10 October 1999

SHORT COURSE ON NONLINEAR PRICING

Robert Wilson *

* Author’s address: Stanford Business School, Stanford, CA 94305-5015.
Phone 650-723-8620; Fax 725-7979. Internet: rwilson@stanford.edu.
These course notes are derived from a book published by Oxford University Press in 1993.
Chapter 1

INTRODUCTION

In the major regulated industries, each posted rate schedule or contract specifies a tariff that indicates the total charge payable by the customer for services provided. The prices (per unit) embodied in tariffs often depend on some aspect of the quantity of services or the rate at which they are purchased. The same feature is seen in many unregulated industries, usually in the form of quantity discounts, rebates, or credits toward subsequent purchases.

The way in which the quantity purchased affects the total charge takes many different forms. For example:

- Railroad tariffs specify charges based on the weight, volume, and distance of each shipment. For instance, discounts on the charge per mile per hundredweight are offered for full-car shipments and for long-distance shipments. In other transport industries such as trucking, airlines, and parcel delivery the rates depend also on the speed of delivery or the time of the day, week, or season.
- Electricity tariffs specify energy charges based on the total kilo-Watt hours used in the billing period, as well as demand charges based on the peak power load during the year. Lower rates apply to successive blocks of kilo-Watt hours and in some cases the demand charges are also divided into blocks. Energy rates for most industrial customers are further differentiated by the time of use, as between peak and offpeak periods.
- Telephone companies offer a variety of tariffs for measured toll service and WATS lines. Each tariff provides the least-cost service for a particular range of traffic volumes. Rates are also differentiated by distance and time of use.
- Airline fares allow “frequent flier” credits toward free tickets based on accumulated mileage. The retail value of a free ticket increases sharply with the number of miles used to acquire it. Further discounts are offered for advance purchase, noncancellation, round trip, weekend stays, and duration.
- Rental rates for durable equipment and space, such as vehicles and parking lots, are lower if the duration is longer. Rates for rental cars are also differentiated by the size of the vehicle and the time of use.
- Newspaper and magazine advertising rates are based on the size and placement of the insertion, the total number of lines of advertising space purchased by the customer during the year, and in some cases the annual dollar billings.

In each of these examples, the significant feature is that the average price paid per unit delivered depends on a measure of the total size of a customer’s purchase. This measure can apply to a single delivery, the total of deliveries within a billing period, accumulated deliveries over an indefinite period, or some combination. It can also depend on a variety of indirect measures of size such as the dollar value of purchases or the maximum single delivery.

Quantity discounts can be offered via smaller prices for marginal units, or via a smaller price for all units if the purchase size is sufficiently large. The size dependence can be explicit in a single tariff, or implicit in a menu of tariffs among which the customer can choose depending on the anticipated volume of purchases.

Several different terms are used to connote this practice. The generic term nonlinear pricing refers to any case in which the tariff is not strictly proportional to the quantity purchased. Tariffs in which the marginal prices of successive units decline in steps are called block-declining or tapered tariffs in several industries (in some instances the marginal prices increase over some range, as in the case of “lifeline” rates). The simplest example of a nonlinear tariff is a two-part tariff, in which the customer pays an initial fixed fee for the first unit (often justified as a subscription, access, or installation charge), plus a smaller constant price for each unit after the first. These and several other standard tariffs are sketched in Figure 1.1. A tariff that charges only a fixed fee is also called a flat-rate tariff. The tariffs in the figure pertain to a single product but in some cases a firm offers a multiproduct tariff that specifies a total charge based on the purchased quantities of two or more products. In the airline industry, for instance, frequent-flier plans offer extra mileage credit on specified routes or at specified times. In the power industry, the total charge for peak and offpeak power often depends on both the maximum power demanded and the average load factor as well as the total energy purchased.

To illustrate, Table 1.1 shows three tariffs for “wide area telecommunications service” (WATS) offered by AT&T, as reported by Mitchell and Vogelsang (1991). Each of the three
1. **INTRODUCTION**

**Figure 1:** Several kinds of tariffs.

Tariffs specify a monthly fixed fee plus rates for each call that depend on the distance; the table shows only the rate charged for each full minute for a distance of 1000 miles. There are two key features. First, each tariff has a block-declining structure indicated by the discounts allowed for the portions of monthly dollar billings above two threshold levels. Second, in choosing to which among the options in this menu of tariffs to subscribe, a customer perceives a more elaborate block-declining tariff obtained as the minimum among these tariffs for each combination of the number, aggregate minutes, and monthly dollar volume of calls.

The term nonlinear pricing is usually restricted to tariffs that are offered on the same terms to all customers in a large class. Thus, each customer pays the same marginal prices for successive units. Although the price for, say, the hundredth unit is the same for all customers, this price may differ from the prices for the tenth and the thousandth units.
1.1. Motivations for Nonlinear Pricing

A nonlinear tariff is much like a product line: offered a menu of quantities and corresponding charges, each customer chooses a preferred quantity and pays the associated charge. Each customer’s decision is essentially to select how many incremental units to buy or, when use is continual, the average rate of usage. Customers typically differ in their valuations of increments, and therefore different customers choose to buy different total quantities depending on the schedule of marginal prices charged for successive increments.

Similar interpretations apply to product lines of related products that differ according to one or more quality attributes. In this case an increment represents an improvement along a dimension of quality. Customers’ differing valuations of quality increments lead therefore to different selections of purchases from the product line, depending on the prices the tariff assigns to successive increments. A product line of machines, for instance, is typically differentiated by attributes such as size or production rate, operating cost or personnel requirements, durability or maintenance requirements, precision or fidelity, et cetera. Many examples from service industries pertain to conditions of delivery: familiar attributes include distance, time-of-day, speed, reliability or availability, and perquisites.
such as comfort or convenience. For instance, for successively higher prices a customer can have a letter or parcel delivered by several classes of regular mail and two classes of express mail, the speediest of which provides guaranteed overnight delivery with computerized tracking of each parcel individually.

In all these examples the net effect of nonlinear pricing is to offer a menu of options from which customers can choose differently depending on their preferences. Agricultural, commercial, and industrial customers account for the majority of sales in the main industries that use nonlinear pricing. Most of these customers are systematic and careful to obtain the maximum net value from service; for example, it is increasingly common to employ specialized managers aided by computerized control systems for services such as communications, power, and transportation. Moreover, these customers usually have stable preferences determined by production technologies, facility configurations, and sales and distribution systems. Among residential customers, an important segment is equally careful to obtain maximum advantage from the menu of options offered; for instance, Mitchell and Vogelsang (1991) report that AT&T’s “Reach Out America” options were selected by 68% of the Plan A customers in the first year and 85% in the third year.

The menu comprises the various quantities and/or qualities, to each of which the tariff assigns a charge. Whenever customers are diverse, a menu with several options is advantageous: it promotes greater allocative efficiency by enabling customers to adapt their purchases to their preferences. In the current jargon, an “unbundled” array of options allows each customer to “self-select” the best purchase. Thus the primary motive for nonlinear pricing is heterogeneity among the population of customers.

Differentiating increments by pricing them differently is advantageous whenever their price elasticities of demand differ and prices exceed marginal cost. This is explained by a standard proposition in economic theory: setting the percentage profit margin on each increment in inverse proportion to the price elasticity of its demand maximizes the aggregate dollar value of customers’ benefits for any fixed revenue obtained by the firm. Heterogeneity among customers virtually assures that different increments have different price elasticities, so the secondary motive for differentiated pricing stems ultimately from some exercise of monopoly power so that prices exceed marginal cost.

Among utilities, the primary motive for pricing above marginal cost is to obtain sufficient revenue to cover periodic operating costs and to repay the cost of capital used to install capacity such as durable equipment. This motive persists in competitive industries such as (deregulated) telecommunications, gas transmission, and rail, trucking, and air
transportation whenever there are economies of scale or the cost of durable capacity is recovered imperfectly by usage charges because demand is variable or stochastic. In oligopolistic industries where entry is limited by the magnitude of fixed costs in relation to profits, nonlinear pricing plays a similar role in maintaining profits sufficient to retain the maximum number of viable competing firms: this works to the ultimate advantage of customers even though it represents an exercise of monopoly power.

1.2. Practical Uses for Nonlinear Pricing

Nonlinear pricing has several roles in practice. In some cases nonlinear pricing promotes more efficient utilization of resources. In others it is used to meet utilities’ revenue requirements. And, it can also be used by firms with monopoly power to increase their profits. For example:

- Nonlinear pricing is often necessary for efficiency. This is the case when the firm’s cost per unit of filling or shipping an order varies with the size of the order. Similarly, if the firm has higher inventory costs than customers do then inducing customers to make periodic large purchases promotes efficiency. In the case of electric utilities, purchasers of large quantities of power typically have higher load factors, so quantity discounts recognize lower costs of idle capacity for these customers. Quantity discounts of this sort are motivated by cost considerations.

- Nonlinear pricing can be used by a regulated monopoly to recover administrative and capital costs. In important contexts, using its monopoly power to obtain operating profits sufficient to meet its revenue requirement is the most efficient means available to a utility or other public enterprise. Efficiency in these contexts refers to avoiding allocative distortions caused by deviations from prices set equal to marginal costs. In particular, if the regulatory objective is to maximize an aggregate of customers’ net benefits from the firm’s operations, then a nonlinear tariff minimizes allocative distortions. In this context, nonlinear pricing is a form of “Ramsey pricing” in which different units of the same commodity are interpreted as different products, as will be explained in §5. In particular, when marginal units of larger orders have greater price elasticities of demand, it is efficient to offer quantity discounts to minimize distortions from allocative efficiency. The distributional effects are not always favorable because customers making small purchases may pay higher prices; however, nonlinear pricing can be modified to avoid these adverse distributional effects. In sum, nonlinear pricing can be motivated by efficient use of monopoly power.
1. INTRODUCTION

to meet a revenue requirement, though perhaps modified to address distributional objectives.

- Nonlinear pricing is often a useful strategy in competitive markets. For example, newspapers and magazines offer quantity discounts to attract large advertisers who have substitutes available in other media such as television. Telephone companies offer quantity discounts for long-distance calls to retain large customers who otherwise might elect to bypass local exchanges. Frequent-flier plans offered by airlines were initially competitive tactics designed to appeal to business travelers. Quantity discounts of this sort are motivated primarily by competitive pressures in submarkets segmented by customers’ volume of purchases.

- Nonlinear pricing can also be used as a means of price discrimination that enables a firm with monopoly power to increase its profits. In this case, it is used mainly as a means of market segmentation in which customers are classified into volume bands. The extent to which it increases profits depends on the degree to which higher-volume customers have higher demand elasticities for incremental units. An example is a product line of machines, such as copiers or printers, that appeal to different volume bands because more expensive machines have higher rates of output and lower marginal costs. Nonlinear pricing of this sort is motivated primarily by market segmentation, but the prices are determined less by competitive pressures to survive than by profit opportunities that arise from differing elasticities of demand in various submarkets.

This list indicates that nonlinear pricing is potentially a vast subject. It is relevant to private firms as well as to public utilities, and its application can stem from considerations of cost, efficiency, or competitive pressures. In §15 we describe a host of other applications that are not recounted here, such as the design of contracts, tax schedules, and regulatory policies. In such contexts, a participant’s selection from a menu of options reflects incentives depending on personal preferences and other private information. In these cases, the role of differentiated pricing of products and services to take account of diversity among customers is supplanted by the design of incentives to take account of participants’ diversity due to their private information. In these as in all other cases, the primary advantage of differentiation stems from heterogeneity in the population.

The Role of Price Discrimination

The connection between nonlinear pricing and price discrimination, as in the last item
1.2. Practical Uses for Nonlinear Pricing

above, calls into question its purported advantages. In general, price discrimination can have adverse distributional effects and it can promote inefficient uses of monopoly power. Distributional effects are severe when different terms are offered to different customers based on observable distinctions. Allocative inefficiencies arise when prices do not minimize distortions required to meet the firm’s revenue requirement. And, productive inefficiencies arise when quantities and qualities are not produced at least cost, or for a given cost greater quantities or qualities could be produced.

