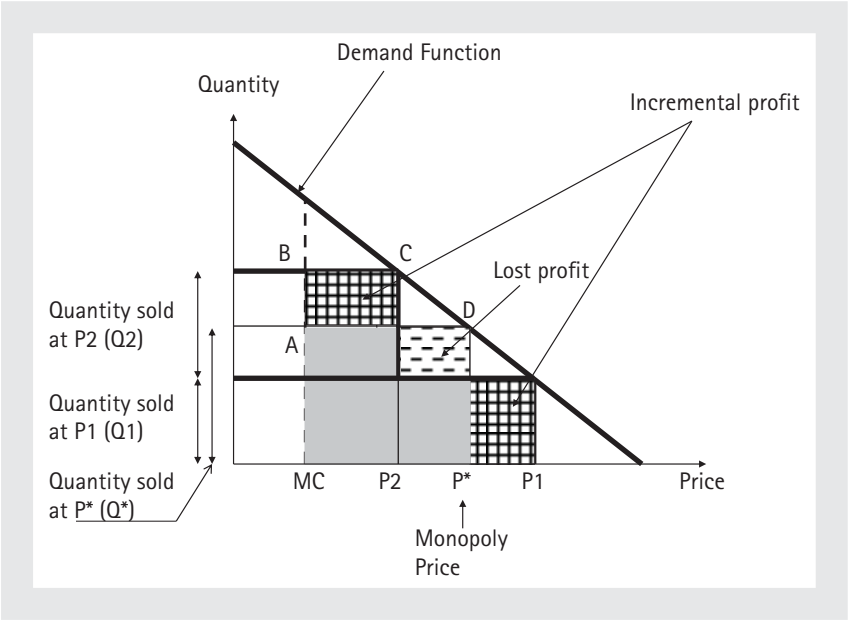

FIGURE 22.1 Increased monopoly profits (incremental minus lost) and social welfare gain (area ABCD) through bifurcated pricing.

or single unit purchases by many customers. In making its pricing decision, the monopoly supplier must trade off increased profits from selling additional units by lowering the price against the lost profits from existing sales. Consequently, the monopoly supplier will set a price above marginal cost, which is suboptimal from a social welfare[1] perspective (since it excludes some customers who are willing to pay more than the product costs). If the monopoly supplier were able to segment the demand and charge two prices for the same product as illustrated in Figure 22.1, more demand valuing the product above marginal cost would be served ($Q1 + Q2 > Q^*$), increasing social welfare. Furthermore, the original monopoly profit $(P^* - MC) \cdot Q^*$, could increase, if the incremental profit exceeds the lost profit (as in Figure 22.1), resulting in a "win–win" proposition.

The difficulty in implementing such market segmentation based on customers' willingness-to-pay is that such information is typically private. Furthermore, such price discrimination would require some means of preventing the high paying customers from purchasing the product at the low price.

Nonlinear pricing, which implements the basic idea illustrated above in a variety of contexts, encompasses basic principles of price discrimination, product differentiation, and

[1] Social welfare (also referred to as social surplus) measures the total benefit to society from production and consumption of a good or service. It is defined as the total benefit from consuming the good or service (as reflected by customers' willingness-to-pay) less the production cost. Social welfare is also the sum of the consumer surplus and producer surplus. The consumer surplus measures the net benefit to consumers from the good or service and is defined as the aggregate willingness-to-pay minus payment. The producer surplus measures producers' total profit (i.e. revenue less production cost) from selling the good or service. Since payments for the good or service constitute a transfer from consumers to producers, prices only affect social surplus to the extent that they affect production or consumption quantities.

market segmentation. However, for all practical purposes, these terms are synonymous and used interchangeably. Unfortunately, the negative connotation of the term "discrimination" often obscures the efficiency gains and Pareto improvement that can be achieved by such practices. For that reason many important contributions to the theory and practice of nonlinear pricing (e.g. Wilson 1993) have tried to disassociate nonlinear pricing from the price discrimination interpretation and the use of the term "Nonlinear Pricing" emphasizes the departure from the classical uniform unit price concept.

The classic economic theory of price discrimination has focused on how to segment the demand for a product or a service and supply them to different segments of the market at different prices. Often, such segmentation requires differentiation of the product or services so that the buyer perceives different values for the different prices. Furthermore, the seller must possess some degree of market power which means that resale markets are limited, either through direct control or due to high transaction costs. For example a volume discount strategy would not be sustainable if customers can combine purchases and share the cost. Likewise a tariff that increases per unit cost with purchase quantity (like lifeline tariffs for electricity or water) could not be implemented if a customer could split its consumption among several meters. Economists have pointed out that introducing product variants aimed at segmenting the market could result in quality degradation and loss of social welfare but here we will not concern ourselves with such consequences.

The principles of price discrimination were introduced by Pigou (1920) who distinguished between three basic forms of price discrimination:

- First degree (Direct) discrimination where prices are based on the purchasers' willingness-to-pay.
- Second degree (Indirect) discrimination where prices are based on some observable characteristics of the purchase (e.g. volume), which is correlated with the customer's preferences.
- Third degree (Semi-direct) discrimination where prices are based on some observable characteristics of the buyer (e.g. geographic location or age).

To illustrate the difference between Semi-direct and Indirect price discrimination consider the example of a children's menu in a restaurant which under a semi-direct discrimination policy can be ordered only by children. By contrast, an indirect discrimination approach would offer on the menu discounted small portions of assorted items that are unlikely to be ordered by an adult but without prohibiting such orders. Nonlinear pricing falls under the category of indirect or second degree discrimination. The efficiency properties of such practices stem from the fact that they induce customers to sort themselves and reveal private information that leads to improved production and allocative efficiencies.[2]

Necessary conditions for sustainability of price discrimination strategies are various forms of nontransferability conditions. In the case of indirect discrimination the demand must be nontransferable, meaning that ~~the~~ one type of purchase, for example high end wine bottles, be met through decanting of discounted jug wine of the same brand. Such a

[2] Production efficiency refers to the extent to which a good or service is produced at least cost while allocative efficiency refers to the extent to which a good or service is allocated to its highest valued use.

possibility would undermine a volume discount strategy. Likewise, semi-direct discriminating requires nontransferability of the product, for example a discounted senior ski ticket cannot be used by a non-senior person. Nontransferability of products (or services) is relatively easy to enforce. Airline restrictions on transfer of tickets represent a classic example of such practices. Nontransferability of demand is harder to enforce but can be facilitated by technological constraints, product differentiation (sometimes at a cost), search cost, and transactions costs. The requirement for a Saturday night stay is an example of product differentiation for the purpose of discriminating between business and recreational travelers at the expense of unutilized plane capacity on Saturdays. Frequent travelers were able for a while to overcome this restriction through overlapping back to back bookings but the airlines were able to curb such practices using sophisticated monitoring of reservations (see Barnes Chapter 3).

Direct discrimination is rare since it requires both types of nontransferability as well as information regarding the customer's preferences and the states of nature upon which such preferences may depend. Nevertheless, pricing of services based on the value of a transaction, for example sale of real estate or pricing of personal services, comes close to direct price discrimination.

In this chapter we will focus primarily on indirect price discrimination, which underlies most of the commercially motivated nonlinear pricing schemes. An exception that will be discussed is Ramsey pricing which discriminates among customer types (e.g. industrial versus residential customers). The objective of such pricing is to achieve cost recovery in regulated utilities with concave cost structures with least efficiency losses due to deviation from marginal cost pricing (known as second best policies).

From an economic theory perspective, the design of nonlinear pricing schemes as indirect price discrimination mechanisms falls into the general category of mechanism design and agency theory (e.g. Tirole 1988) where the seller can be viewed as the principal who designs an incentive scheme that will induce desired purchase behavior by its customers who are the agents.

An indirect price discrimination mechanism must first identify target characteristics, which differentiate customers and develop a sorting mechanism that separates customers according to the target characteristic such as quantity choice, time of use, time value, or level of use. In order to implement such a mechanism we must have disaggregated demand data specifying customer preferences with regard to various product attributes. Assembling such data requires that at a minimum we are be able to specify the following aspects:

- What is a customer? (For instance regarding frequent flyer plans, the customer and the billing account may not be the same.)
- Dimension of the tariff (physical units, number of transactions, dollar amount).
- Units of purchase (kWh, KW, metric cube)
- Quality dimensions (time of use or interruptibility for electricity service, advance reservation, and flexibility for airline tickets)
- Method of billing (low daily rate with mileage charge versus flat daily rate with unlimited miles). Terms of the contract and method of billing may be sometimes interpreted as quality attributes.

In the following we will discuss in more detail five generic nonlinear pricing schemes that will illustrate the underlying theory and practical applications of such methods:

- Bundling
- Quantity discounts
- Ramsey pricing
- Quality differentiation
- Priority pricing and efficient rationing

The objective of this chapter is neither to be exhaustive in surveying nonlinear pricing practices and methods nor to be comprehensive in terms of the theoretical foundation of the nonlinear pricing methods discussed and the related literature. For an extensive treatment of nonlinear pricing the reader is referred to Wilson (1993) that provides a deep analysis of such methods along with a detailed bibliographic survey and historical review of the area. This chapter is written primarily as a tutorial with the objective of conveying the philosophical basis for nonlinear pricing and highlighting thematic application areas, key ideas, and the basic methodologies used in designing such tariff structures.

22.2 Bundling

Bundling is the most basic form of nonlinear pricing and indirect price discrimination which segments the market by offering commodities either separately or in a bundle which is offered at a price below the sum prices of the components. There is a fine line between bundling and "tying" which is illegal in the USA. Under tying, customers are forced to buy one thing as a condition for being able to buy another popular or essential product or service. Companies often use tying as a mechanism to monitor usage of the essential product, which will enable them to discriminate based on usage. For instance IBM used to force their customers who bought IBM computers to buy only IBM punch cards. By controlling the price of the punch cards they were effectively able to charge their computers different prices based on use. Similarly Xerox was forcing their customer to use only Xerox toner in their copiers and more recently HP was trying to force their customers to buy HP maintenance services for their HP computers. These practices are now considered illegal.