Historically, the principle problem is quality degradation undertaken solely to enhance product differentiation. This problem was noted by Dupuit (1844) in the earliest treatise on pricing by public utilities and it persists currently. For instance, nonrefundable advance-purchase airline fares restrict a customer’s option to change an itinerary even when it would be costless for the firm — indeed, even as the plane departs with empty seats — and weekend-stay requirements that are costly for customers are enforced only because they are especially onerous for business travelers. Similarly, airlines offer different fares to different customers for identical services.¹

A standard example of quality degradation is temporal price discrimination in which a seller, such as a publisher, holds stock in inventory and offers a declining sequence of prices designed to obtain earlier sales at higher prices to more impatient customers. Even if storage costs are nil this is inefficient because it diminishes customers’ net benefits by the imputed costs of delay. Another standard example is a manufacturer who uses an inefficient product design based on limited durability or planned obsolescence. Various propositions in economic theory attempt to show that inefficient quality degradation and accompanying price discrimination will not occur or will not succeed, but the evident prevalence of such practices shows that the assumptions used are much too restrictive to be entirely realistic.²

Unfortunately, the adverse welfare effects of inefficient price discrimination are by-

¹ For a recent flight, the lowest fare available as a private individual was $328; if my host university made the reservation it was $288; and if my own university travel office made the reservation it was $200. For a second trip two months later in the midst of a price war, the fare was $66.

² These propositions are referred to collectively as Swan’s theorem and the Coase conjecture; cf. Swan (1972), Coase (1972), Bulow (1982), and Gul, Sonnenschein, and Wilson (1986). However, Bulow (1986) shows that Swan’s theorem is false when a formulation analogous to Gul, Sonnenschein, and Wilson is used; in particular, quality degradation in the form of inefficient product durability is an optimal strategy for a monopolist manufacturer. Also, Gul (1987) shows that the Coase conjecture is false for firms in an oligopoly.
passed entirely in this book. Careful analysis of the merits of allowing inefficient uses of monopoly power to meet a firm’s revenue requirement is a topic too complex and lengthy to be included here. Consequently, we consider only efficient uses of monopoly power in the form of Ramsey pricing. In particular, (a) we consider only cases in which the tariff offers the same terms to all customers; (b) the quality specifications of the firm’s products or services are supposed to be fixed; and (c) the firm’s costs and production technology are fixed — and presumably operated efficiently.

1.3. Feasibility of Nonlinear Pricing

The implementation of nonlinear pricing requires that four preconditions are satisfied. This section elaborates these basic requirements and indicates the variety of practical situations in which nonlinear pricing is feasible.

The four preconditions can be summarized briefly as follows: 3

- The seller has monopoly power.
- Resale markets are limited or absent.
- The seller can monitor customers’ purchases.
- The seller has disaggregated demand data.

Each is discussed in turn.

Monopoly Power

In severely competitive markets, prices are driven down close to the direct costs of supply. This excludes nonlinear pricing except for quantity discounts based on actual cost savings in production or delivery of large orders. Few markets are perfectly competitive, however. In monopolistically competitive markets, firms’ products are differentiated sufficiently that each enjoys some power to set its prices above direct costs. In oligopolistically competitive markets, firms’ products are close or perfect substitutes but the number of firms is sufficiently small to enable positive profits. These profits are limited to normal returns on irreversible investments in capacity if there is a persistent threat of entry.

---

3 To this list, for firms in unregulated industries in the United States, might be added legal feasibility. Although the Robinson-Patman Act has not been enforced in recent years, it prohibits quantity discounts that injure competition by giving large commercial customers a competitive advantage over smaller ones competing in the same retail market for commodity products (not services). For discussions of the Robinson-Patman Act in this connection see §5.2, Scherer (1980, §21) and Varian (1989, §3.7).
Nevertheless, nonlinear pricing is usually feasible in markets that are imperfectly competitive. We shall see in §12, however, that the degree of competition limits the extent of nonlinear pricing, and moreover, nonlinearities in the tariff are erased if there are many firms. Monopoly power may differ substantially among submarkets. Airlines, for instance, encounter competitive pressures that vary substantially among different routes.

A firm in a regulated industry is usually assured a monopoly regarding distribution (not necessarily production or generation) of services in its allotted district. The largest industrial customers are an exception in partially deregulated industries: in Georgia and England, for instance, electric utilities compete for customers with loads exceeding 900 kilo-Watts; and in the telephone industry in the United States there is strong competition for the largest commercial customers, partly from resellers who purchase bulk service from AT&T. The monopoly distribution franchise includes a service obligation along with control of investments, prices, and net revenues by the regulatory agency. The extent that monopoly power is exercised in rate design depends critically on the magnitude of the revenue requirement. A water company, for instance, may find no advantage in nonlinear pricing while an electric utility or a telephone company with substantially larger capital requirements may find it necessary.

**Resale Markets**

A necessary condition for nonlinear pricing is that resale markets are absent, limited, or controlled by the original supplier. If resale is freely available to customers, then nonlinear pricing enables some customers to profit from arbitrage. For example, if large orders are offered at a lower average price per unit then small ones, then a customer purchasing a large order can profit from breaking a large order into several smaller lots and selling them in the resale market. Allowance of resale by bulk purchasers has been a major feature of deregulation of long-distance telephone transmission. Alternatively, if the price schedule is increasing then a customer might benefit from opening multiple accounts and purchasing a small amount from each one. The net result for the original supplier is that all sales are at the lowest average price per unit it offers.

Excluding or limiting resale is therefore essential if nonlinear pricing is to be effective. Tariffs for electric power and communications, for instance, include explicit provisions prohibiting resale whenever regulatory policies allow. If resale is feasible but expensive, nonlinearities in the tariff are restricted to levels that cannot be arbitraged by customers. If bulk resale is excluded, then to exclude multiple accounts it suffices to ensure that the
price schedule is not increasing; i.e., the tariff is concave. The least restrictive requirement is that the tariff is subadditive: if the tariff charges \( P(q) \) for a purchase of size \( q \) then subadditivity requires that \( P(q_1 + q_2) \leq P(q_1) + P(q_2) \), which ensures that purchasing from two accounts is not cheaper than purchasing from a single account.

Resale exclusion is essentially an elaboration of the requirement of monopoly power, because it prevents competition from customers acting as secondary suppliers in resale markets. Competition from resale markets is pervasive for many durable goods even if the manufacturer of original equipment has a monopoly. Resale and rental markets for items of capital equipment offer customers a choice between purchasing new equipment, or buying or renting older equipment. Thus, each sale by the original manufacturer creates a potential competitor. Firms such as IBM and Xerox leased rather than sold their machines for this reason, until the practice was interpreted as anticompetitive under the U.S. antitrust laws. Other items are less susceptible to arbitrage: a large diamond or a large computer cannot be profitably divided into smaller ones.

On the other hand, resale and rental markets are precluded in many industries. Resale is impossible or expensive in many service industries, especially if it involves direct labor services. In capital-intensive industries, the technology may exclude resale if the seller controls essential equipment, such as switching and trunk lines in the case of communications and power, and aircraft in the case of airlines. In other industries it is excluded by regulatory provisions or business practices. For instance, resale of telephone service was prohibited until several years before deregulation of the industry, when the FCC required AT&T to allow MCI to resell long distance transmission.\(^4\) In the electric power industry, municipal utilities have long had the privilege of reselling power to local customers, but with this exception distribution has usually been a regulated monopoly, except as noted previously for the largest industrial customers in a few partially deregulated jurisdictions. The role of restrictive business practices is especially prominent in the airline industry, partly as a residual from many years of regulation; in particular, nontransferable tickets and reservations are issued to customers by name for specific flights. Similar practices are used in various markets for leased equipment, such as rental vehicles.

\(^4\) Reselling is sufficiently costly in telecommunications that even after the FCC directive allowing resale of switched services AT&T still has profitable WATS services that are priced nonlinearly and only the highest volume segment has encountered appreciable competition from resale. I am indebted to David Sibley for this observation.
Monitoring Purchases

An essential ingredient of any nonlinear tariff is a system to monitor customers’ purchases. Monitoring includes identifying customers, measuring their purchases, and billing. The system depends on several parameters of the tariff, such as:

- What is a customer? Customers are usually identified with transactions or billing accounts when uniform pricing is used, but when a nonlinear tariff is used the definition of a customer can have a major effect on tariff design. Extreme examples are the frequent-flier plans offered by airlines. These plans interpret the traveler as the customer, even if tickets are billed to the traveler’s employer, and rebated tickets are usually restricted to the traveler’s family. This definition of customers hinders employers from garnering the rebates allowed their employees. Somewhat similar is the practice among publishers of interpreting advertising agencies as customers, rather than the firms whose products are advertised. Multiple accounts may need to be excluded if the price schedule has an increasing segment, as in the case of lifeline rates for small purchases.

- What are the dimensions of the tariff? Purchases can be denominated in physical units, number of transactions, or dollar amounts. Several dimensions can apply simultaneously: for example, magazines offer discounts for large single advertisements and also for total annual billings; and electric utilities offer discounts on both demand charges (for the maximum power load) and energy charges. Similarly, the billing period is a crucial parameter of the tariff.

- What are the units of purchase? Quantity discounts can apply to single orders, the rate of purchases (over a billing period, or a longer span such as year), or cumulative purchases. Any of these can be measured in physical units, number of orders, or dollar amounts. Dollar measurements are especially useful for aggregating heterogeneous items. Often the choice is affected by the source of cost economies (such as a single shipment) or by the natural time frame over which purchases tend to be stable from period to period. Telephone and electricity tariffs are usually based on monthly billing periods or annual rates of consumption, whereas airline frequent-flier plans use annual rates and also cumulative purchases measured by mileage.

- What are the quality dimensions? Nonlinear pricing depends sensitively on whether a spectrum of qualities is offered as a differentiated product line (i.e., unbundled)
or instead prices are based on a single average quality. For instance, an electric utility can interrupt customers randomly or in rotation as needed in times of scarce supplies, so that all customers have the same chances of suffering an interruption, or it can offer differentiated services that provide a spectrum of service priorities. Differentiation usually entails more sophisticated monitoring: electric power differentiated by time of use requires metering and billing systems that separately record usage at different times, and differentiated services for interruptible or curtailable power require more elaborate dispatching and control systems.

- What is the method of billing? Many tariffs specify charges or rebates for single transactions or periodic billings, but others invoke elaborate procedures. The customer may be required to offer proof of purchases or to apply for discounts or rebates. Frequent-flier plans provide discounts only as rebates in kind via free tickets, quality upgrades, and accessory services such as privileged lounges. In some cases the discount is implicit, as in the case that the seller absorbs part or all of the cost of transport and delivery. A common procedure depends on a menu of contracts from which each customer chooses initially and then is billed accordingly at the end of the billing period. For example, a customer might choose between banking services at either a fixed price per check, or a monthly fee plus a lower price per check; for rental cars, between two daily rates with differing mileage charges; and for a parking space, between a monthly rate and a daily rate. Product lines of machines often take this form: the customer can choose among several increasingly expensive machines that have successively lower operating costs. If customers are uncertain what their consumptions will be, then the seller can reduce their risks by billing according to the most favorable contract the customer might have selected; e.g., several rental-car companies follow this practice and AT&T does so for some of its optional plans.

This partial list of the considerations involved in tariff design refers mainly to quantity discounts, but similar considerations apply to quality attributes. However, tariffs based on qualities impose special problems of measurement and contractual performance. For example, nonlinear pricing of service reliability, speed of delivery, or product durability cannot usually be implemented directly. Service reliability is more easily implemented by offering customers a choice of priority classes for access to scarce capacity; speed of delivery is often expressed vaguely (e.g., overnight delivery) or via queuing priorities;
and product durability is better assured via maintenance or replacement warranties.