By contrast, bundling refers to the practice where products or services are sold together as a package providing a discount relative to component pricing. Pure bundling means that only the package is offered whereas mixed bundling means that both the package and the components are available. To see how bundling can be beneficial consider the following example adapted from Stigler (1963). Suppose that we have two products X and Y and two types of customers A and B. The products are unique and are produced by a monopoly supplier at zero marginal cost. Table 22.1 summarizes the willingness-to-pay (WTP) of each customer type for each of the products and the resulting market outcomes. We observe that by offering the bundle the monopolist is able to increase its profits from \$19 to \$20, by exploiting the negative correlation in preference among the two customer types.

Table 22.1 Mixed bundling example

	WTP by A	WTP by B	Monopoly price	Profit
Product X	\$8.0	\$7.0	\$7.0	\$14.0
Product Y	\$2.5	\$3.0	\$2.5	\$5.0
Bundle X + Y	\$10.5	\$10.0	\$10.0	\$20.0

When we have a continuum of customers that are characterized by their willingness-to-pay for product X and Y we can identify the regions in which consumers will buy the separate products and the bundle as illustrated in Figure 22.2. We denote the prices and corresponding costs of the component products X and Y as P_X,P_Y,C_X,C_Y respectively, and the bundle price for one unit of X and one unit of Y as P_b. Furthermore we consider the case where a customer will only consider buying at most one unit of each product (e.g. travel and lodging for a vacation). If a bundle is not offered then customers in the area AJEC would buy only product Y since their willingness to pay for product X is less than its price. Likewise, customers in the area KEGM would only buy product X and customers in the area JEK would buy the two products since their willingness-to-pay for each product exceeds the price. With the bundle we increase sales of the two products by essentially offering them at a discount through the bundle to customers who buy both. We are also

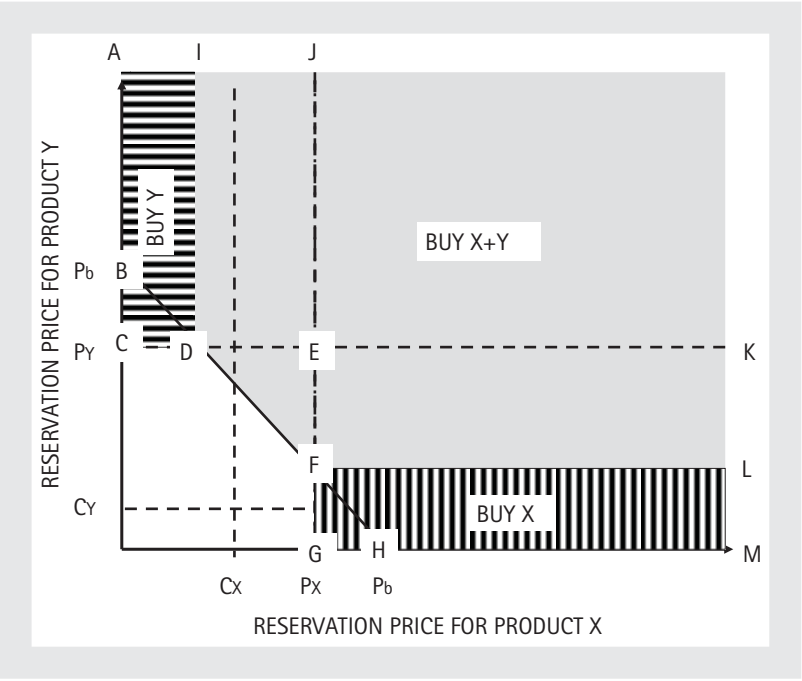

FIGURE 22.2 Illustration of customer choices under mixed bundling

able to sell the bundle to customers in the area DEF who would not buy anything if the bundle was not offered. The optimal price for the bundle and the component products can be determined by formulating an optimization problem that will maximize the seller's profit given the distribution of customers' willingness-to-pay for the two products. Since offering the bundle at a price that equals the sum of the components is a feasible solution of such optimization, mixed bundling is guaranteed to yield at least as much profit as simple linear pricing of the component products.

While the analysis of bundling can be extended to more than two products the graphical representation gets messy and we will not pursue it any further for bundling arbitrary products. However, we can analyze in more generality special kind of bundles consisting of multiple units of the same product. In such a case the bundling strategy is referred to as quantity discounts.

22.3 Quantity discounts

In order to analyze quantity discount strategies, we have to extend our concept of a demand function to capture the divergence among customer types with regard to purchasing of multiple units of a product. If we assume that all units are sold at the same price, as in basic economic theory, we do not care if ten units are purchased by ten different customers or by one customer. However, if we want to use purchase quantity as a means for screening customers by type, a more disaggregated demand model is needed. We can do it by defining a demand profile, $N(q, p)$, that describes how many customers will buy q units or more of the product at price p. Alternatively, we may think of each incremental unit of purchase as a separate product so $N(q, p)$ may be interpreted as the demand function describing the demand for the qth unit purchased by a customer as a function of the price charged for the qth unit. This will allow us to set the price of each incremental unit of purchase separately and obtain a price function, $p(q)$, which specifies the marginal price for the qth unit purchased by a customer.[3] In practice, volume discounts take the form of block declining tariffs characterized by break points at discrete quantity levels as shown in Figure 22.3.

The lower part of the figure shows the marginal unit price, which changes as purchase quantity increases while the upper part shows the cumulative payment as function of quantity. Note that we can also have a fixed charge such as a monthly fee for phone service on top of which we have a per-minute charge, which declines with usage. A two-part tariff consisting of a fixed charge and a constant per unit charge is simply a declining block tariff with a single block.

In order to sustain such a pricing scheme it must be impossible or costly for buyers to get together and buy a larger quantity at a discount and split it among themselves. Often, when dealing with packaged goods like cereal not all quantities are available and the supplier offers just a few box sizes. If you look, however, at the price per ounce you will note a quantity discount as the box size increases.

[3] Volume discounts are specified sometimes in terms of a uniform price $P(q)$ applied to all the units purchased, which is declining with purchase quantity. Such a uniform price can be interpreted as the average price corresponding to the marginal price function $p(q)$ and calculated as $P(q) = \frac{1}{q}\int_0^q p(\alpha)d\alpha$.

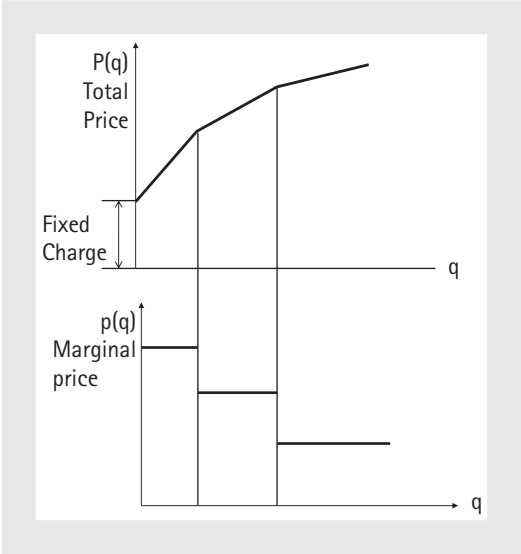

FIGURE 22.3 A block declining tariff structure.

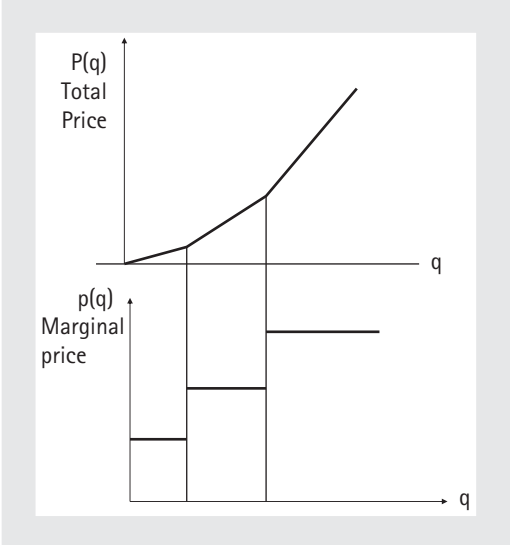

FIGURE 22.4 A block increasing tariff structure.

For some commodities, such as electricity or water, where the objective of price discrimination is to promote conservation and "tax the rich", the marginal price function is actually increasing with quantity as shown in Figure 22.4.

The lower consumption blocks that are billed at a lower per unit price are sometimes called "life line" rates. In such cases it will be to the advantage of a household to get two

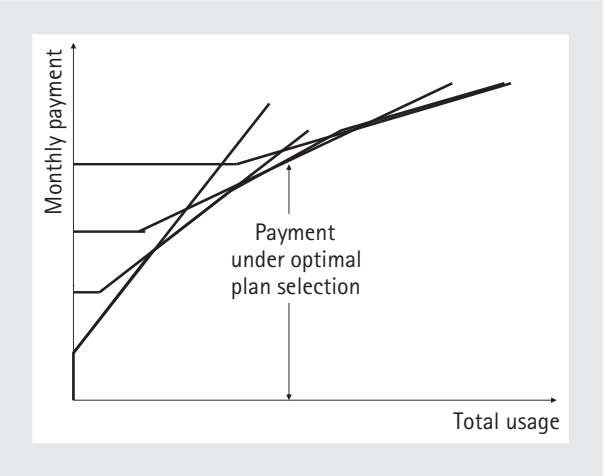

FIGURE 22.5 Implementing a volume discount policy through a menu of optional two part tariffs.

water or electricity meters and pretend to be two households, hence it is essential in order to sustain such a pricing policy to prevent splitting of the demand.

In many cases quantity discount pricing is implemented through optional two-part tariff contracts. For example, in the case of a mobile phone service a consumer can choose among several plans with increasing monthly payments and declining per minute cost (for minutes above the free ones). Figure 22.5 illustrates that when multiple two-part tariff options are offered and the customer is assumed to self-select the best plan for its usage rate then the entire price menu replicates a block declining price structure.