A practical aspect of tariff design is the need to make the tariff simple enough that it can be understood by customers and sales representatives. In 1980, this author found the pricing manuals used by salesmen for a copier manufacturer and a telephone company (for WATS lines) so lengthy (each were 3 inches thick) and complex as to be indecipherable, as indeed was the opinion of a salesman in each case. Among the many tariffs offered for WATS lines, a third appeared to be more expensive for each volume of the customer’s usage than some combination of the others. This situation has changed dramatically with increasing deregulation of the long-distance telephone industry: AT&T now offers a simple array of WATS options that represents essentially a single four-part tariff.

**Disaggregated Demand Data**

The fundamental motivation for nonlinear pricing is heterogeneity among customers. It is because different customers value successive increments differently that a seller’s optimal product design and pricing policy differentiates according to purchase size. There are some markets (such as household durables like refrigerators and cars) in which purchase size is a moot consideration since customers rarely purchase more than a single item, but in many markets quantity is an important dimension and there is great dispersion among customers’ purchases.

Adapting the pricing policy to take account of the underlying heterogeneity among customers amounts essentially to designing a product line that is differentiated by quantity (or quality) of purchases. The essential demand data required for the design task are records or estimates of customers’ purchases classified according to both the price paid and the quantity purchased. Alternatively, the design can be based on models that rely on other measures of heterogeneity among customers, such as sales volume or production rates for industrial customers and socio-demographic indices for residential customers.

Many firms do not routinely accumulate data that is disaggregated to this degree, and therefore do not recognize or cannot measure the advantages of nonlinear pricing. In other cases, individual customer records are maintained but the data reflect so little variation in prices that demand elasticities for various purchase sizes are difficult to estimate. A wealth of data is obtained after an initial implementation of nonlinear pricing has been tried, but considerable risks may afflict the initial trial.
1. INTRODUCTION

It is important to recognize that nonlinear pricing need not involve fine differentiation to be advantageous. Typically it suffices to identify a few volume bands for differentiated prices in order to obtain most of the gains that a finely differentiated tariff could obtain. This is fortunate, because fine differentiation complicates the tariff and imposes costs of monitoring and billing. In practice it is often sufficient to offer a menu of several two-part tariffs.

In the next chapters we describe the basic features of nonlinear pricing and describe how a nonlinear price schedule is constructed. We emphasize the interpretation that it amounts essentially to market segmentation in terms of volume bands of customers. This interpretation provides an intuitive explanation of nonlinear pricing; in addition, it provides a systematic method of analyzing a wide variety of problems. In §2 several examples of nonlinear pricing illustrate how it is applied in practice. In the next section we introduce such applications with a few illustrations from the power industry, and then the final sections summarize the main conclusions derived in subsequent chapters.

1.4. Illustration: The Electric Power Industry

Nonlinear pricing has an important role in many industries. The exposition therefore presents the subject in sufficient generality to be widely applicable. Also, the examples are drawn from several industries in order to illustrate the variety of contexts in which nonlinear pricing is used. Nevertheless, the text includes many topics that are familiar and important aspects of rate design in the electric power industry.

It is instructive first to recall that purchases of generating equipment put a utility in the position of a customer facing a nonlinear pricing schedule. Consider a merit order of four generators: hydro, nuclear, coal, gas. These four have increasing costs of energy ($/kWh) and decreasing capital costs per unit of power ($/kW). For a one kilo-Watt load, therefore, as the duration of the load (hrs./year) increases, the merit order typically selects these four generation sources in reverse of the order listed. In effect, the utility faces a schedule of generation costs depending on duration that is nonlinear. This schedule is not offered by any one supplier, and moreover the utility may purchase several types of generators to meet its load-duration profile, but it illustrates that choosing from a nonlinear schedule is familiar in the power industry from the customer’s side of the transaction.

Nonlinear pricing of services from vendors is increasingly common in the power industry as an integral part of long-term power supply contracts allotted via auctions.
Whether designed by the utility or proposed in an auction by the supplier, payment schedules are typically differentiated by power level, duration, and total energy, and also by a variety of quality attributes such as availability, ease and assurance of dispatch, and other measures of reliability. Indeed, auctions especially attract contracts with elaborate pricing rules in order to meet the utility’s objectives at least cost to the supplier while ensuring sufficient incentives for compliance on both sides.

From the supplier’s side, utilities in the power industry have long offered a variety of nonlinear rate schedules, especially block-declining tariffs for commercial and industrial customers. A customer incurs energy charges each month, as well as a demand charge based on the customer’s peak power, that accumulate nonuniformly depending on the blocks in the rate schedule. Thus, the actual tariff combines a demand charge with an energy charge as in a two-part tariff but with significant nonuniformities due to the several blocks in the rates for usage and peak power. Wright tariffs enable an especially elaborate scheme of nonlinear pricing of both energy and maximum power level. Tariffs of both types are illustrated in §2.2 and analyzed in §11.

An alternative view of nonlinear pricing interprets the tariff as depending on an index of quality rather than quantity. This is familiar in the power industry as peakload pricing. That is, power in shoulder and peak periods has higher quality, in the sense that customers want more power at such times and some are willing to pay more. As with nonlinear pricing of quantities, the price per unit of energy varies systematically with the time of use, leading to higher prices in peak periods if capacity is scarce or generation costs are higher.

Other kinds of quality differentiation are also important. An important one in demand side management programs, as they are called in the power industry, is differentiation of rates based on the possibility of interruption or curtailment of service in times of scarce supply. This too has a cost justification because interruptible demands substitute for costly capacity to serve peak loads. Rate design based on differentiated service priorities is illustrated in §2.3 and analyzed in §10.

Of particular interest in the power industry is the burgeoning use of nonlinear pricing to retain large customers. This practice has increased in the telephone industry since deregulation in an attempt to deter large customers from electing to bypass local exchanges. But it is also an important tactic for utilities concerned about losing large customers to the form of bypass peculiar to the power industry, namely cogeneration. Quantity discounts for large purchases also have a cost justification because large cus-
tomers tend to have higher load factors and therefore impose relatively lower costs for capacity to serve peakloads. An example in §5.3 illustrates the use of a nonlinear tariff by a telephone company to offer favorable rates to large customers without imposing disadvantages on small customers. The key feature is that large customers with opportunities to bypass have, in effect, greater price elasticities of demand due to their opportunities to elect bypass — and it is precisely this greater elasticity that makes nonlinear pricing effective.

Lastly, we note that nonlinear pricing is essentially an application of Ramsey pricing in which market segments are identified with volume bands. In view of the pervasive role of the principles of Ramsey pricing for rate design in the power industry, this interpretation is useful for the light it sheds on the full implications of Ramsey pricing in rate design. Although regulatory agencies usually invoke arguments based on Ramsey pricing in the limited context of uniform prices, the method has a far greater range of application.

1.5. Overview of the Chapters

This book has four main parts. Part I is intended for general readers, whereas Parts II-IV are increasingly technical and they require tolerance for more mathematical symbolism. Part I includes Chapters 2 through 5. It presents the basic ideas of nonlinear pricing in elementary terms. Mathematical notation is kept to the minimum necessary for accuracy: it is used mostly for concise reference to various quantities, prices, costs, and revenues. The device that makes this elementary presentation feasible is the representation of demand data in terms of the demand profile defined in Chapter 3. The demand profile summarizes data in a partially aggregated form that retains information only about how the distribution of customers’ purchases is affected by the prices charged. This is the usual form in which demand data is available, and it is also the most aggregated form that retains sufficient information to construct a nonlinear tariff.

Chapter 4 demonstrates how the demand profile can be used to construct a nonlinear tariff for a single product offered by a profit-maximizing monopoly firm. Chapter 5 extends this analysis to a regulated firm that maximizes the aggregate of customers’ net benefits subject to the requirement that its net revenues are sufficient to recover its full costs. In addition, it shows how the tariff can be constrained to ensure that no customer is disadvantaged by nonlinear pricing, as compared to the uniform price that yields the same net revenue for the firm. General readers may also be interested in the other
applications of nonlinear pricing described in elementary terms in Chapter 15, and the short history of the subject in Chapter 16.

Part II comprises Chapters 6 through 8, which rely on completely disaggregated models in which each customer’s demand behavior is specified explicitly. It is intended for readers interested in the technical aspects of nonlinear pricing, and the density of mathematics is greater. Chapter 6 complements Part I’s exposition in terms of the demand profile with a parallel analysis based on a disaggregated demand model in which customers or market segments are identified by a single parameter affecting their preferences or demands. This chapter describes the design of multipart tariffs, fully nonlinear tariffs, and associated access charges or fixed fees. Chapter 7 uses several examples to illustrate the consequences of income effects. Chapter 8 considers several technical aspects and indicates how the previous analyses generalize to more complex models.

Part III considers the design of tariffs for single products with one or more auxiliary quality attributes, such as the time or reliability of delivery. Chapter 9 presents two versions of nonlinear pricing for products with multiple attributes, such as the quantity and also the quality, as measured by the time, speed, or other conditions of delivery. Chapters 10 and 11 focus on applications affected by supply or capacity limitations: in the first, aggregate capacity is rationed by pricing service priorities; and in the second, a customer’s capacity allotment is priced separately from actual usage.

Part IV addresses the design of tariffs for multiple products priced jointly. Chapter 12 considers first the case of separate tariffs for several products. It also shows how the analysis extends to competition among several single-product firms in a single industry. Chapters 13 and 14 develop the general theory of nonlinear tariffs for multiple products, first in terms of disaggregated demand models and then in terms of a multiproduct version of the demand profile adapted to computations.

Two supplementary chapters provide additional material of general interest. Chapter 15 outlines applications of nonlinear pricing in other contexts, such as contracting in insurance and labor markets, which are similarly affected by the feature that each participant’s selection from a menu of options is affected by personal attributes or private information. Chapter 16 provides a synopsis of the literature on nonlinear pricing.

The following list of the chapters includes brief summaries of their contents. Each is accompanied by a “bottom line” statement of the main qualitative conclusion derived.

**Part I: Fundamentals of Nonlinear Pricing**
2. Illustrations of Nonlinear Pricing: This chapter describes the practice of nonlinear pricing by firms in several industries. Several competitive industries are included, such as publishing, airlines, and express mail, as well as regulated monopolies, such as electric power and telephone.

Conclusion: In both regulated and competitive industries, various forms of nonlinear pricing are used to differentiate or “unbundle” quantity and quality increments.

3. Models and Data Sources: This chapter describes the data used to construct nonlinear tariffs. It emphasizes the representation of demand data in summary form in terms of an estimate of the demand profile.

Conclusion: Because the advantages of nonlinear pricing stem from heterogeneity among customers, implementation relies on estimates of the price elasticity of the distribution of customers’ purchases, which can be obtained from data summarized in the demand profile.

4. Tariff Design: This chapter shows how the demand profile is used to construct a nonlinear tariff for a profit-maximizing monopoly firm. Construction of the schedule of marginal prices is illustrated both for demand data in tabular form, and for parameterized models estimated from demand data. Nonlinear pricing can also be interpreted as a special case of bundling in which different prices are charged for various combinations or bundles of items.

Conclusion: For a profit-maximizing firm, the schedule of marginal prices is derived by optimizing the price charged for each increment in the purchase size. With modifications this principle applies also to access fees and multipart tariffs.

5. Ramsey Pricing: This chapter shows that with only slight modification, the same principles of tariff design apply to a regulated firm whose price schedule maximizes the aggregate of customers’ net benefits subject to recovery of the firm’s full costs. In practice, however, it is often important to modify the tariff by imposing the constraint that no customer is disadvantaged compared to an existing tariff, such as the uniform price that recovers the same revenue for the firm. Typically this is done to ensure that efficiency gains from a nonlinear tariff obtained by customers purchasing large amounts do not impose unfavorable consequences on other customers purchasing small amounts.

Conclusion: Ramsey pricing is based on essentially the same principles of tariff design, but to ensure that no customer is disadvantaged (compared to the uniform
price that meets the same revenue requirement), customers can still purchase initial increments at the revenue-equivalent uniform price.

**Part II: Disaggregated Demand Models**

6. *Single-Parameter Disaggregated Models:* This chapter reconsiders the topics in Chapter 4 in terms of parameterized models of customers’ benefits or demands. Such models are used in most technical treatments of the subject. A single type parameter is used to describe the differences among customers or among market segments. Optimal tariffs are derived for versions in which the type parameter is discrete or continuous. The construction of optimal fixed fees and multipart tariffs are also derived.

*Conclusion:* A model that identifies distinct market segments via a single demand parameter enables exact characterization and computation of optimal multipart and nonlinear tariffs.

7. *Income Effects:* This chapter describes the consequences for nonlinear pricing of income effects derived from customers’ budget constraints. Three different types of income effects are illustrated with numerical examples. The computations required to calculate an optimal tariff are more complicated, but the main qualitative features are not altered.

*Conclusion:* Examples indicate that income effects can be severe, depending on their form and the variance of the income elasticity of demand in the population, but if the tariff leaves customers with large residual incomes and the income-elasticity variance in the population is small then income effects have little effect on tariff design.

8. *Technical Amendments:* This chapter examines several technical aspects of nonlinear pricing. The first two sections develop conditions sufficient to ensure that a nonlinear tariff meets all the requirements for optimal behavior by customers and by the firm. The third demonstrates that an optimal multipart tariff, or a menu of optional two-part tariffs, closely approximates the optimal nonlinear tariff. In particular, the profit or surplus lost by using only \( n \) segments or options is approximately proportional to \( 1/n^2 \). A menu with only a few options is therefore sufficient to realize most of the potential gain from a nonlinear tariff. The fourth shows that the main results from previous chapters are valid also for disaggregated demand models in which customers’ are described by multiple type parameters. The fifth describes an extension of the demand-profile formulation to encompass dependencies on the total tariff charged. The main new feature is that prices for small purchases may be
reduced in a fashion similar to lifeline rates.

**Conclusion:** (a) To ensure that demand projections are accurate, the assignment of customers’ types to purchases must be monotone, which can be assured by using an averaging procedure to eliminate nonmonotonicities. (b) A multipart tariff with only a few segments obtains most of the gains from nonlinear pricing. (c) The main features of analyses based on one-dimensional type parameters are preserved in models with many parameters. (d) If demands are sensitive to the total tariff then charges for small purchases may be reduced to retain the optimal market penetration.

### Part III: Multidimensional Tariffs

**9. Multidimensional Pricing:** This chapter presents the simplest form of multiproduct nonlinear pricing in which each customer assigns, to each unit purchased, a preferred combination of quality attributes. This form of nonlinear pricing is used when customers select delivery conditions or other aspects of service quality. Chapter 4’s method of constructing a nonlinear tariff from the demand profile applies equally to this case, but some applications involve substitution effects requiring more elaborate calculations.

**Conclusion:** Nonlinear pricing can be extended straightforwardly to multiple quantity and quality dimensions if customers select arbitrary sets of increments; however, the tariff is affected by substitution effects in customers’ demands.

**10. Priority Service:** This chapter describes an application of multidimensional pricing to the case that customers are rationed based on supply availability. Customers select their service priorities from a menu of options priced according to a nonlinear tariff.

**Conclusion:** When supply is limited by capacity constraints and demand or supply is stochastic, nonlinear prices for service priorities and quantities jointly can be designed using the principles of Ramsey pricing in a multidimensional formulation.

**11. Capacity Pricing:** This chapter extends multidimensional pricing to include charges for capacity as well as usage. The charges can take several alternative forms corresponding to either time-of-use usage charges accompanied by uniform demand charges for capacity, or nonlinear pricing of capacity increments accompanied by nonlinear pricing of the duration of usage of each capacity unit (called a Wright tariff in the power industry).

**Conclusion:** When customers purchase sets of load-duration increments reflecting peakload effects, nonlinear pricing can be used to design unified tariffs for usage
and capacity requirements.

**Part IV: Multiproduct Tariffs**

12. *Multiple Products and Competitive Tariffs:* This chapter illustrates several extensions of nonlinear pricing to multiproduct contexts. In the first case considered, a single firm offers a separate tariff for each of several products. The optimal tariffs can be derived from a multiproduct version of the demand profile in which substitution effects are represented explicitly. The calculations can be done via a simple gradient procedure. In the second case, each firm in a competitive industry offers a separate tariff for its own product. One model considers pure price competition among firms whose products are imperfect substitutes; a second considers firms offering identical products but competition is affected by the firms’ supply or capacity commitments. A third case mentioned briefly considers competition among multiproduct firms who adapt their pricing policies to changing demand conditions.

*Conclusion:* Nonlinear pricing can be used for multiple products offered by a single firm, or by several firms in a competitive industry. Customers’ opportunities to substitute one product for another have pronounced effects on tariff design. The form of implementation depends on whether multiple products are priced separately or jointly and on the competitive structure of the industry.

13. *Multiproduct Pricing:* This chapter describes the construction of a single comprehensive tariff by a multiproduct monopolist or regulated firm. The formulation relies on a model of customers’ benefits or demand functions with one or more parameters that describe customers’ types. When there is more than one type parameter, computations are complicated by the technical requirement that the marginal price schedules assess cumulative charges independent of a customer’s pattern of accumulation. Examples indicate that a significant feature of multiproduct tariffs is an emphasis on bundling; that is, the price schedule for each product is strongly dependent on the quantities the customer purchases of other products. For instance, prices for power in peak periods can depend on the customer’s offpeak load, as in the case of so-called load-factor discounts.

*Conclusion:* The design of a multivariate tariff for multiple products follows the same general principles as for a single product but calculations are complicated. Bundling is important: each product’s prices depend on purchases of other products.

14. *Multiproduct Tariffs:* This chapter constructs an approximation of the optimal multivariate tariff that is derived directly from the multiproduct demand profile. The
approximation is based on omitting consideration of customers’ participation constraints, and therefore incomplete account is taken of opportunities for multiproduct bundling. The tariff can be constructed from the “price differentials” charged for multiproduct increments. A simple gradient algorithm suffices for the calculations.

Conclusion: The multiproduct demand profile can be used directly to compute an approximately optimal multiproduct tariff, and for many problems this is the easiest approach.

Supplementary Chapters

15. Other Applications of Nonlinear Pricing: Previous chapters emphasize applications of nonlinear tariffs to pricing in product markets. This chapter describes a few of many applications to other markets, such as insurance, contracting, and work incentives. In product pricing, nonlinear tariffs are useful because customers are diverse, whereas in other applications the role of diversity is supplanted by the privacy of participants’ personal information or actions. That is, the diversity of customers’ types is represented by the diversity of what a single participant might know.

Conclusion: In other contexts of product pricing, contracting, regulation, and taxation, the design of the tariff or other menu of options recognizes that each participant’s selection depends on personal preferences or private information.

16. Bibliography: This chapter provides a brief history of the development of the theory of nonlinear pricing, and it provides references to important contributions. The large technical literature stems from Ramsey’s formulation of product pricing to meet a revenue requirement and Mirrlees’ applications to governmental taxation. The subsequent literature addresses a wide variety of applications of nonlinear pricing to quantity and/or quality differentiated products, and related applications to contracting affected by incentives and private information.

References: All the references cited in the text or mentioned in the bibliography are collected together at the end, along with a selection of uncited articles of related interest.

The illustrations and applications described in the text represent a substantial selection bias in favor of the major regulated industries that use nonlinear pricing most heavily. In particular, my own background is mostly in the electric power industry and so it is used often as a case study.
Most of the chapters include numerical examples that indicate either the method of calculation or the character of the results. These examples are chosen purposely to be quite simple, since they are used only for illustrative purposes. In practice, of course, rate design deals with more complex pricing problems. Actual applications involve substantial tasks of data analysis, model formulation and estimation, and complicated calculations that are treated incompletely here.

1.6. Summary

The main themes of the book can be summarized as follows.

- The illustrations and applications in the text indicate that nonlinear pricing is widely used, often in disguised forms. The advantages of nonlinear pricing derive ultimately from heterogeneity among customers’ valuations of successive increments of the quantity or quality purchased. This heterogeneity in the population of customers allows product differentiation to be based on segmenting the market according to volume bands or quality-differentiated product lines. Some degree of monopoly power, reinforced by limitations on resale markets, is essential for nonlinear pricing to be effective. Implementation requires that the seller can identify customers and monitor their purchases.

- A nonlinear tariff can be constructed from ordinary demand data obtained from uniform pricing. The observed distribution of purchases at various prices, summarized in the demand profile, provides the requisite demand data. For a single product, elementary arithmetical calculations suffice to design the price schedule for a nonlinear tariff. This is done by interpreting the tariff as charging prices for a product line consisting of successive increments of a customer’s purchase. Nonlinear tariffs offered by competing firms can be analyzed by the same methods. The calculations are comparably simple for multidimensional pricing, but for general multiproduct pricing the calculations are complicated by the effects of substitution among products.

- Nonlinear pricing can be based on explicit consideration of each customer’s predicted behavior and often this is done by using parameterized models that identify distinct market segments. At this level of detail, an optimal tariff can be characterized exactly, at least in principle. Because sufficient data to estimate such models is rarely available, however, portions of the exposition rely on a partially aggregated
formulation in terms of the demand profile, which summarizes how the distribution of customers’ purchases depends on the prices charged.

- Practical applications of nonlinear pricing use a multipart tariff or a menu of optional two-part tariffs. This is sufficient because a multipart tariff with a few segments realizes most of the advantages of nonlinear pricing. Multipart tariffs and optional two-part tariffs are constructed in the same way except that the optimality condition represents an average over each segment of the conditions that pertain to a fully nonlinear tariff.

- Nonlinear pricing is an efficient way for a regulated monopoly to obtain sufficient revenue to recover its full costs. However, because a nonlinear tariff benefits mostly large customers via quantity discounts, small customers can be disadvantaged compared to the uniform price that obtains the same revenue. A useful expedient to eliminate these unfavorable distributional effects is to cap the marginal prices at the revenue-equivalent uniform price.

Part I

FUNDAMENTALS OF NONLINEAR PRICING
Chapter 2

ILLUSTRATIONS OF NONLINEAR PRICING

Quantity discounts have a long history in business practice. The number of deliberate applications of the principles of nonlinear pricing has, however, increased markedly in recent years. Three causes account for this growth. The elementary explanation is that an increasing number of firms offer standardized products or services in mass markets. Because these markets comprise diverse market segments, but a single tariff, price schedule, or menu of options is offered, there are advantages to tariffs that differentiate products or services according to observable aspects of customers’ purchases, such as purchase size or quality selection. Moreover, the rudimentary tasks of monitoring purchases and of accounting and billing have been simplified by computerized data processing. A further explanation is that major industries such as transportation, communications, and power (electric generation and bulk gas supply) have been substantially deregulated. The competitive pressures and high capital costs in these markets require product differentiation for survival, and nonlinear pricing is an important means of differentiation. A deeper explanation, however, is the recognition that product differentiation generally and nonlinear pricing in particular are necessary ingredients of efficient product design and pricing in many industries. This explanation is developed in more detail in §5 when we study Ramsey pricing by a regulated firm with a revenue recovery requirement, and in §12 when we study imperfectly competitive markets. Here our intent is to introduce some of the practical aspects of nonlinear pricing as it is practiced in several industries.

This chapter describes several applications of nonlinear pricing. Section 1 describes the structure of rates for advertising used by two major weekly magazines in a highly competitive market. Sections 2 and 3 describe the tariffs offered in France and California by public utilities supplying electric power. These applications illustrate several practical aspects of nonlinear pricing. In both cases, the firms use elaborate tariffs that invoke nonlinear pricing on several dimensions simultaneously, some of which are dimensions of quantity and others, of quality. Second, these tariffs are motivated partly
by the firm’s cost considerations and partly by customers’ demand behaviors — in each case, in ways peculiar to the technologies of the supplier and the customers in the industry. And third, one application is drawn from a competitive industry while the others pertain to regulated monopolies, thus illustrating a wide spectrum of possibilities for applying nonlinear pricing. Sections 4, 5, and 6 provide synopses of nonlinear pricing in the telephone, express mail, and airline markets, all of which have been substantially deregulated in recent years.