For analytical convenience we will assume that the price function $p(q)$ is continuous and show how it can be determined given the disaggregated demand profile $N(q, p)$. We also assume that for any quantity q, $N(q, p)$ is declining in p (fewer customers will buy the qth unit as the unit price increases, that is $\partial N(q, p)/\partial p < 0$. For any price level, p, the number of customers who will buy the qth unit declines with q, that is $\partial N(q, p)/\partial q < 0$ and the rate of decline decreases with p, $\partial^2 N(q, p)/\partial q \partial p > 0$. The last condition is a common technical assumption (often referred to as a "single-crossing") ~~that~~ guarantees that demand functions for different units q will not cross. Under the single-crossing assumption, the profile $N(q, p)$ will look as shown in Figure 22.6.[4]

Let us now consider the problem of a monopolist who wants to determine a unit price function $p(q)$ that will maximize its profit assuming that each unit costs c to produce. For simplicity, let us assume that q can only take integer values. The profit of the monopolist is given by:

$$\pi = \sum_{q=1}^{Q} N(q, p(q))(p(q) - c)$$

[4] The term "single-crossing" refers to the fact that any monotone function of p will cross the demand function corresponding to any unit q at most once.

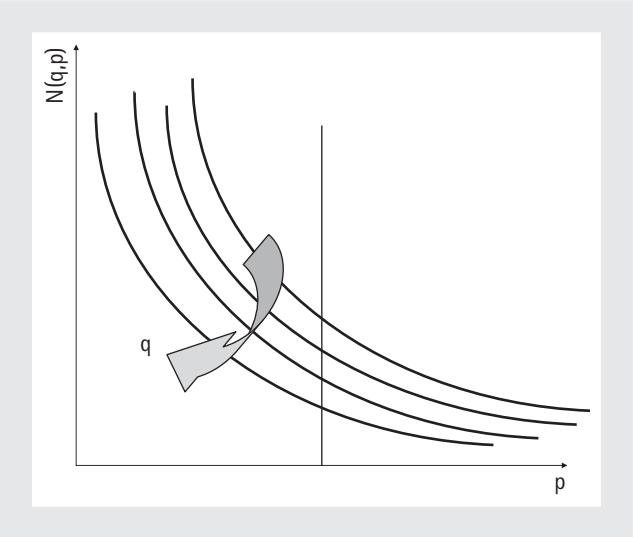

FIGURE 22.6 Demand functions for incremental purchase units.

We note, however, that this profit function is separable with respect to q so in order to maximize this function with respect to $p(q)$ we need to maximize each of the terms. Thus the necessary conditions for maximum profit are:

$$\frac{\partial}{p(q)}\{N(q, p(q))(p(q) - c)\} = 0 \text{ for } q = 1, 2, 3 \ldots$$

This gives the optimality condition:

$$(p(q) - c) \cdot \frac{\partial N(q, p(q))}{\partial p(q)} + N(q, p(q)) = 0$$

We can define the elasticity of demand for the qth unit as:

$$\varepsilon(q) = -\frac{\partial N(q, p(q))/\partial p(q)}{N(q, p(q))/p(q)} = -\frac{\partial N(q, p(q))/N(q, p(q))}{\partial p(q)/p(q)}$$

Then the optimality condition becomes:

$$\frac{p(q) - c}{p(q)} = \frac{1}{\varepsilon(q)}$$

The, so called, "inverse elasticity rule" implied by this optimality condition is that the optimal "percentage markup" for each incremental unit should be inversely proportional to the demand elasticity for that unit. The intuitive justification for this rule is that high demand elasticity entails stronger response (i.e. larger demand decrease) to the same percentage increase in price. Thus, a monopoly, that must tradeoff between reduced sales versus increased profit per sale in choosing the optimal markup, will opt for a lower percentage markup when demand is more elastic.

Let us consider now a special case where $N(q, p) = a \cdot p^{-\eta q}$, $a > 0$, $\eta > 1$
For this case

$$\varepsilon(q) = -\frac{-a \cdot \eta \cdot q \cdot p^{-\eta q - 1}}{a \cdot p^{-\eta q}/p} = \eta \cdot q$$

$$\frac{c}{p(q)} = 1 - \frac{1}{\eta q}$$

$$p(q) = \frac{c\eta q}{\eta q - 1}$$

$$\text{as } q \to \infty \quad p(q) \to c$$

The resulting price function will have the form shown in Figure 22.7.

To illustrate the win–win aspect of volume discounts, consider now a special case of the above where customers can only buy either one or two units of a product, the disaggregated demand profile $N(q, p)$ is as in the example above with the demand elasticity parameter $\eta = 2$ and marginal cost $c = 1$. In this case the monopoly will charge for the first unit $p_1 = 2$ and sell the second unit at a discounted price $p_2 = \frac{4}{3}$. The monopoly seller's total profit in this case is:

$$profit = N(1, p_1) \cdot (p_1 - c) + N(2, p_2) \cdot (p_2 - c) = a \cdot (2c)^{-2}(c) + a \cdot \left(\tfrac{4}{3}c\right)^{-4}\left(\tfrac{1}{3}c\right)$$

$$= \tfrac{a}{4}\left(1 + \tfrac{27}{64}\right) = 0.36a$$

and the corresponding total number of units sold is:

$$units = N(1, p_1) + N(2, p_2) = a \cdot (2c)^{-2} + a \cdot \left(\tfrac{4}{3}c\right)^{-4} = 0.56a$$

For comparison, suppose that the monopoly seller could not use volume discounts because it was unable to restrict resale (i.e. buyers that only want one unit can form coalitions and buy two units at discounted prices and then split them). In that case, the monopoly seller would set a uniform price for all units treating the demand as a single demand function given by:

$$x(p) = N(1, p) + N(2, p) = a \cdot (p^{-2} + p^{-4})$$

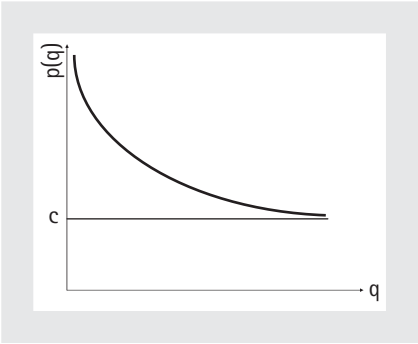

FIGURE 22.7 Optimal unit price function vs. purchase quantity.

The optimal monopoly price is still given by the inverse elasticity rule:

$$\frac{p-c}{p} = \frac{1}{\varepsilon}$$

Except that here the elasticity is based on the aggregate demand function $x(p)$ so

$$\varepsilon = -x'(p) \cdot p/x = \frac{2p^2+4}{p^2+1}$$

To determine the price, we solve the equation:

$$\frac{p-1}{p} = \frac{p^2+1}{2p^2+4},$$

which reduces to the polynomial equation $p^3 - 2p^2 + 3p - 4 = 0$ having a root at $p = 1.65$. Thus, instead of pricing the first purchase unit at $p = 2$ and the second unit at 1.33, the monopoly seller will price all units at $p = 1.65$. The total demand corresponding to that price is $x(1.65) = 0.5a$ and the total profit is given by $0.5a(1.65 - 1) = 0.325a$. So the ability to discriminate based on purchase quantity increases the monopolist profit by about 11 percent and increases social welfare since more customers who value the product above its marginal cost of production will be able to enjoy it although some will pay more for it (those who only buy one unit). It should be noted, however, that price discrimination does not always result in increased social welfare. As shown by Varian (1985), under fairly general conditions, a necessary condition for a social welfare increase due to price discrimination is an increase in output (or consumption). Thus, to the extent that a nonlinear pricing scheme could result in reduced consumption such a strategy would also reduce social welfare.

22.4 Ramsey pricing

As mentioned in the introduction, Ramsey pricing is a form of semi-direct price discrimination. Its purpose is to enforce a total profit constraint while incurring the least social cost. It is presented here because, in spite of the different motivation and apparent philosophical differences, the methodology used to derive Ramsey pricing and the end results bear remarkable similarity to those presented in the previous section in deriving optimal volume discount schedules.

Here instead of differentiating among the demands for the first, second, and third … unit of consumption the regulated monopoly seller (with the blessing of the regulator) differentiates among the demand of different customer classes; say, commercial and residential. This type of price discrimination was common in the old days when AT&T had a monopoly over long distance phone service. Suppose that the demand functions for phone call units in each customer class are given by $x_c(p)$ and $x_r(p)$, respectively.

Again if the monopoly wants to maximize total profit and is able to charge different prices to the two customer classes then the problem is separable and the optimal price charged to each class is determined by the inverse elasticity rule:

$$\frac{p_c - c}{p_c} = \frac{1}{\varepsilon_c} \text{ and } \frac{p_r - c}{p_r} = \frac{1}{\varepsilon_r},$$

where ε_c and ε_r denote the elasticity of the corresponding demand functions, $x_c(p)$ and $x_r(p)$.

The example given in the previous subsection for price discrimination between the first and second unit of purchase can be relabeled to reflect discrimination between commercial and residential demand for long distance phone calls since the commercial demand function is in general less elastic than residential demand just as the demand for the first unit of purchase was assumed to be less elastic than that for the second unit. Thus, by discriminating between the two classes of service the monopolist's profits go up, total usage increases and therefore social welfare increases (because we assume constant unit cost). Furthermore the residential customers will end up paying less for their calls while commercial customers pay more.

When the seller is a regulated monopoly (as AT&T was), the regulator may put a limit on the profits that the monopoly can earn based on the cost of investment made by the regulated monopoly in building the infrastructure. This is called rate of return regulation where the monopoly profits are limited to an annual percentage return on investment cost. In such a case, the regulated-monopoly-pricing problem is formulated as one of maximizing social welfare subject to a profit constraint.