2.1. Time and Newsweek’s Advertising Rates

Extreme examples of nonlinear pricing are found in the rate schedules offered by newspapers and magazines for advertising insertions. In this section we describe the schedules used by *Time* and *Newsweek* magazines, two of the major news weeklies competing in the national market in the United States. ¹ Because these schedules are long and complicated, we summarize the main features.

Both magazines use nonlinear pricing on several dimensions, including the size of an individual advertisement, the portion of total circulation purchased per issue, and the customer’s cumulative dollar amount of advertising purchased per year. In addition, each magazine differentiates its schedule further according to quality attributes, including the color content and the audience characteristics. Each of these is described in turn.

Each magazine uses a three-column format; consequently, prices for single advertising insertions (“ads”) are quoted in terms of multiples of half-columns or sixths of a page, and differentiated according to whether the number of colors (in addition to black) is 0, 1, or 4 in order of increasing quality and price. We refer to a full-page four-color ad as 1P4C, et cetera. The complexity of the resulting schedules can be reduced by dividing each price by both the size of the insertion and the price of a 1P4C ad, thus converting prices to a full-page four-color basis, which we call *normalized* prices. Dividing further by the magazine’s *reach* (the size of the magazine’s audience) puts the prices in terms of CPM: the cost per thousand readers reached. Although *Time*’s normalized rates are over 40% more expensive than *Newsweek*’s, its CPM-rates are only about 10% higher for 1P4C ads. *Time*’s normalized prices for each color quality decrease as the size of the ad increases, and so do the marginal prices of incremental sizes with two exceptions:

¹ This section is based on Timothy McGuire, “Nonlinear Pricing and Unbundling by the Major Newsweeklies”, Stanford Business School, December 1987. The description is based on the January 5, 1987, rate cards #71 and #58 of *Time* and *Newsweek*, respectively.
for 0C and 1C ads the incremental normalized price of moving from a third-page to a half-page ad increases. (This is presumably due to the higher cost of composing a page that includes an ad taking up the top or bottom half of the page; for Newsweek, even the normalized prices increase at these same points.) For each size the incremental cost of 1C is positive, and the further incremental cost of 4C is also positive, but these increments decline with the size of the ad — and more so for 4C than for 1C. Time’s normalized prices for third-page and full-page ads reflect quantity discounts for full pages of 17% and 23% for 0C and 4C respectively, and Newsweek’s discounts are slightly less. For each size, each magazine adds a fixed percentage to the 0C price to obtain the 1C price (Time’s percentage is smaller), and a percentage that declines with size to obtain the 4C price.

The second category of nonlinear pricing is based on the circulation purchased. Each magazine offers advertisers various options: the national edition, various overlapping regional and “major markets” editions, and several special editions aimed at selected demographic categories such as income level or occupation. Time and Newsweek have 113 and 94 regional editions, as well as many special editions. (The magazine a subscriber receives includes the ads from all editions for which he qualifies based on geographical location and demographic characteristics.) Typically the prices for these editions have slightly higher targeted-CPMs compared to the national edition, reflecting the higher “quality” audience obtained. For 4C, Time charges $6390 for the first geographical edition, $3495 for the second, and zero thereafter up to 100,000 circulation, with additional circulation priced at a declining marginal rate with 15 blocks that varies from $35.46 down to $27.40 per thousand copies for more than 2.9 million copies. These rates are uniformly higher than the “professional-managerial audience” targeted-CPM of about $18 (and the “adult audience” CPM of about $4.50) for the national edition, but the special editions obtain nearly the same CPMs for such selected audiences. Newsweek charges a flat rate of $820 per regional edition up to the tenth and then zero, but its schedule of marginal charges per thousand copies over 196 ranges of circulation shows almost no pattern; indeed, the marginal charge is lowest at 0.4 million copies ($24) and significantly higher ($27) at 2.0 million. The irregularity of Newsweek’s marginal circulation charges might be due to the tabular form of the rate schedule, comprising three tables each with 196 rows and 10 columns, which tends to obscure the marginal charges that concern customers.

The third category of nonlinear pricing is based on the customer’s total annual purchases of advertising, and it occurs in three parts. Time, for example, offers an 11-
block schedule of discounts based on total annual purchases that is stated in terms of the percentage rebate that applies to the entire year’s purchases: these rebates range from 0% up to 15% for purchases exceeding $8.9 million per year. Second, it offers a 4-block schedule of rebates for matching the purchases in the prior year: these range from 0% up to 4% for purchases exceeding $8.9 million. And third, further rebates of 3% or 5% are given for exceeding the prior year’s spending by $0.375 or $0.750 million. Whereas the latter two types of rebates are evidently aimed to preserve market share and capture customers’ annual increments in their advertising budgets, the first type is aimed at a special feature of the advertising industry. If identical or similar ads are placed in successive issues of a magazine, the reach (the number of readers who see the ad in any issue) increases but at a sharply decreasing rate; similarly, the average frequency (the number of times a specific reader sees the ad) increases at a decreasing rate. Further, frequency has a diminishing effect on the chances a reader will purchase the advertiser’s product. For example, Time’s reach from 52 insertions over a year is only 3 times the single-issue reach, and the frequency is only 18 times as high. Thus, advertisers encounter sharply decreasing returns from multiple insertions in the same magazine. Presumably Time’s rebate schedules reflect this feature of customer’s demands for multiple insertions. An additional possibility pertains to customers who sell multiple products: the rebates give some incentive to purchase advertising about secondary products for which the gains from advertising are less.

In sum, the rate schedules offered by these two magazines address several different dimensions of customers’ decision processes. For a single insertion, the rates offer decreasing marginal prices for increments in the size of ads, as well as for quality attributes such as the number of colors. Further quantity discounts are offered for the circulation of the issue, and for the regional and demographic quality of the audience reached. Finally, rebates on annual expenditures are offered to counter the effects of diminishing marginal reach and frequency. Combining all these features into the rate schedule produces a long and elaborate tariff that is seemingly bewildering in its complexity, at least for the uninitiated. Nevertheless, in this intensely competitive industry these pricing practices have survived, and in fact are used similarly by both major news magazines.

2.2. EDF’s Tariffs for Electric Power

The public utility that produces and distributes electric power in France is Electricité de France, generally known as EDF. For many years its tariffs have been based on an
elaborate system of nonlinear pricing. The tariffs are color-coded according to a scheme familiar to customers. We discuss the following three:

- **Tarif Bleu**: The blue tariff offers one schedule for residences and farms, and another for professional offices with power loads up to a maximum of 36 kilo-Volt-Amperes (kVA).
- **Tarif Jaune**: The yellow tariff applies to customers with loads between 36 and 250 kVA.
- **Tarif Vert**: The green tariff offers a variety of rate schedules for industrial and commercial customers with loads less than 10,000 kilo-Watts (kW) in series A, a second set of schedules for customers with loads between 10,000 and 40,000 kW in series B, et cetera.

These tariffs use nonlinear pricing in different ways that we describe briefly.\(^2\)

**The Blue Tariff**

The blue tariff for professional offices offers three options among which each customer can select. Each option consists of a fixed charge per month plus a charge for the actual energy consumed, as measured in kilo-Watt-hours (kWh). The fixed charge, based on the power rating of the customer’s appliances, is computed according to a nonlinear schedule in which customers with higher power ratings pay disproportionately more. Moreover, the schedule of monthly fixed charges depends on whether the customer selects the basic, “empty hours”, or “critical times” option. Compared to the basic option, the monthly charge is about 25% higher for the empty-hours option, and about 50% lower for the critical-times option. The basic option has a constant energy charge (50 centimes/kWh) regardless of the circumstances in which the energy is consumed. In contrast, the empty-hours option allows a discount of nearly 50% for energy consumed at night in the designated off-peak hours (10 PM to 6 AM) and the critical-times option has a 36% lower energy charge (32 centimes/kWh) applicable to all hours, but imposes a surcharge of over 800% for energy consumed after the utility has broadcast an announcement that power is in scarce supply. The blue tariff for residences and farms offers the same options and imposes the same energy charges, but the fixed charge is less than half as much; moreover, the fixed charge is payable annually.

\(^2\) The following description is based on the tariffs effective in February 1987.
The blue tariff can be interpreted as a collection of three two-part tariffs, each consisting of a monthly fixed charge plus an energy charge. The basic fixed charge increases with the maximum power level at an increasing rate, although it is appreciably less for residences and farms. The likely cost basis for this schedule is that higher maximum power levels in offices are associated with lower load factors and sharper peaks that require the utility to maintain greater reserve margins of capacity to serve peak loads. Residential and agricultural customers have higher load factors, considering that the peak is in the winter and is due substantially to heating loads: another possibility is that the lower fixed charges for residences and farms reflect subsidies.

The three options offer each customer a choice of which two-part tariff to elect. The empty-hours option incurs a higher fixed charge but allows the customer substantial savings on the charges for energy consumed in night-time hours. This option reflects the cost structure of baseload generation. The critical-times option saves appreciably on both the fixed and energy charges but the surcharge that is applied when supplies are scarce essentially requires the customer to curtail demand at these critical times. Thus, customers who elect the critical-times option are according lowest priority for service. Because these times are substantially unpredictable and the warning time is short, the customer must be able to interrupt operations quickly and inexpensively to take advantage of this option. The lower charges reflect the fact that such customers require essentially no reserve margin and their energy can be supplied by baseload capacity or by idle peaking capacity; indeed, as more customers elect this option the utility can alter its mix of generation technologies in favor of more baseload sources such as nuclear reactors.

The Yellow Tariff

The yellow tariff applicable to small industrial facilities is similar to the blue tariff but differs in several respects. First, the energy charges in all categories differ between the summer and winter seasons: energy charges are three to four times higher in winter than in summer. This difference reflects higher peak loads in winter due to the prevalent use of electric power for space heating. These peak loads are met with generators that use fossil fuel, which is substantially more expensive than the hydroelectric and nuclear sources used to meet base loads. A second feature is that the fixed charge, payable annually for each kVA of power demanded, is based on a fixed rate (rather than a nonlinear schedule).

The third and key feature, however, is that the fixed and energy charges are set
2.2. EDF’s Tariffs for Electric Power

nonlinearly based on the duration of each increment of the load during the year; tariffs of this form are called Wright tariffs in the United States. To understand this scheme, imagine that the customer’s meter comprises many small meters: the first kVA of the customer’s load is assigned to the first meter; the second kVA, to the second meter; et cetera. Whenever the customer’s power load is, say, six kVAs, the first six meters are turning. At the end of the year, the first meter records the number of hours in which a first kVA was being used, the second meter records the number of hours in which at least two kVAs were being used, et cetera. The yellow tariff assigns each kVA to one of two categories, designated medium utilization and long utilization, corresponding to whether that kVA was used for less or more than a certain number of hours per year. This assignment can also be interpreted as a classification of the customer’s load into peakload and baseload segments. For a kVA designated as part of the peakload segment in winter during daytime hours, the annual fixed charge is less than a third as high but the energy charge can be nearly 50% greater to reflect the lower capacity costs and higher fuel costs of the generators used to serve peak loads. The critical-times option is available only for long utilizations and only during the winter; during the summer it is identical to the empty-hours option.

The structure of the yellow tariff reflects the fundamental features of the cost structure of electricity generation. The utility can choose among several kinds of generators, such as hydroelectric, nuclear, coal, and oil or gas. Within this list, the energy cost is increasing and the capital cost is decreasing. For example, a nuclear generator has a lower energy cost and a higher capital cost than a gas turbine. For base loads, therefore, it is less expensive to use a nuclear generator, and to use a gas turbine only to meet the incremental loads that occur in periods of peak demand. To see this, it is easiest to recognize that whichever generator is used to meet infrequent peak loads will be idle most of the time, and since the gas turbine has the lowest cost of construction, its imputed cost of being idle is also the lowest. Thus, even though gas turbines are expensive to operate, they are cost-effective in meeting infrequent short bursts of demand. The yellow tariff recognizes this cost structure by imposing on peakload segments of the customer’s demand, a lower fixed charge (used to recover the capital costs of constructing generators), and a higher energy charge. The net effect is a tariff that uses nonlinear pricing of load duration in the form of a piecewise-linear tariff with two segments:

- In the first segment, for durations (hours per year) short enough to be classified as
part of the customer’s peakload segment, the fixed charge (francs per kVA per year) is low and the energy charge (centimes per kWh) is high.

- In the second segment, for durations long enough to be classified as part of the customer’s baseload segment, the fixed charge is higher and the energy charge is lower.

Thus, the key feature is that for each increment of the customer’s load, the yellow tariff depends nonlinearly on the accumulated hours that the increment is demanded during the year.

The Green Tariff

The green tariff applicable to medium-sized industrial and commercial customers offers a wide variety of rate schedules among which each customer can select. Here we describe only a few of these schedules for series A and B.

Schedule A5 for customers with loads less than 10,000 kW offers a basic option and a critical-times option similar to the corresponding options in the yellow tariff, but with three refinements. First, an additional charge is imposed for reactive energy (centimes per kVArh). Second, the energy charge in winter is differentiated according to three periods: in addition to the nighttime and daytime periods, a third designation of peak periods within the daytime period is added. Compared to the regular daytime energy charge, the peak-period energy charge is 28% to 50% higher depending on the duration of utilization (higher for shorter utilizations, since on average there is a greater dependence on peaking generators). Third, utilization is divided into four intervals: in addition to the medium and long utilizations, short and very long utilizations are specified. As in the yellow tariff, longer utilizations incur higher annual fixed charges and lower energy charges. Figure 2.1 shows how the annual charges under the green tariff Schedule A5 basic option for winter peak hours depends on the annual hours per year. The four intervals of annual utilizations are numbered 1 to 4 in the figure. The summary effects of these features is to make Schedule A5 of the green tariff a refined version of the yellow tariff in which more detailed account is taken of the effects of the customer’s usage on the utility’s costs of capacity and energy.

Schedule A8 is similar except for the addition of further distinctions that allow even lower off-peak rates in July and August (August is when nearly all of France is on

---

3 Only two duration intervals are used for the critical-times option.
2.2. EDF’s Tariffs for Electric Power

Figure 1: EDF Green Tariff A5 for the basic option in winter peak hours depending on the annual utilization.

vacation), and intermediate rates in Autumn and Spring; thus, it includes still further refinements. These additional refinements are also included in Schedule B for larger customers with loads between 10,000 and 40,000 kW. However, compared to Schedule A5, in Schedule B the fixed charges are lower by about 20% and the peak energy charges are lower by 7% (very long utilization) to 27% (short utilization). Taken together, therefore, we see that Schedules A5 and B constitute nonlinear pricing along the dimension of the customer’s load. That is, for any prescribed length of utilization and time of use, the fixed charge and the energy charge are both lower if the customer’s load exceeds 10,000 kW.

Summary

EDF’s tariffs illustrate several important features of nonlinear pricing as it is applied in
ILLUSTRATIONS OF NONLINEAR PRICING

practice.

1. Different tariff schedules are offered to different classes and power loads of customers. Each schedule offers a menu of options from which the customer can choose depending on its circumstances. Often these options take the form of two-part tariffs — although taken together for any one schedule, or considered in the aggregate for several schedules, they make the total available menu substantially nonlinear. The menu offered to a small customer may include fewer refinements than one offered to a large customer because the advantages of such refinements are not worth the combined cost to the utility and the customer of implementing them.

2. Nonlinear pricing occurs along several dimensions simultaneously. EDF’s tariffs, for example, use the dimensions of duration (length of utilization) and power (magnitude of load). Further incentives are offered for usage in off-peak periods, and for quick curtailment of usage at critical times of scarce supply. These dimensions reflect the underlying cost structure of the industry. In EDF’s case, the key tradeoff is between capacity and energy costs, and this in turn is determined by load duration. As predicted by the theory, quantity discounts are also offered to the largest customers, although possibly these too have a basis in the cost structure if large customers tend to have more stable loads with higher load factors.

3. The tariffs serve several purposes simultaneously. One purpose is to recover the utility’s costs of current operations (mostly energy costs) and amortization of costs of investments (mostly capital cost of generation and distribution equipment). In the long run, however, the tariffs also promote more efficient utilization of resources by customers. Because the tariffs reflect the long-run cost structure of power generation, they encourage customers to avoid wasteful usage. In EDF’s case, for example, the tariffs encourage customers to:

- Consume power in offpeak rather than peak periods, thus reducing investments in peaking equipment and reducing energy costs. Lower energy charges in offpeak periods provide this incentive.
- Curtail demand quickly in response to supply scarcity, thus diminishing peak loads. Lower energy charges for normal hours and much higher rates at critical times provide this incentive.
- Smooth consumption over time, thus increasing average duration of utilization and thereby altering the mix of power generation in favor of cost-effective
baseload sources. Lower energy charges (though higher annual demand charges) for longer utilizations provide this incentive.

In practice, few applications use nonlinear pricing in the elaborate detail and systematic conformity to costs implemented by EDF. Wright tariffs are rarely used in the United States currently although they provide incentives for consumption smoothing that are precisely consistent with the long-term cost structure of generation. Although it was popular in the early years of the industry, it fell into disuse until the recent introduction by a few utilities of so-called load-factor tariffs that are essentially Wright tariffs by another name. American regularity agencies in particular have opposed Wright tariffs, favoring instead a philosophy of “immediate causation” of short-term costs, and therefore time-of-use tariffs based on the capacity and operating costs of the marginal generator. These two kinds of tariffs are compared and studied in §11.

2.3. PG&E’s Curtailable Service

In California, Pacific Gas and Electric Company offers a tariff for large industrial customers that has explicit options for curtailable and interruptible power service. These options provide several quality dimensions related to service priority of the sort studied in §10.

Curtailable and interruptible service are two forms of priority service that differ mainly in the extent of load reduction that a customer is obligated to accept when the utility encounters a deficiency of generating capacity. Curtailable service requires the customer to reduce its load: although the load reduction is partly negotiable, PG&E requires that the load is reduced to at least 500 kW below the customer’s lowest average peak-period load among the previous six summer months (summer is PG&E’s peak season). Interruptible service requires the customer to eliminate its load entirely; further, the customer must accept automatic interruptions whenever the line frequency is deficient, which is often a signal that a generation shortage is imminent. Both options impose penalties if the customer fails to comply: these penalties are proportional to the excess load imposed by the customer, and they are substantially higher the second, and higher again the third time, within a year that the customer fails to comply. Both options limit load reductions to 6 hours on each occasion.

The service reliability associated with each option can be chosen by the customer from a menu of four alternatives. Each offers ordinary “firm service” as well as three
alternatives that are distinguished by limits on the utility’s prerogative to request a load reduction. These alternatives are shown in Table 2.1 as types A, B, and C under the firm-service tariff E20. Considering only curtailable service, for example, alternative A requires the utility to provide a warning 60 minutes in advance, to limit such curtailments to 15 times per year, and to limit the cumulative duration per year to 50 hours. Alternative C, in contrast, allows 10 minutes warning, 45 curtailments per year, and 200 hours cumulative duration. These specifications do not specify exact service reliabilities; instead, they are cast in terms of restrictions on easily verifiable actions of the utility. Altogether, the E20 tariff specifies seven different service conditions among which customers can choose.

The collection of tariffs shown in Table 1 can be interpreted as a single comprehensive tariff offering a menu of eleven options to each nonresidential customer. The charges in this super-tariff depend on three different dimensions:

- **Load size.** For example, the three firm-service tariffs A1, A10, and E20 offer successively lower marginal energy charges but successively higher fixed charges and demand charges. Also, A11 and E20 differ mainly in terms of the higher load that qualifies the customer for the lower energy charges under E20.

- **Time of use.** For example, A1 and A10 have energy charges independent of the period, whereas the other tariffs differentiate significantly among periods. A6 and A11 in particular provide substantial differences between peak and offpeak rates.

- **Service Reliability.** Alternatives A, B, and C under the E20 curtailable tariff offer successively lower demand and energy charges in all periods in exchange for acceptances...
### Table 2.1
Pacific Gas and Electric Company
Price Schedules

<table>
<thead>
<tr>
<th>Type</th>
<th>Fixed Chg. $/Mo.</th>
<th>Demand Peak $/kW</th>
<th>Energy Charge Peak Cents/kWh</th>
<th>Offpk Cents/kWh</th>
<th>Limits Warn Min.</th>
<th>Penalty Dur. No. Hrs.</th>
<th>1st $/kW</th>
<th>2nd $/kW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Firm Service:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A1</td>
<td>10.404</td>
<td>10.404</td>
<td>10.404</td>
<td>10.404</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A6</td>
<td>24.895</td>
<td>12.447</td>
<td>6.473</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A10</td>
<td>50</td>
<td>2.85</td>
<td>8.658</td>
<td>8.658</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A11</td>
<td>50</td>
<td>8.10</td>
<td>8.658</td>
<td>2.85</td>
<td>12.746</td>
<td>5.405</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E20</td>
<td>100</td>
<td>8.10</td>
<td>7.633</td>
<td>7.269</td>
<td>7.633</td>
<td>4.264</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Curtailable Service (E20):</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>290</td>
<td>4.87</td>
<td>2.85</td>
<td>7.631</td>
<td>7.266</td>
<td>4.265</td>
<td>60</td>
<td>15</td>
</tr>
<tr>
<td>B</td>
<td>290</td>
<td>3.04</td>
<td>2.85</td>
<td>7.626</td>
<td>7.261</td>
<td>4.264</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>C</td>
<td>290</td>
<td>2.85</td>
<td>7.494</td>
<td>7.136</td>
<td>4.226</td>
<td>10</td>
<td>45</td>
<td>45</td>
</tr>
<tr>
<td>Interruptible Service (E20):</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>300</td>
<td>2.85</td>
<td>6.886</td>
<td>6.557</td>
<td>4.128</td>
<td>60</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>B</td>
<td>300</td>
<td>2.85</td>
<td>6.392</td>
<td>6.088</td>
<td>4.036</td>
<td>30</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>C</td>
<td>300</td>
<td>2.85</td>
<td>5.701</td>
<td>5.428</td>
<td>3.908</td>
<td>10</td>
<td>45</td>
<td>45</td>
</tr>
</tbody>
</table>

of increasingly more numerous and longer curtailments with shorter warnings and higher penalties for noncompliance. The choice between curtable and interruptible service further differentiates in terms of the magnitude of the load reduction required.

The annual demand charges represent further differentiation that conforms with the analysis of capacity pricing in §11. Nine of the options impose the same demand charge for annual maximum demand ($2.85/kW) in association with a time-of-use schedule for energy charges. The firm-service tariffs A11 and E20 impose additional demand charges ($8.10/kW) for the peak-period peakload, but offer lower energy charges, as an incentive for load reduction in peak periods: these charges reflect the fact that in peak periods the customer’s load contributes directly to the system’s capacity requirements (the demand charge for the annual maximum is lower because capacity is a shared resource in non-
peak periods).\(^5\) Finally, these peak-period peakload demand charges are substantially discounted or eliminated if the customer elects curtailable or interruptible service. Curtailable or interruptible service enables the utility to avoid serving the customer in times of tight power supplies, which in turn reduces capacity requirements.

In sum, PG&E’s tariff structure reflects differentiation along several dimensions of quantity and quality simultaneously. The eleven tariffs shown in the table allow customers substantial choices in terms of load size and service reliability, and within nine of these tariffs the energy charges differentiate by time of use. It is, therefore, an example of a multidimensional tariff; moreover it includes capacity charges of the sort examined in §11.