As explained in the introduction, the social welfare resulting from consuming an incremental unit of a product or service (i.e. the surplus to society) is given by the difference between the consumers' willingness-to-pay for that unit (given by the inverse demand function $P(x)$ in Figure 22.8) and the unit's production cost c. This difference, which varies with the total consumption level x, is represented by the vertical slices shown in Figure 22.8. The purchase quantity of the product or service offered at a uniform price p^* is given by

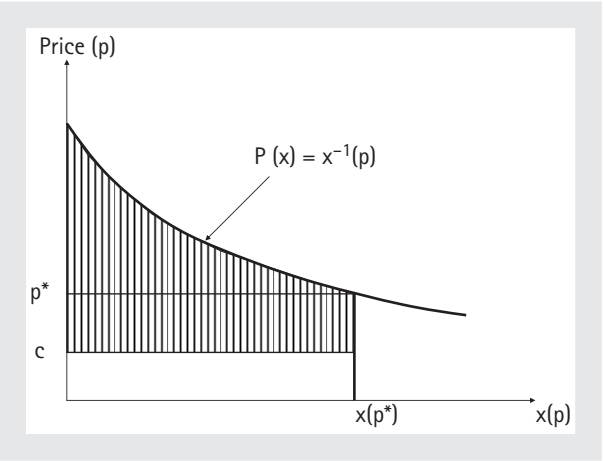

FIGURE 22.8 Illustration of social value resulting from consumption at unit price p*.

x ($p^\star$) at which point consumers' willingness-to-pay $P(x(p^\star)) = p^\star$. Hence the aggregate social welfare is given by the area between the marginal cost c and the willingness-to-pay function $P(x)$ to the left of $x(p^\star)$, as shown in Figure 22.8.

In our setting we assume that commercial and residential customers have willingness-to-pay functions $P_c(x)$ and $P_r(x)$ respectively, and the monopoly supplier is regulated so that its net profit is set to a predetermined value π (e.g. based on some allowed rate of return on capital investment). Then the optimization problem for setting the socially optimal prices p_c and p_r for the two customer classes, subject to the regulated profit constraint is:

$$\max_{p_c,\ p_r} \{ \int_0^{x_c(p_c)} (P_c(x) - c)dx + \int_0^{x_r(p_r)} (P_r(x) - c)dx$$

Subject to: $x_c(p_c) \cdot (p_c - c) + x_r(p_r) \cdot (p_r - c) = \pi$

To solve this problem we write the Lagrangian:

$$L(p_c, p_r, \lambda) = \int_0^{x_c(p_c)} (P_c(x) - c)dx + \int_0^{x_r(p_r)} (P_r(x) - c)dx + \lambda\{x_c(p_c) \cdot (p_c - c) + x_r(p_r) \cdot (p_r - c) - \pi\}$$

This Lagrangian is separable so the optimality conditions are given by

$$\frac{\partial}{\partial p_i}\{ \int_0^{x_i(p_i)} (P_i(x) - c)dx + \lambda \cdot x_i(p_i) \cdot (p_i - c)\} = 0 \quad i = \{c, r\},$$

and the profit constraint. This reduces to:

$$(p_i - c) \cdot x_i'(p_i) + \lambda \cdot (p_i - c) \cdot x_i'(p_i) + \lambda \cdot x_i(p_i)\} = 0 \quad \text{for} \quad i = \{c, r\},$$

which can be rewritten as

$$\frac{p_i - c}{p_i} = \frac{\lambda/(\lambda + 1)}{\varepsilon_i} \ \ i = \{c, r\}.$$

The above result tells us that the optimal regulated monopoly prices should be set so that the percentage markup in each customer group is proportional to the inverse elasticity. In other words, the more elastic the demand is, the lower the markup should be (price that the market will bear). The Lagrange multiplier factor $\lambda/(\lambda + 1)$, which scales the percentage markup is determined so the profit constraint is satisfied.

In other words the ratio of percentage markup rule among classes of customers in the regulated monopoly problem is the same as in the profit maximizing monopoly problem (and the same as in an oligopoly) but the prices are different because of the profit constraint. This pricing rule is called the *Ramsey pricing* rule. The intuition behind this

rule is that social welfare is affected by consumption. Transfer of money between members of society does not affect the social welfare of society as a whole. Therefore, if we need to generate a certain level of profit so as to recover the supplier's investment costs and fair return on capital, we charge more to those customers whose demand will be affected the least by a higher price.

22.5 Quality differentiation

In this section, we will discuss nonlinear pricing that is based on differentiating products or services so as to exploit customers' heterogeneous preferences for specific product attributes. Such differentiation can be based on exogenous product characteristics such as speed, convenience, and packaging, or can be induced through pricing that results in self-segmentation or rationing schemes that create supply uncertainty for the service or product.

22.5.1 Pricing exogenous quality attributes

Quality differentiation in the context of nonlinear pricing is done through unbundling quality attributes of products or services for which customers have heterogeneous preferences, for the purpose of market segmentation and indirect price discrimination. Typical unbundled attributes include product features, packaging, distribution channels, or delivery conditions such as time of use, class of service in airlines, speed of delivery in mail service, bulk versus retail.

The basic idea is to capitalize on the dispersion in customer preferences (i.e. willingness-to-pay for the different attribute levels) and create an offering that gives customers a tradeoff between attribute level and price. In general not all customers rank attribute options in the same way. For example, choice between points of delivery will be ranked differently by customers based on where they live. Similarly time of use of a service may be ranked differently by different customers. Location and time of use fall under the general category of locational attributes. On the other hand, attributes such as speed of mail delivery, priority of service in a queue, or comfort levels in a plane are ranked the same by all customers even if they differ in how much they are willing to pay for different levels of these attributes. Attributes for which customers have the same preference rankings are called "quality attributes". A general property of quality attributes is that they are "downward substitutable", that is you can always use a higher quality level to serve demand for a lower quality level. For instance, a 2 GHz processor can always replace a 1 GHz processor in a computer and a first class seat in a plane can be used to accommodate a customer that paid for a coach seat.

We characterize the quality dimension by a parameter, s, so that a larger value of s represents higher quality. For example, s may represent speed of delivery for mail service (defined as the inverse of time en route), or the speed of memory chips. In general, quality can be multi-dimensional, reflecting different aspects of a product or service affecting customers' preferences. In this chapter, we restrict ourselves to a single quality dimension to simplify the exposition. The demand function is characterized by a function, $N(s, p)$ that

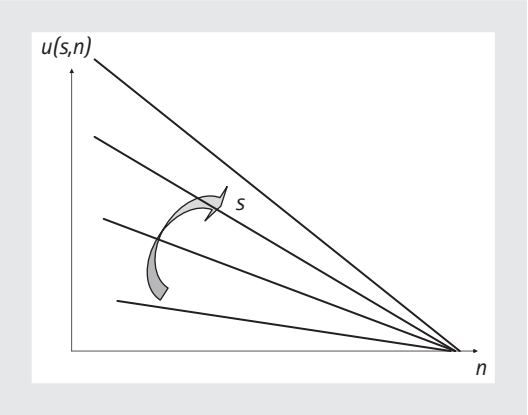

FIGURE 22.9 Illustration of inverse demand function (utility) for different quality levels.

defines the demand function for quality s as function of price, if that was the only product quality offered. Alternatively, we can arrange the units of demand in decreasing order of willingness-to-pay and define the inverse demand function, $u(s, n)$, so that $N(s, u(s, n)) = n$.

Figure 22.9 illustrates the inverse demand functions for different quality levels. We assume that the inverse demand functions satisfy the following properties:

$$\frac{\partial u(s,\ n)}{\partial n} < 0,\ \frac{\partial u(s,\ n)}{\partial s} > 0,\ \frac{\partial u(s,\ n)}{\partial s \partial n} < 0.$$

These inequalities imply that willingness-to-pay for any quality level decreases with n (we sort the customers so that this is true). Willingness-to-pay by any customer n increases with quality and the sensitivity of customers to quality decreases with n. The last condition is again a "non-crossing" condition ensuring that the demand functions for different quality levels do not cross. The commonly used multiplicative utility function form $u(s, n) = g(s) \cdot w(n)$, where $g(s)$ is increasing and $w(n)$ is decreasing, is a special case that satisfies these properties. Figure 22.9 illustrates the case where $w(n)$ is linear.

Suppose that a discrete set of quality levels, $s_1 > s_2 > , \ldots . > s_k$, is being offered at prices, $p_1 > p_2 > , \ldots . > p_k$. Thinking of each unit n of demand as a separate customer, we can write the so-called self-selection and individual rationality conditions for customer n as:

$$i(n) = \arg\max_i \{u(s_i, n) - p_i\}$$

and

$$u(s_{i(n)}, n) - p_{i(n)} > 0$$

These conditions state that each customer n selects the quality level $i(n)$ that maximizes his surplus (defined as utility minus price) provided that the surplus is positive otherwise no product is chosen yielding zero surplus. These conditions are illustrated graphically in Figure 22.10.

The customers will divide themselves among the different product qualities by selecting the quality that maximizes their surplus. Thus the demand for each quality level s_i is given

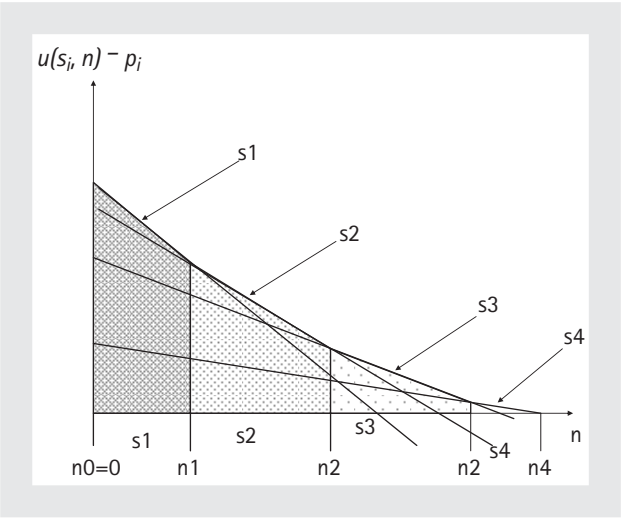

FIGURE 22.10 Customers' maximum surplus for different quality levels.

by the difference, $(n_i - n_{i-1})$, where the boundary points, n_i, are defined by the indifference relations $u(s_i, n_i) - p_i = u(s_{i+1}, n_i) - p_{i+1}$, $i = 1, 2, \ldots k$, and the individual rationality condition $u(s_i, n_i) - p_i \geq 0$. We may assume that s_{k+1} is a default free quality level for which the utility of all customers is zero. The above model characterizes the demand function and cross substitution among the different quality levels. A monopoly offering a product line consisting of quality levels $s_1 > s_2 > , \ldots . > s_k$ with corresponding unit costs $c_1 > c_2 > , \ldots . > c_k$ can determine the profit-maximizing prices for each quality level by solving the optimization problem:

$$\max_{p_1, p_2, \ldots, p_k} \sum_{i=1}^{k} (n_i - n_{i-1})(p_i - c_i)$$

$$s.t.$$

$$u(s_i, n_i) - p_i = u(s_{i+1}, n_i) - p_{i+1}, \; i = 1, 2, \ldots, k$$

$$u(s_i, n_i) - p_i \geq 0, \; i = 1, 2, \ldots, k$$

The above framework can also be used to solve the problem of a supplier that wants to introduce a new product offering a new quality level in a market that is already divided between existing quality levels serving the demand. For example, it would apply to a provider of two-day delivery service in a market already served by cheap US Postal Service and FedEx, which offers next day delivery at a much higher price. In that case, the supplier of the new service can solve the above optimization problem to determine its optimal price, while taking the prices of the existing quality levels as given. Interestingly, he only needs to consider the qualities adjacent to his since all the other terms in the objective function and constraints are not affected by his decision and will drop out of the optimization.