The explicit differentiation in PG&E’s tariffs reflects a new development in the power industry. As recently as October 1985, for instance, all of Florida Power and Light’s fixed-rate and time-of-use nonresidential tariffs (except for demands below 20 kW) imposed the same demand charge ($6.50/kW) whereas energy charges were based only on the customer’s power demand, as shown in Figure 2.2. All of FPL’s curtailable tariffs allowed the same reduction ($1.70/kW) in the demand charge for the curtailable portion of the customer’s load, defined by “customer will curtail demand by 200 kW or more upon request of utility from time to time.”\(^6\) More recent developments at FPL and elsewhere emphasize tariff designs that are explicitly differentiated along several dimensions.

An example is the interruptible service plan proposed by Georgia Power Company.\(^7\) This plan provides an annual credit or rebate on the customer’s demand charge. For each kilo-Watt of interruptible power (in excess of the customer’s firm power level), this credit is proportional to a factor that varies on two dimensions. One dimension measures the number of interruptions allowed per year; thus, it is related inversely to the service reliability. The second dimension measures the “hours use of demand” for that kilo-Watt; thus, it is roughly proportional to the customer’s load factor. For example, a kW eligible for only 30 hours of interruption and used less than 200 hours per year earns a 23% credit on the demand charge, whereas at the other extreme a kW eligible for 240 hours of interruption and used more than 600 hours per year earns an 84% credit. The

---

\(^5\) Differentiated demand charges of this sort can also be interpreted as invoking the principles of a Wright tariff to encourage load leveling.

\(^6\) Source: Casazza, Schultz & Associates, *Electric Rate Book*, 1988. Penalties were also assessed for noncompliance with curtailment requests.

net effect is essentially a Wright tariff in which net demand charges decrease for more hours of interruption and/or more hours of usage. This is one version of the priority service tariffs addressed in §10.2.

2.4. MCI’s and AT&T’s Telephone Tariffs

The MCI Card tariff offered by the MCI Communications Corporation illustrates several of the complications involved in selecting the terms in which price schedules are quoted. 8

An important complication in telephone tariffs is that marginal prices ($/minute) depend on the origin and destination, as well as the time of day and the service mode, according to detailed schedules filed with the U.S. Federal Communications Commission. 8

Figure 2: FPL’s fixed and time-of-use energy charges.

8 MCI Tariff FCC #1, as quoted in “MCI Card Savings”, MCI Communications Corporation, February 1988.
There are many such combinations (MCI’s tariff FCC #1 is over an inch thick) and therefore the major vendors of long distance service usually provide customers only with illustrations of applicable rates for a few examples. These illustrations convey the dependence of rates on the distance involved, and the differing rates for the three main time periods: day (peak), evenings, and nights and weekends (offpeak).

Quantity discounts take two forms in the MCI tariffs. First, each call is charged according to a two-part tariff that in addition to the marginal charge per minute specifies a fixed charge per call ($0.55 for calls from touch-tone telephones). Second, a volume discount is offered depending on monthly billings of domestic daytime (and all international) calls: for a monthly bill between $50 and $100 this discount is 2%, and for bills over $100, it is 5%. Note that this discount applies to domestic calls only in peak periods, aggregated over different origin-destination pairs and different call durations by summing the dollar amounts of their charges. Further, the discount applies to the total dollar amount, rather than applying only to those charges exceeding a threshold. Presumably the simplicity of this tariff is designed to ensure compact presentation and easy interpretation by customers. In addition, it reflects the important feature that most customers place calls to so many destinations that more specific terms would be useless.

Other tariffs are also offered as options by the several interexchange carriers. For example, AT&T’s 1990 basic tariff for interstate Measured Toll Service (MTS) uses nearly uniform prices differentiated by three times of day, eleven distance categories, and surcharges by the type of call (more for operator-assisted and person-to-person). For a distance of 1000 miles, the day, evening, and night/weekend rates are $0.249, $0.1496, and $0.1300 per minute, respectively. But AT&T also offers several “Reach Out America” options. One such plan charges a monthly fee of $8.70 and provides one hour per month of free nighttime and weekend calls and lower marginal rates of $0.11 per minute thereafter, independent of distance, plus discounts on other calls depending on the time of day (10% day, 25% evening) or the destination (intrastate or international). The net effect of these two AT&T plans, the basic Measured Toll Service and the Reach Out America option, is a piecewise-linear tariff: customers obtain a quantity discount by subscribing for a Reach Out America plan. For calls made only in evening or daytime hours and

---

10 Mitchell and Vogelsang (1991, p.148) report that the first year (1984-85) of the Reach Out America plan, “increased the mean minutes of night/weekend calling by 41.9%. The overall price elasticity of the group of [optional calling plan] subscribers is of the order of
for a distance of 1000 miles, the Reach Out America plan is cheaper than Measured Toll Service if the number of hours per month exceeds four or six. MCI and Sprint offer similar options with slightly lower rates and smaller discounts, and Pacific Bell offers a similar Call Bonus plan for intrastate calls. In addition, in 1990 MCI offered an optional three-part tariff for daytime calls: for a monthly fee of $12 a customer receives one hour of free calls with additional usage billed at $0.20 per minute.

AT&T offers commercial customers an explicit piecewise-linear tariff for wide-area service (WATS). For daytime calls, the 1990 PRO WATS tariff provides discounts of 8% for monthly billings exceeding $200 and 21% for billings exceeding $2000. For larger customers with their own trunk lines, the Megacom WATS tariff charges $50 per month and provides discounts of 5% for billings over $7500 per month, and 10% for billings over $30,000 per month. These discounts apply to daytime rates at 1000 miles that, after the initial connection charge, are $0.173 per minute (in 6 second increments), compared to $0.249, $0.254, and $0.239 for MTS, PRO WATS, and High-Volume WATS. Multipart tariffs of this sort are addressed in §6.4.

The book by Mitchell and Vogelsang (1991) provides an excellent comprehensive description of the role of nonlinear pricing in the design of telephone tariffs, including many specific aspects not addressed here.¹¹

2.5. Federal Express’ Mail Rates

Companies offering express mail and facsimile transmission services use nonlinear pricing along several dimensions, including business volume, item size, and several dimensions of quality, such as speed and priority of delivery. In this section we describe the “discount” price schedule offered by Federal Express Corporation to high-volume customers for its Priority 1/Courier-Pak overnight delivery service, and for its Standard Air delivery service.

Table 2.2 shows the marginal prices per pound for these two services.¹² When the

¹¹ This area is changing rapidly. For example, on 1 August 1991 the Federal Communications Commission adopted new rules “loosening restrictions that have made it difficult for [AT&T] to offer volume discounts . . . for large corporate customers”; however, restrictions remain on local carriers wanting to reduce “charges for high-volume customers such as AT&T.” (San Francisco Chronicle, 2 August 1991, page B1.)

¹² Source: “U.S. Domestic Express Rates – Sample Only”, Federal Express Corporation, 1989. The per-package rates described are for Priority 1/Courier-Pak Discount Rates, providing next-day delivery, and for Standard Air Discount Rates. According to this
tariffs for these two services are graphed they show a somewhat complicated pattern. However, they show a nearly consistent pattern when graphed on a logarithmic scale, as shown in Figure 2.3. On a logarithmic scale the Courier-Pak schedule is revealed to consist essentially of the minimum of three logarithmically-linear tariffs. The first segment corresponds to the minimum charge of $9.00, the second to the interval between 1 and 50 pounds, and the third to the interval above 50 pounds. Similarly, the Standard Air schedule has two segments, the first corresponding to the minimum charge of $6 and the second to a logarithmically-linear schedule for items over 5 pounds. Note that for heavy items the difference between the two schedules consists of a fixed percentage.

### Table 2.2
Federal Express Discount Rates

<table>
<thead>
<tr>
<th>q-th Pound</th>
<th>Priority 1/Courier-Pak</th>
<th>Standard Air</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$9.00/Lb.</td>
<td>$6.00/Lb.</td>
</tr>
<tr>
<td>2</td>
<td>2.50</td>
<td>0.00</td>
</tr>
<tr>
<td>3-4</td>
<td>2.75</td>
<td>0.00</td>
</tr>
<tr>
<td>5</td>
<td>1.75</td>
<td>0.50</td>
</tr>
<tr>
<td>6-10</td>
<td>1.65</td>
<td>1.00</td>
</tr>
<tr>
<td>11-50</td>
<td>0.75</td>
<td>1.00</td>
</tr>
<tr>
<td>51-100</td>
<td>0.90</td>
<td>0.85</td>
</tr>
<tr>
<td>101-300</td>
<td>1.02</td>
<td>0.86</td>
</tr>
<tr>
<td>301-500</td>
<td>0.99</td>
<td>0.84</td>
</tr>
<tr>
<td>501-1,000</td>
<td>0.96</td>
<td>0.82</td>
</tr>
<tr>
<td>1,000+</td>
<td>0.90</td>
<td>0.77</td>
</tr>
</tbody>
</table>

2.6. Delta Airlines’ Frequent Flier Rebates

Soon after deregulation of the U.S. airline industry, “frequent flier” plans became a central part of the airlines’ marketing strategies. These plans offer rebates, in the form of free tickets, depending on the number of miles accumulated by the customer. The rebates schedule, the discount is the greater of $11.50 and 40% for Priority 1/Courier-Pak, and for Standard Air it is $10 up to 150 Lbs. or 14% for larger items. Not shown in the table is a maximum charge of $86 for 100 Lbs. with Standard Air service, which implies a zero marginal price for pounds 96-100.
vary nonlinearly with mileage, offering rebates that increase more than proportionately with the customer’s accumulated mileage. Because they impose costs on customers who might otherwise divide their travel among several airlines, frequent flier plans encourage a customer to use fewer airlines; that is, they impose switching costs. Incidentally, the rebates accrue directly to the traveler even if the fare is paid by an employer, and therefore they serve as an inducement to attract business travelers. The plans enable the airlines to monitor the frequency and length of each customer’s trips; consequently, further perquisites are offered to frequent business travelers who accumulate sufficient mileage within a specified period such as a year, especially on otherwise undiscounted fares.

Figure 2.4 depicts the rebates offered by Delta Airlines.\footnote{Source: “Frequent Flyer Program”, Delta Airlines, Inc., 1989.} This figure displays only

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure3.png}
\caption{Federal Express’ discount tariffs.}
\end{figure}
the mileage that is deducted for coach or economy class tickets; a fractional ticket, such as 0.25, indicates that the customer receives a 25% discount on the purchase of the ticket for the mileage specified. The rebate tickets allow any domestic U.S. origin and any destination (round trip or one way) served by the airline within one of the three regions (Domestic, Hawaii, Japan, or Europe) specified.

Except for the one kink at 30,000 miles, Delta’s domestic rebate schedule is essentially a two-part tariff (with a 20,000 mile fixed fee). This aspect is explicit in other cases. For instance, United Airlines’ Special Awards rebates in Winter 1991 required 8, 14, or 20 flight segments for 1, 2, or 3 domestic tickets, equivalent to a fixed fee of 2 segments and a marginal price of 6 segments for each ticket; Continental’s rebates were identical except that the marginal price of a fourth ticket was 4 segments.

The airlines also offer discounts on several other dimensions. Among the most
important are advance-purchase discounts, which play an important role in segmenting the market between business travelers and others. For example, Delta’s regular daytime midweek roundtrip coach fare between San Francisco and Atlanta in May 1989 was $1,104, or $884 at night, but with 3, 7, or 14 days advance purchase it was $940, $398, or $338. The last two fares were $40 higher on weekends, plus a further premium on popular flights both midweek and weekends. Similarly, one-way unrestricted coach fares were $552, $470, and $435 with 0, 3, and 7 days advance purchase. These fares are displayed in Figure 2.5 for a roundtrip midweek.

A second dimension is the demand for the flight, reflecting a kind of spot pricing or peakload pricing. Delta’s fares described above allowed for peak-demand premia by offering a menu of fares. The 14-day advance-purchase daytime midweek roundtrip coach fare of $338, for example, was one of three fares for otherwise identical service at
$338, $358, and $398, each with a quota of seats available: each customer was assigned to the lowest-priced unfilled fare class still available.