Pricing of a product line consisting of quality-differentiated products has been extensively addressed in the revenue management (RM) literature. However, the traditional RM approach usually characterizes the demands for different product variants or quality levels as exogenous independent stochastic processes. Modeling cross-substitution among different products based on the underlying customer choice behavior is a relatively recent trend in the RM literature, pioneered by Talluri and van Ryzin (2004). While these efforts have yet to make the connection and capitalize on the rich literature on multiproduct pricing, in marketing science and economics, this is a promising development. Characterizing the customer preference structure underlying the demand for variants of differentiated products is essential for understanding the impact of relative prices and how a new entry might impact an existing product line.

22.5.2 Price induced endogenous qualities

Differential quality of service can sometimes be created by inducing customers to sort themselves through differential pricing in situations where quality is affected by the demand, for example through congestion. To illustrate such phenomena, consider a situation where 100 customers need to be served, each taking 1 minute. All arrive at once and are served at random by two servers that charge $2 per customer. Assume this is the prorated cost of providing the service which in total costs $200. The average waiting time of each customer is 25 minutes. Let us assume now that the customer population consists of 75 students whose time is worth $6/hour and 25 professors whose time is worth $60/hour. Table 22.2 summarizes the costs and benefits incurred by each of the customer types, the supplier and society as a whole under the random service policy.

Suppose now that we offer service at one server for free while the other server charges $10 per customer. We do not restrict access to any of the servers but provide a forecast of an equilibrium average waiting time of 12.5 minutes for the $10 server and 37.5 minutes for the free server. Customers will self-select which server they want to use based on the calculation in Table 22.3. Accordingly, students will self-select the free line while professors will select the $10 server, so the waiting time forecast will be realized and everyone is better off than before.

Price induced quality differentiation is common in pricing products and services where customers incur personal cost in addition to the tariff (e.g. waiting time cost). It has been proposed, for instance, as a mechanism for increasing the utilization of underutilized carpool lanes on the freeways by allowing drivers to buy permits for these lanes at high prices (in addition to permits for carpools and gasoline efficient cars). To some extent such

Table 22.2 Costs under uniform price

	Students	Profs	Supplier	Society
Cost	$6 \times 25/60 = \$2.5$	$60 \times 25/60 = \$25$	$200	
Charge	$2	$2	($200)	
Total	$4.50	$27	0	$4.5 \times 75 + 27 \times 25 = \$1012.5$

Table 22.3 Cost distribution and service qualities under dierential pricing

	Students		Professors		Supplier	Society
	Serv A	Serv B	Serv A	Serv B		
Cost	$6 \times 12.5/60 = \$1.25$	$6 \times 37.5/60 = \$3.75$	$60 \times 12.5/60 = \$12.5$	$60 \times 37.5/60 = \$37.5$	\$200	
Charge	\$10	0	\$10	0	(\$250)	
Total	\$11.25	\$3.75	\$22.50	\$37.50	(\$50)	$3.75 \times 75 + 22.5 \times 25 - 50 = \$793.75$

a policy is implicitly implemented through enforcement policies that determine the probability of a citation for illegal use of carpool lanes, setting the price to the expected value of the fine. Student nights at movie theaters at reduced ticket prices is another example of price induced quality.

22.5.3 Rationing-based quality differentiation

When the supply of a product is limited by scarcity or limited capacity, it is possible to use supply uncertainty as a mechanism for quality differentiation. Such an approach is particularly useful when the demand function is such that using a single price will result in monopoly prices that underutilize available supply. This may occur when the profit function as a function of supply quantity is non-monotone so that the monopoly supplier may be induced to withhold available capacity.

Consider a promoter of a rock concert in a sports arena that can accommodate 10,000 people. The cost of putting up the event is \$300,000. Market research data suggest that the market for such an event consists of two segments. There are about 5,000 customers in the area who will pay up to \$100 per ticket and another 55,000 potential customers who are willing to pay up to \$20 per ticket. It is impractical to have assigned seats so a simple option is to have a uniform price for all tickets. If the price is set at \$100 per ticket 5,000 tickets will be sold at a net profit, after covering expenses, of \$200,000. The corresponding social welfare as measured by willingness-to-pay minus cost is also \$200,000. Alternatively, if ticket prices are set so as to fill up the venue they can be sold at \$20 on a first come first serve basis over the internet. This pricing scheme will make some people happy but will result in a \$100,000 loss for the promoter. Furthermore, at \$20 per ticket the chance of any customer getting a ticket is on average 1/6 so the expected social welfare of such a strategy is $(5000 \times 100 + 55{,}000 \times 20)/6 - 300{,}000 = -\$33{,}000$. Clearly the first option of pricing the tickets at \$100 is superior both from a profit-to-promoter and a social welfare perspective. However, the thought of having half the venue empty while there are 55,000 potential customers out there willing to pay \$20 per tickets is bothersome.

Figure 22.11 illustrates the demand function and revenue function, which create the dilemma faced by the promoter. The important aspect of that revenue function is its non concavity in the region where the available capacity falls.

The solution to the promoter's dilemma is to introduce two types of tickets: reserved tickets at \$90 and lottery tickets at \$20. All the reserved tickets can then be sold in advance

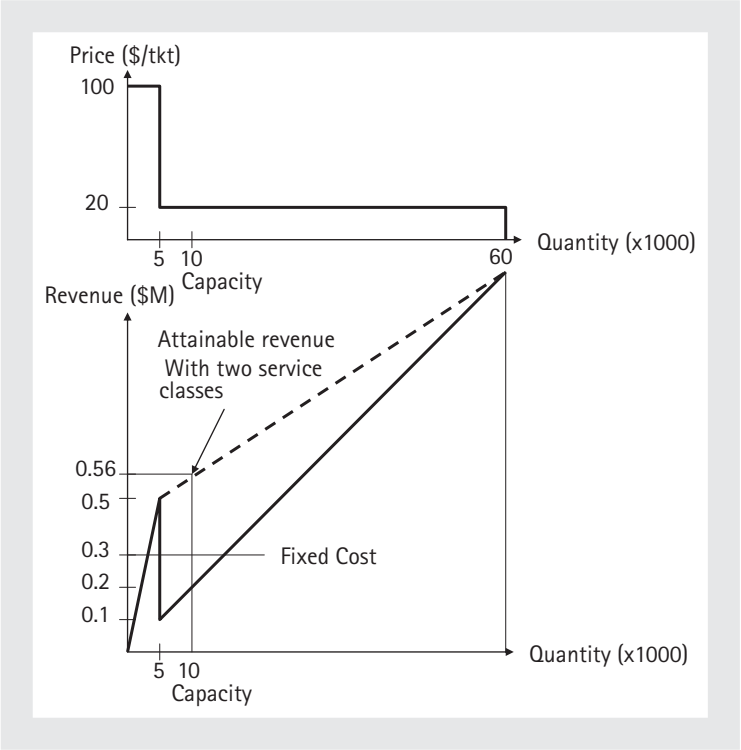

FIGURE 22.11 Demand and revenue functions for tickets.

to the customers who are willing to pay $100 while the rest of the tickets are released on the internet the day before the show at $20 with an average probability of 1/11 of getting one. For this to work, however, one must assure nontransferability of the demand of the potentially high paying customers, that is induce such customers not to opt for the cheap tickets. This is guaranteed by the above prices since $100 - 90 > (100 - 20)/11$ so that a customer whose willingness-to-pay is $100 will maximize his/her expected utility by purchasing the reserved ticket, while customers who are willing to pay $20 will compete for the standby tickets (perhaps we should give them a break and sell the tickets for $19.) With this strategy the promoter can increase its profits and the social welfare by $100,000 and make some additional customers happy.

Ferguson (1994) provides an elegant proof showing that the above approach will increase the monopolist profit whenever the profit function is non-concave and the capacity limit falls in a rising non-concave portion of the profit. Figure 22.12 illustrates the profit as a function of quantity sold at a single price taking into consideration that the price that will sell quantity q is given by the inverse demand function $P(q)$. The dashed line shows the attainable profit when we introduce a second offering with uncertain delivery.

To achieve the higher profit we will offer q_1 units with guaranteed supply at a price p_1 and offer the remaining $Q - q_1$ units on a lottery basis at a price $p_2 = P(q_2)$. The quantities q_1 and q_2 are exactly the tangency points on the curve. This can be proven by starting with

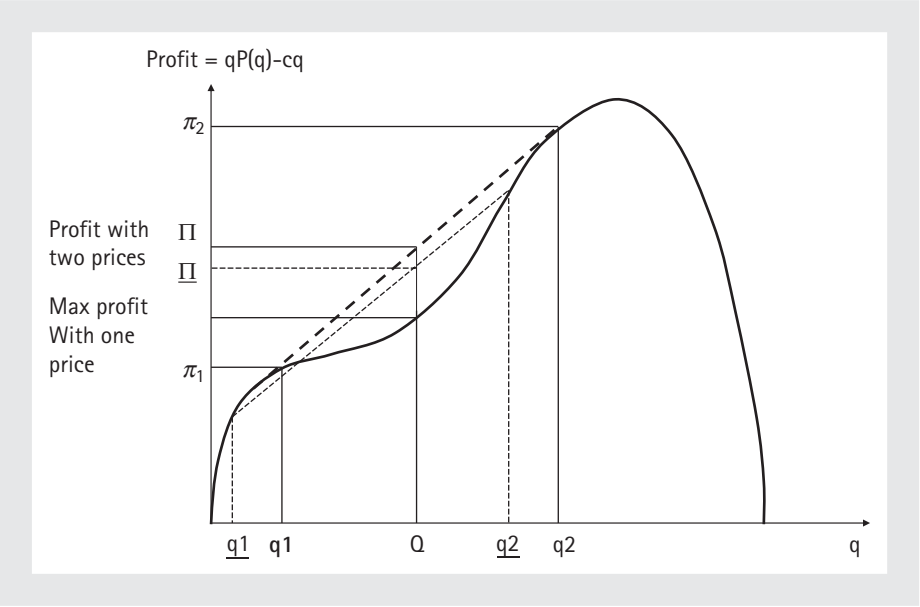

FIGURE 22.12 Improving profits by introducing a product with uncertain delivery.

arbitrary quantities, q_1 and q_2, and maximizing the profit function with respect to these quantities. The tangency condition follows from the optimality conditions. Given the above structure, the probability that a standby customer gets the product is given by $r = (Q - q_1)/(q_2 - q_1)$ so now we can calculate p_1 so that the first q_1 customers will prefer the guaranteed supply option, that is $P(q_1) - p_1 \geq r \cdot (P(q_1) - p_2)$. The monopolist will want to set p_1 as high as possible. Thus, $p_1 = (1 - r) \cdot P(q_1) + r \cdot P(q_2)$ and the corresponding total profit is therefore:

$$\Pi = p_1 \cdot q_1 + p_2 \cdot (Q - q_1) - c \cdot Q = p_1 \cdot q_1 + r \cdot p_2 \cdot (q_2 - q_1) - c \cdot Q$$

$$[(1 - r) \cdot P(q_1) + r \cdot P(q_2)] \cdot q_1 + r \cdot P(q_2) \cdot (q_2 - q_1) - c \cdot Q$$

Therefore, $\Pi = (1 - r) \cdot \pi_1 + r \cdot \pi_2$

where r is such that $Q = (1 - r) \cdot q_1 + r \cdot q_2$

The above derivation is valid for arbitrary values of q_1 and q_2, not just the tangency points (see dashed lines in Figure 22.12) but it is easy to see from the figure (or prove algebraically) that choosing the tangency points maximizes the supplier's profits. We may further conclude that such a strategy is beneficial only if the available capacity falls in a region where the profit function is increasing and there is a gap between the profit function and its concave hull. Under such circumstances, a single product with uncertain delivery will suffice to attain the potential profit, given by the concave hull of the original profit function at full capacity utilization.

22.6 Priority service pricing and efficient rationing

In the previous section we introduced the idea of quality differentiation through uncertain supply when capacity is scarce and fixed. This basic concept is expanded by *priority pricing*. This pricing mechanism is a quality differentiation and enables an efficient rationing in situations where supply is both scarce and uncertain. It enables customers to pay different prices based on the order in which they are served or probability of getting the product. In the case of electricity supply, for instance, customers can sign up for an option of being curtailed when supply is scarce in exchange for a discount on their electricity bills. Another example of priority pricing is the practice of the discount clothing store Filene's Basement, which posts on each item a series of increasing percentage discounts on the item and the date on which each discount level will go into effect. Customers must trade off the option of a larger discount against the probability that someone else will purchase the item they want.

The basic principle is that an efficient priority-pricing scheme will result in customers being served in order of willingness-to-pay. Therefore, under efficient rationing the qth unit of demand is served if and only if the available supply is q or larger. Therefore, the probability that the qth unit is served $r(q) = 1 - F(q)$ where $F(q)$ denotes the cumulative probability that the available supply level is q. We assume now that each unit of demand corresponds to a customer demanding one unit and the inverse demand function representing the willingness-to-pay of customer q for the product is given by $v(q)$. Since the demand is monotone in q we can, without loss of generality, characterize customers in terms of their valuation v and define directly the probability of service for a customer with valuation v as $r(v) = 1 - F(q(v))$ where $q(v)$ is the demand at price v. Since the supplier does not know how much a particular customer is willing to pay for the product all he can do is set prices based on probability of delivery or equivalently the place in line for delivery. Thus, the price structure will be of the form $P + p(r)$ where P is a uniform fixed charge applied to all customers and $p(r)$ depends on the probability of service selected by the customer. The challenge here is to design the price function to induce each customer v to select her designated efficient service priority $r(v)$. The self-selection condition and individual rationality condition for customer q are:

$$\max_r \{r \cdot v - P - p(r)\}$$
$$r \cdot v - P - p(r) \geq 0$$

Customers whose optimal r does not satisfy the second condition will not buy the service. We assume in the above formulation that a customer pays even if she does not get the service but the formulation can be easily changed so that payment is made only if service is obtained.

The first order necessary condition for the customer's self-selection is: $dp(r)/dr = v$ and we want to induce the customer to select $r = r(v)$. We will determine the price function $p(r)$ indirectly by first defining $\hat{p}(v) = p(r(v))$. Thus,

$$\frac{d\hat{p}(v)}{dv} = \frac{dp(r)}{dr} \cdot \frac{dr(v)}{dv} = v \cdot \frac{dr(v)}{dv}$$

so,

$$\hat{p}(v) = \int_0^v \omega \cdot dr(\omega) = v \cdot r(v) - \int_0^v r(\omega) d\omega$$

The price function can now be obtained as $p(r) = \hat{p}(v(r))$ where $v(r)$ is the inverse of the function $r(v)$. Note that the expected social surplus from offering priority $r(v)$ to the customer with valuation v is given by $v \cdot r(v)$. Out of this total surplus the supplier collects $P + \hat{p}(v)$. From the individual rationality condition a customer will buy only if,

$$v \cdot r(v) - P - \hat{p}(v) = \int_0^v r(\omega) d\omega - P \geq 0.$$

Thus, the fixed charge can be mapped onto a cutoff value v_o so that

$$P = \int_0^{v_o} r(\omega) d\omega.$$

Customers with valuation below v_o are excluded and $r(v_o)$ is the lowest probability of service being offered. The consumer surplus to a customer with valuation v under this scheme is

$$CS(v) = \int_{v_o}^v r(\omega) d\omega$$

Figure 22.13 illustrates the distribution of surplus between the supplier and the customer.

So far, we have characterized the pricing scheme that will induce efficient rationing through self-selection. The only degree of freedom in that price structure is the fixed charge, P, which determines the cutoff level for customers that will be served. This level can be set based on the objective of the supplier, whether it is to maximize social welfare, recover the supply cost, or maximize profit in the case of a monopoly.

To illustrate the implications of the above results we now specialize them to the case where the probability of supply is described by a uniform distribution on $[o, Q]$,

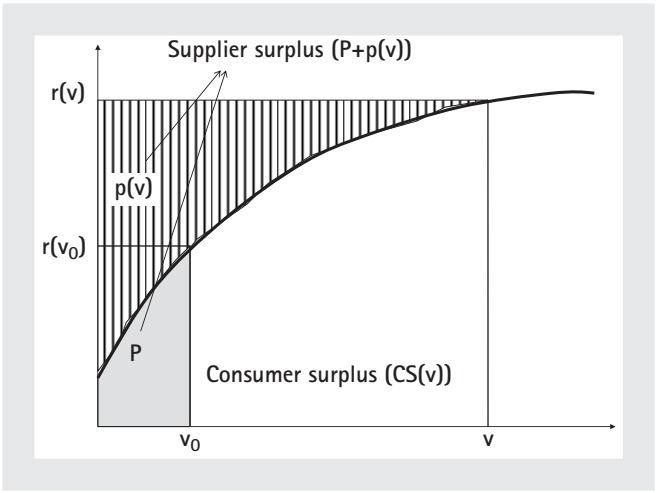

FIGURE 22.13 Allocation of the social surplus $v \cdot r(v)$ due to serving a unit with valuation v.

that is $F(q) = q/Q$ and the inverse demand function is given by $v(q) = 1 - q/Q$. This implies that the probability of having enough supply to serve a customer with valuation v under an efficient rationing scheme is $r(v) = v$. Plugging this into the above results gives:

$$\hat{p}(v) = v \cdot r(v) - \int_0^v r(\omega)d\omega = v^2 - \frac{v^2}{2} = \frac{v^2}{2}.$$

Hence $p(r) = r^2/2$, $P = v_0^2/2$ And the fraction of served demand is $q_0/Q = 1 - \sqrt{2P}$. The total supplier revenue is given by

$$\Pi = P \cdot q_0 + \int_0^{q_0} p(r(v(q))dq = P \cdot q_0 + \int_0^{q_0} \frac{(1 - q/Q)^2}{2} dq = P \cdot q_0 - \tfrac{Q}{6}[1 - \tfrac{q}{Q}]^3|_0^{q_0}$$

$$= P \cdot Q\left(1 - \sqrt{2P}\right) - \tfrac{Q}{6}\left[\left(\sqrt{2P}\right)^3 - 1\right] = Q \cdot \left[P - \tfrac{4}{3}P\sqrt{2P} + \tfrac{1}{6}\right]$$

Maximizing the profit with respect to ~~F~~ P yields ~~F~~ $P = 1/8$ and consequently $q_0/Q = 1/2$, that is the optimal strategy of a monopolist is to price out half of the demand by imposing a fixed charge $P = 1/8$ and a priority charge $p(r) = r^2 / 2$ for values of r between 0.5 and 1. The monopolist profit will then be $\Pi = \frac{5Q}{24}$

The total social welfare is given by

$$SW = \int_0^{q_0} v(q) \cdot r(v(q))dq = \int_0^{q_0} (1 - q/Q)^2 dq = -\tfrac{Q}{3}[1 - \tfrac{q}{Q}]^3|_0^{q_0} = \tfrac{Q}{3}\left[1 - \left(\sqrt{2P}\right)^3\right].$$

Thus the social welfare under a monopoly regime is $SW_m = \frac{7Q}{24}$ and consequently the total consumer surplus is $CS_m = \frac{Q}{12}$.

A social welfare maximizing entity, however, will impose no fixed charge so that no customer is excluded (this is often called Universal Service), achieving a social welfare of $SW = \frac{Q}{3}$ but customers will still be charged a priority price $p(r) = r^2/2$, which yields a profit of $\frac{Q}{6}$ (substitute $P = 0$ in the profit formula above). An interesting question is whether a universal service scheme with priority pricing is better for consumers then free universal service with random rationing. To address this question we compare the individual consumer surplus for both cases. With free random rationing, every customer has a probability $R = 1/2$ of being served and there is no charge. In that case, a customer with valuation v gets an expected benefit of $v/2$. With priority service, a customer with valuation v gets an expected consumer surplus of $v \cdot r(v) - p(r) = v^2/2$ (since under efficient rationing $v = r$). Thus the net gain in consumer surplus from priority pricing for a consumer with valuation v is $v(v-1)/2$ which is negative for all v in the interval [0,1] so all customers are worse off.

If the social welfare maximizer is a cooperative that returns all profits to the consumers as a uniform dividend, then allocating the profit of $\frac{Q}{6}$ to the Q units of consumption results in a dividend of $\frac{Q}{6}$ per unit and a net consumer surplus gain (over the free universal service approach) of $v^2/2 - v/2 + 1/6$. This net gain attains its minimum at $v = 1/2$. In other words, the least advantaged customer is the one with valuation 1/2 who will receive the same service reliability of 1/2 with both approaches. For that customer, the net gain in consumer surplus is $\frac{1}{8} - \frac{1}{4} + \frac{1}{6} = \frac{1}{24}$. Therefore, all customers are better off with the revenue neutral priority service approach. The above result was shown by Chao and Wilson (1987) to be true in general not just for uniform distributions.

The social welfare for free universal service with a single priority of service (i.e. uniform service) is given by:

$$SW_1 = \int_0^Q R \cdot v(q)dq = Q\tfrac{1}{2} \cdot \tfrac{v^2}{2}\big|_0^1 = \tfrac{Q}{4}$$

Hence, the social welfare loss due to inefficient rationing of a uniform service is $\frac{Q}{3} - \frac{Q}{4} = \frac{Q}{12}$ which represents a 25 percent efficiency loss.

So far, we have considered a continuum of priorities but in practice we may be able to segment customers into a limited number of discrete priorities. One question is how much of the welfare gains from priority service we lose if we only have a discrete number of priority classes. We start by segmenting the customers into two halves $v \in [0, \frac{1}{2}]$ and $v \in [\frac{1}{2}, 1]$, and offer to the first (lower valuation group) probability of service $R_1 = \frac{1}{4}$ and to the second (higher valuation group) probability of service $R_2 = \frac{3}{4}$. This is feasible since the average probability of service across all customers is $\frac{1}{2}$ which is how much the system can provide. To enforce such market separation through self-selection we will charge the low priority group a uniform price p_1 and the high priority group a higher price p_2. Incentive compatibility and individual rationality conditions require that:

$$\text{for } v \in [0, \frac{1}{2}],\ v \cdot R_1 - p_1 \geq 0 \text{ and } v \cdot R_1 - p_1 \geq v \cdot R_2 - p_2$$

$$\text{for } v \in [\frac{1}{2}, 1],\ \ v \cdot R_2 - p_2 \geq 0 \text{ and } v \cdot R_2 - p_2 \geq v \cdot R_1 - p_1$$

Since the lowest value customer in the low priority group has valuation zero we must have $p_1 = 0$. Then we can determine p_2 by applying the incentive compatibility condition to the boundary customer with valuation $v = 1/2$ who will be indifferent between getting the higher reliability at the higher price or the lower reliability at the lower price. Thus, $\frac{1}{2} \cdot \frac{1}{4} - p_1 = \frac{1}{2} \cdot \frac{3}{4} - p_2$, which results in $p_2 - p_1 = \frac{1}{2}(\frac{3}{4} - \frac{1}{4}) = \frac{1}{4}$ so $p_2 = \frac{1}{4}$. One can easily verify that these prices satisfy the incentive compatibility and individual rationality constraints above.

Now let us calculate the social welfare of the two priority schemes and compare it to the free universal service approach and the continuous priority pricing. Denote the social welfare corresponding to the continuous priorities as $SW_\infty = \frac{Q}{3}$ (infinite number of priorities), and as shown above, for the single priority $SW_1 = \frac{Q}{4}$.

For the two priority cases the social welfare is given by:

$$SW_2 = \int_0^{Q/2} R_2 v(q)dq + \int_{Q/2}^{Q} R_1 v(q)dq = Q\left[\tfrac{1}{4} \cdot \tfrac{v^2}{2}\Big|_0^{1/2} + \tfrac{3}{4} \cdot \tfrac{v^2}{2}\big|_{1/2}^1\right] = \tfrac{10Q}{32}$$

Thus the relative welfare loss of the two priority cases as compared to the single priority case is:

$$\frac{SW_\infty - SW_2}{SW_\infty - SW_1} = \frac{\frac{Q}{3} - \frac{10Q}{32}}{\frac{Q}{3} - \frac{Q}{4}} = \frac{1}{4} = \frac{1}{2^2}$$

In other words, going from one to two priorities reduced the welfare loss by a factor of 4. Using the approach described above it is possible to extend the result to n priorities and show that for the special case studied above the relative welfare loss for n priorities is $1/n^2$. In simple terms, the above implies that with two priorities we can capture 75 percent of the welfare gains achievable with an infinite number of priorities and with three priorities 91

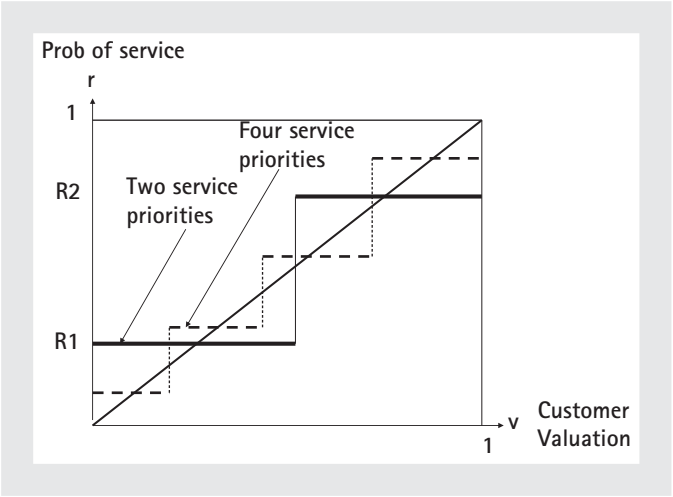

FIGURE 22.14 Priority service pricing with discrete priority levels.

percent of the achievable gain. Figure 22.14 illustrates the market segmentation with discrete priority service. Note that in deriving the welfare loss above we assumed equal partitioning of the demand into priority classes which is optimal when the valuations are uniformly distributed. In general, however, the optimal partitioning may be non-uniform and can be optimized to achieve maximum efficiency gains. Chao and Wilson (1987) have shown that in general the welfare loss with n discrete priority level is of order $1/n^2$.

22.7 Concluding remarks

In this chapter, we described various nonlinear pricing schemes that exploit disaggregated demand data and revealed heterogeneity of customers' preferences. The common theme in the methodological treatments presented is a strong emphasis on modeling the preference structure that underlies the demand heterogeneity. This approach is based on a vast literature in economics of information and game theory dealing with price discrimination, mechanism design, principal agent theory, and incentives. For ease of presentation, all examples and theory were presented for the cases where a customer's heterogeneity can be characterized by a single dimension. However, the theory can be generalized to multi-dimensional customer types as shown in Wilson (1993).

The above approach differs from the growing body of literature on revenue management that takes the heterogeneous demand as given (but subject to stochastic variations) and focuses instead on more detailed modeling on the supply side which is typically modeled simplistically in the aforementioned economics literature. Supply side aspects such as inventories and the news-vendor problem, have been typically abstracted in the economics literature dealing with mechanism design and in the nonlinear pricing

literature. On the other hand, by not modeling the underlying structure of the demand side, the revenue management literature has been limited ~~to~~ addressing the problem of cross-impact among existing products and pricing of new products attempting to penetrate existing markets. Recent work by Talluri and van Ryzin (2004), Su (2007), and by Lutze and Ozer (2008) are good examples of an emerging trend to bridge the gap between the two approaches. Such research should continue to develop models that have realistic representations of supply side aspects along with a fundamental representation of preference structures and incentives on the demand side, which drive the demand for diverse products and services.

From a practical applicability perspective, sophisticated nonlinear pricing schemes have become technologically feasible in many service industries due to the proliferation of advanced metering and control technologies at low cost. In the electric power industry, for instance, we are witnessing massive deployment of smart meters that will facilitate demand response through price incentives and contracted load control options that enable differentiation of service quality. Opportunities for facilitating load response through nonlinear pricing schemes have also spurred new business opportunities for retail intermediaries (often referred to as aggregators) that package load control options into wholesale products that are offered to the grid operator as operating reserves or offered into the balancing market auction (see ~~the chapter by Robert Wilson~~ Chapter 4 ~~in this book on electricity markets~~). In the airline industry, nonlinear pricing has been common and enabled by the technological advances in online reservation systems (see Chapter 3 by Barnes). Likewise telecom services such as mobile phone services are provided with a multitude of billing and service options (see Chapter 9 by Zimmerman). Nonlinear pricing methods have also become more prevalent in retail over the past two decades largely due to sophisticated scanning and penetration of radio-frequency identification (RFID) tagging that supports modern inventory management and automatic "mark down" policies.

Given the technical feasibility of nonlinear pricing approaches, an open question for practitioners is how much differentiation is appropriate when taking into account the ability of consumers to process information and possible adverse reaction to what may be perceived as unstable prices. Some pricing policies such as real time pricing of electricity face political scrutiny and in the telecom industry we are witnessing a return to tariffs that provide unlimited service at flat rates. Theoretical models of customer choice, traditionally used in the economics and marketing literature, often assume that customers are perfectly rational and have unlimited computational capabilities. However, a growing literature in behavioral economics (see Camerer et al. 2004 and Chapter 20 by Özer and Zheng) suggests that customers' rationality and ability to determine their optimal choice are limited while human judgment is affected by numerous biases that can be manipulated. Future research on nonlinear pricing accounting for customers' preferences and strategic choice behavior should attempt to integrate new empirically validated models of choice behavior emerging from the rapidly growing field of behavioral economics. Such research will hopefully provide insight and practical guidance with regard to tradeoffs between the pursuit of efficiency versus realistic limitation on product variety and pricing complexity in designing nonlinear pricing schemes.

22.8 Bibliographical notes

The purpose of this section is to provide a brief historical perspective and some key references that were omitted in the text for sake of continuity. This bibliographic review is by no means comprehensive and the reader is referred to the book by Wilson (1993) for a more complete review of the literature.

The theory of price discrimination dates back to Pigou (1920). Cassady (1946a,b), Phlips (1983), and Varian (1985) provide detailed reviews and interpretations of the theory and practice of price discrimination. The example of bundling given in this chapter is due to Stigler (1963). The analysis of two-product bundling with continuous willingness-to-pay is due to Adams and Yellan (1976). Optimal two-part tariffs, which represent the simplest form of quantity-based nonlinear pricing, were analyzed by Oi (1971) and many others. The analysis of nonlinear pricing for continuous quantities has been influenced primarily by Mirrlees' (1971) work on optimal taxation, which is rooted in the work of Ramsey (1927). A sample of key contributions and expositions addressing the optimal structure of quantity based nonlinear tariffs under alternative competitive conditions, their welfare implications and various extensions of the theory include: Brown and Sibley (1986), Goldman et al. (1984), Katz (1982), Mirman and Sibley (1980), Oren, Smith and Wilson (1983,1984,1985), Roberts (1979), Spulber (1981), Stiglitz (1977), Willig (1978). The derivation of quantity-based nonlinear tariffs using profile function, used in this chapter, is due to Wilson (1993). This derivation is more transparent since it avoids the use of customers' utility functions parametric on customer type, which is the common approach in the literature. Laffont et al. (1987) and Oren et al. (1985) developed special cases of nonlinear pricing when customers' types are multi-dimensional. Early contributions to nonlinear pricing in the management science and marketing science literature began in the mid-1980s, including Jucker and Rosenblatt (1985), Monahan (1984), Moorty (1984), Lal and Staelin (1984), Braden and Oren (1994). One of the early contributions to the analysis of quality differentiated nonlinear pricing is Mussa and Rosen (1978) which focuses on the welfare implication of such differentiation by a monopoly supplier. Subsequent work by Oren et al. (1982, 1987), Chao et al. (1986) and by Smith (1986, 1989) emphasizes the development and applications of quality differentiated price schedules, particularly in the context of electric power service and high tech products. The recent books by Talluri and van Ryzin (2005) and by Phillips (2005) provide an extensive review of the alternative treatment of quality differentiated pricing in the growing revenue management literature. Marchand (1974) and Tschirhart and Jen (1979) describe the early analysis of interruptible electricity pricing. The concept of priority pricing has been introduced by Harris and Raviv (1981) and extended by Chao and Wilson (1987) with a special emphasis on application to the electric power service. Wilson (1989a) generalized the idea of priority service to a general theory of efficient rationing and Wilson (1989b) combines the concepts of priority service with Ramsey pricing.

References

Adams, W. and Yellan, Janet (1976) "Commodity Bundling and the Burden of Monopoly", *Quarterly Journal of Economics* 90: 475–98.

Braden, David J. and Oren, Shmuel S. (1994) "Nonlinear Pricing to Produce Information", *Marketing Science* 13/3: 310–26.

Brown, Stephen J. and Sibley, David S. (1986) *The Theory of Public Utility Pricing*. Cambridge: Cambridge University Press.

Camerer C. F., Lowenstein, G., and Rabin, M. (2004) *Advances in Behavioral Economics*. Princeton, NJ: Princeton University Press.

Cassady, Ralph (1946a) "Some Economics of Price Discrimination under Non-perfect Market Conditions", *Journal of Marketing* 11: 7–20.

—— (1946b) "Techniques and Purposes of Price Discrimination", *Journal of Marketing* 11: 135–50.

Chao, Hung Po, Oren, Shmuel S., Smith, Stephen A., and Wilson, Robert B. (1986) "Multi-Level Demand Subscription Pricing for Electric Power", *Energy Economics* 8: 199–217.

—— and —— (1987) "Priority Service: Pricing, Investment and Market Organization", *American Economic Review* 77: 899–116.

Ferguson, D. G. (1994) "Shortages, Segmentation and Self-selection", *The Canadian Journal of Economics* 27/1: 183–97.

Goldman, M Barry, Leland, Hayne E., and Sibley, David S. (1984) "Optimal Nonuniform Pricing", *Review of Economic Studies* 51: 302–19.

Harris, Milton and Raviv, Arthur (1981) "A Theory of Monopoly Pricing Schemes with Demand Uncertainty", *American Economic Review* 71: 347–65.

Jucker, James V. and Rosenblatt, Meir (1985) "Single-Period Inventory Models with Demand Uncertainty and Quantity Discounts: Behavioral Implications and New Solution Procedures", *Naval Research Logistics Quarterly* 32: 537–50.

Katz, Michael L. (1982) "Nonuniform Pricing, Output and Welfare under Monopoly", *Review of Economic Studies* 50: 37–56.

Laffont, Jean-Jacque, Maskin, Eric and Rochet, Jean-Charles (1987) "Optimal Nonlinear Pricing with Two-Dimensional Characteristics", in T. Grove Radner and Reiter (eds) *Information, Incentives and Economic Mechanisms*. Minneapolis, MN: University of Minnesota Press, 256–66.

Lal, Rajiv and Staelin, Richard (1984) "An Approach for Developing an Optimal Quantity Discount Policy", *Management Science* 30: 1524–39.

Lutze, Holly and Ozer, Ozalp (2008) "Promised Lead Time Contracts under Asymmetric Information", *Operations Research* 56/4: 898–915.

Marchand, M. G. (1974) "Pricing Power Supplied on an Interruptible Basis", *European Economic Review* 5: 263–74.

Mirman, Leonard J. and Sibley, David S. (1980) "Optimal Nonlinear Prices for Multiproduct Monopolies", *The Bell Journal of Economics* 11: 659–70.

Mirrlees, James A. (1971) "An Exploration in the Theory of Optimal Taxation", *Review of Economic Studies* 38: 175–208.

Monahan, J. P. (1984) "A Quantity Discount Pricing Model to Increase Vendor Profits", *Management Science*: 30720–726.

Moorty, K. Shridar (1984) "Market Segmentation, Self-Selection, and Product Line Design", *Marketing Science* 3: 288–307.

Mussa, Michael and Rosen, Sherwin (1978) "Monopoly and Product Quality", *Journal of Economic Theory* 18: 301–17.
Oi, Walter J. (1971) "A Disneyland Dilemma: Two-Part Tariffs for a Mickey Mouse Monopoly", *Quarterly Journal of Economics* 85: 77–96.
Oren, Shmuel S., Smith, Stephen A., and Wilson, Robert B. (1982) "Linear Tariffs With Service Quality Discrimination", *The Bell Journal of Economics* 13/2: 455–71.
—— and —— (1982) "Nonlinear Pricing in Markets with Interdependent Demand", *Marketing Science* 1/3: 287–313.
—— and —— (1983) "Competitive Nonlinear Tariffs", *Journal of Economic Theory* 29/1: 49–71.
—— and —— (1984) "Pricing a Product Line", *Journal of Business* 57/1: S73–S99.
—— and —— (1985) "Capacity Pricing", *Econometrica* 53/3: 545–67.
—— and —— (1987) "Multiproduct Pricing for Electric Power", *Energy Economics* 9/2: 104–14.
Phillips, Robert (2005) *Pricing and Revenue Optimization*. Stanford, CA: Stanford University Press.
Phlips, Louis (1983) *The Economics of Price Discrimination*. Cambridge: Cambridge University Press.
Pigou, Arthur Cecil (1920) *The Economics of Welfare*. London: Macmillan Press, Ltd. 4th edn 1932.
Ramsey, Frank P. (1927) "A Contribution to the Theory of Taxation", *Economic Journal* 37: 47–61.
Roberts, Kevin W. S. (1979) Welfare Considerations of Nonlinear Pricing," *Economic Journal* 89: 66–83.
Smith, Stephen A. (1986) "New Product Pricing in Quality Sensitive Markets", *Marketing Science* 5/1: 70–87.
—— (1989) "Efficient Menu Structures for Pricing Interruptible Electric Power Service", *Journal of Regulatory Economics* 1: 203–23.
Stigler, George J. (1963) United States v. Loew's, Inc.: A note on Block Booking, Sop. Ct. Rev. 152.
Spulber, Daniel (1981) "Spatial Nonlinear Pricing", *American Economics Review* 71: 923–33.
Stiglitz, Joseph E. (1977) "Monopoly, Nonlinear Pricing and Imperfect Information: The Insurance Market", *Review of Economic Studies* 44: 407–30.
Su, Xuanming (2007) "Intertemporal Pricing with Strategic Customer Behavior", *Management Science* 53/5: 726–41.
Talluri, Kalyan T. and Van Ryzin, Garrett J. (2004) "Revenue Management under General Discrete Choice Models of Consumer Behavior", *Management Science* 50/1: 15–33.
—— and —— (2005) *The Theory and Practice of Revenue Management*. New York: Springer Press.
Tirole, Jean (1988) *The Theory of Industrial Organization*. Boston MA: MIT Press.
Tschirhart, J. and Jen, F. (1979) "Behaviour of a Monopoly Offering Interruptible Service", *Bell Journal of Economics* 10: 244–58.
Varian, Hal (1985) "Price Discrimination and Social Welfare", *American Economic Review* 75: 870–5.
Willing, Robert D. (1978) "Pareto-Superior Nonlinear Outlay Schedules", *The Bell Journal of Economics* 9: 56–69.
—— (1989a) "Efficient and Competitive Rationing", *Econometrica* 57: 1–40.
—— (1989b) "Ramsey Pricing of Priority Service", *Journal of Regulatory Economics* 1: 189–202.
—— (1993) *Nonlinear Pricing*. New York: Oxford University Press.