A third dimension is the percentage of the fare that is refundable if the trip is postponed or canceled after the fare has been paid. Delta’s fares described above with 0, 7, and 14 days advance purchase were 100%, 75%, and 0% refundable; consequently, some customers might choose the $398 fare, even if the purchase is made 14 days in advance, in order to obtain the higher percentage of refundability.

2.7. Summary

The illustrations in this chapter motivate the topics studied in this book. Readers familiar with pricing policies in the several industries described above may find these illustrations redundant because unraveling the intricate complexities of rate design is a constant challenge. For others, the pervasive role of nonlinear pricing may be a surprise since as a customer one is often unaware of the full menu of options involved when subscribing to power, telephone, express mail, and airline flights — as well as many other services such as rentals (equipment, vehicles, space, rooms) and banking and financial services — and products such as publications and capital equipment (computers, copiers). It is easy to focus on the option that meets one’s immediate need and to take little account of the overall design. One way to appreciate the implicit role of nonlinear pricing in the design of prices for increments in performance (speed, precision, durability) for a series of items in a product line is to examine a catalogue from an equipment manufacturer or a general supplier such as Sears. Similarly, quantity discounts and time-of-use differentiated rates are specified in your monthly bills from utilities, in the fares and frequent-flier rebates offered by airlines (and more recently hotels and rental car companies), and in the implicit price schedule for viewing films in a theatre, on video, on cable television, or “free” on commercial television.

Some MBA students in my course on Pricing have been stymied initially about finding a topic for a term paper on nonlinear pricing, but when encouraged to draw on personal experience, every one has found an interesting topic. The range extends from pricing package sizes of grocery items to the terms for corporate mergers, and from contracts for original-equipment auto parts to membership fees for a flying club. My favorite is an application to pricing space, power, and launch priorities for a proposed commercial laboratory facility in earth orbit.

Space limitations prevent a longer list of illustrations than those provided here from
the major regulated and recently deregulated industries. These were chosen because they fit the focus adopted in subsequent chapters. Parts I and II emphasize the advantages of nonlinear pricing as an efficient use of monopoly power to recover costs in regulated and other capital-intensive industries. Also, these illustrations display the varieties of multidimensional and multiprodut pricing addressed in Parts III and IV.
Chapter 3
MODELS AND DATA SOURCES

The illustrations in §2 are drawn from several different industries but they share important features. In each case the price schedule is differentiated along one or more dimensions so that lower marginal and average prices are offered for larger quantities. These discounts stimulate demand and enhance efficiency, provided the net prices exceed marginal cost. For instance, EDF’s lower marginal prices, and prices differentiated by service times and conditions, provide more accurate signals to customers about marginal costs of supply. Compared to a single price applicable to all units, profits increase; or in a regulated industry, the firm’s revenue requirement can be met with greater net benefits for customers.

Some concepts that enable unified analyses of nonlinear pricing in these and other industries are developed in this chapter and the next. This chapter presents the basic tools used subsequently. We emphasize the description of customers’ demand behaviors in terms of an appropriate aggregate measure of demands for successive increments in the quantity purchased, depending on the marginal prices charged. This measure, called the demand profile, provides the minimum disaggregated information about customer’s demands. This is the information needed to design nonlinear tariffs for either profit maximization in unregulated industries or efficient recovery of revenue requirements in regulated industries.

Models of customers’ demands can be formulated at various levels of aggregation, including levels corresponding to individuals and to more or less finely differentiated market segments. Models of individuals’ demands can be based on projected benefits from enduses or on specified demand functions, and this same range of modeling options is applicable to analyses of more aggregated market segments. Empirical estimation of demand behaviors depends similarly on the choice of a level of aggregation. Reliable data on individual demands are rarely available, however. It is important as a practical matter, therefore, to adapt the analysis to the most aggregated data that still provide the requisite information for tariff design. In this and the remaining chapters of Part I the
analysis is formulated in terms of this ideal level of aggregation of demand data that is appropriate for nonlinear pricing. Part II repeats this analysis at the fully disaggregated level of individual customers using explicit models of benefits or demand functions.

The advantages of nonlinear pricing stem ultimately from heterogeneity in the population of customers. For this reason, completely aggregated demand data are insufficient because they mask the relevant heterogeneity among customers. For instance, data about aggregate demand in the form of the total number of units sold at each price are insufficient because they cannot distinguish whether these quantities were bought by one or many customers; that is, they obscure the distribution of purchase sizes, which typically vary greatly among customers. Because the essential feature of nonlinear pricing is differentiation by size of purchase, on the other hand, one need not preserve information about other kinds of heterogeneity. It suffices to retain only data about the distribution of customers’ purchase sizes at each price.

The demand profile introduced in this chapter provides data at the right level of aggregation. It does so via a convenient interpretation. Data is typically accumulated by observing customers’ purchases in response to a specified price, and in this form the data specify directly the distribution of purchase sizes at each price. One can imagine a spreadsheet with each price assigned to a corresponding row and each purchase size assigned to a corresponding column. Then the data along each row records the distribution of purchase sizes observed at that price, namely the number of customers buying each successive incremental unit. The trick of interpretation is to realize that data in this format also provides down each column the demand function for this increment. That is, each column records the number of customers purchasing that increment at each of the prices indicated by the rows.

For nonlinear pricing, therefore, it suffices that demand data are disaggregated only to the extent of retaining information about customers’ (aggregate) demands for successive increments in the size of purchase. It is usually practical to obtain demand data at this level of aggregation. Fortunately, data about customers’ responses to uniform prices are sufficient: the firm need not have data about responses to nonlinear tariffs in order to examine the merits of nonlinear pricing. An important caveat is that the data must be sufficient to provide estimates of how demands will change as prices change. This is the binding constraint in practice whenever the firm’s prior experience provides little information about demands at prices other than the current one, or other means of estimating demand elasticities.
Section 1 reviews the basic approach to modeling demand behavior that underlies both linear and nonlinear pricing. Two subsections are designated by an asterisk * indicating that they contain optional technical material. Section 2 describes the data requirements and estimation procedures for the demand models commonly used in tariff design. Section 3 mentions some of the welfare considerations affecting regulated firms. Section 4 lists several cautionary considerations that affect nonlinear pricing generally, as well as those that pertain to the restrictive assumptions invoked in this exposition. An optional Section 5 summarizes technical assumptions used in subsequent chapters.

Regarding Mathematical Notation

This and later chapters use notation from calculus. Various marginal measures are represented by derivatives. For instance, marginal cost is the rate at which the firm’s total cost increases as the supply or output rate increases, and the marginal price is the rate at which the customer’s total charge (according to the tariff) increases as the purchased quantity increases. Similarly, aggregates such as sums and averages are represented by integrals. The firm’s profit contribution is the sum of the profit margins on the units sold, and customers’ aggregate benefit is the sum of the differences between the buyer’s valuation and the price charged.

Some readers may be deterred by this mathematical symbolism, but my experience has been that writing out the corresponding differences and summations is a greater deterrent because the formulas then appear awesomely long and complicated. Because this detracts from the main purpose of conveying the basic concepts of nonlinear pricing, a compromise has been adopted. The notation of the calculus is often used, but in Part I the results are not materially affected by this notation. Each marginal measure can be interpreted and evaluated as a ratio of discrete differences, and each integral can be interpreted and evaluated as a sum of discrete terms.

The reader can interpret each marginal measure as representing the increment obtained in the dependent variable if a small increment is made in the independent variable. For example, suppose \( U(q, t) \) indicates the gross benefit obtained by a customer described by a parameter \( t \) (indicating the customer’s “type”) from a purchase of size \( q \). Specifically, \( U(q, t) \) is the customer’s maximum willingness-to-pay in dollar terms for a purchase of size \( q \). Then the marginal benefit \( v(q, t) \) is defined as the (partial) derivative \( \frac{\partial U(q, t)}{\partial q} \) as the quantity \( q \) varies, leaving the customer’s type parameter \( t \) fixed. This marginal benefit is interpreted as the customer’s willingness-to-pay for the
3.1. Descriptions of Demand Behavior

The several levels of aggregation at which demand data can be analyzed are familiar from the process of setting uniform prices. Typically these include aggregation at the level of the entire market, segments of the market, and individual customers or types of customers. Each level of aggregation is useful for different purposes. For a profit-maximizing firm setting a uniform price, it may be sufficient to use aggregated market data to estimate the aggregate demand function — or perhaps only the price elasticity of this demand function to examine whether a price change would be profitable. However, analysis of various market segments or types of customers may reveal more information relevant to product differentiation. For regulated firms and public enterprises this finer information is necessary to assess the distribution of benefits among customers, which is a matter of special concern to regulatory agencies.

These same levels of data analysis are also useful for nonlinear pricing. Moreover, by recording and processing the data appropriately, essentially all the information required to design a nonlinear tariff can be obtained. Although we defer the description of how to construct an optimal nonlinear tariff to §4, it is possible to specify the informational requirements on the basis of general considerations that are applicable whether or not tariff design aims to optimize.

The Demand Function

When a uniform price is used, the major purpose of data analysis is ultimately to predict
the total quantity $\bar{D}(p)$ that would be sold at each price $p$. That is, $\bar{D}(p)$ is the demand function ordinarily used in applications of economic theory to rate design. For many purposes, moreover, it is sufficient to obtain an estimate of the price elasticity of demand, which indicates the percentage reduction in demand that would ensue in response to a one percent increase in the price. Interpreting the demand function as a differentiable function of the price, this elasticity is given by the formula

$$\eta(p) = -\frac{p}{\bar{D}(p)} \frac{d\bar{D}(p)}{dp}.$$  

For a discrete price change $dp$, the numerator on the right is the change in total demand: $d\bar{D}(p) \equiv \bar{D}(p + dp) - \bar{D}(p)$. This formula is sometimes written in the form

$$\eta(p) = -\frac{d\bar{D}(p)/\bar{D}(p)}{dp/p}$$

to convey the fact that it represents the ratio of percentage changes.

Estimating a demand elasticity is never trivial. Even if actual demand $\bar{D}(p)$ in response to the price $p$ is directly observed (and stable over time amidst many vicissitudes), an estimate of the demand elasticity requires that responses to other prices must be observed too. A firm may have tried other prices and therefore has such data, but more commonly the task of estimating responses to other prices requires the techniques of marketing research to obtain data from surveys or panel studies, or of benefit-cost studies to estimate customers’ willingness-to-pay based on analyses of end uses. The output of a demand analysis, such as the demand elasticity, is simple compared to the inputs mustered to obtain a reliable estimate.  

\[\text{This notation reflects a standard convention. A bar over a demand function represents the total quantity sold in response to a uniform price, or it can be interpreted as the average demand per potential customer. Later we use } \bar{D}(p; t) \text{ to represent the demand function for an individual customer of type } t, \text{ and } N(p; q) \text{ to present the number of customers purchasing at least } q \text{ units in response to the uniform price } p. \text{ Thus, demand functions indicated by a symbol } \bar{D} \text{ are measured in terms of units of the product sold, whereas the demand profile indicated by a symbol } \bar{N} \text{ is measured in terms of the number or fraction of customers purchasing an incremental unit.}\]

\[\text{Mitchell and Vogelsang (1991) describe AT&T’s repeated underestimates during the 1980s of the numbers of customers who would elect its “Reach Out America” optional calling plans. For a thorough (though now dated) survey of the methodology and several empirical studies that estimate the price elasticity of aggregate demand for electricity, see Taylor (1975). Koenker and Sibley (1979) calibrate demand estimates for their study of nonlinear pricing of electricity against two of the elasticity estimates reported in Taylor’s survey. For recent studies of demands for telecommunication services, see De Fontenay, Shugard, and Sibley (1990).}\]
The Demand Profile

Nonlinear pricing requires the same kind of data, but in a partially disaggregated form. Recall that a nonlinear price schedule charges a possibly different price for each increment in the customer’s purchase size. The required prediction is therefore, for each increment (say the $q$-th unit) the demand for that increment at its marginal price $p(q)$. This demand is usually measured, not in terms of units of the product, but rather in terms of the number of customers purchasing this increment. That is, unlike an ordinary uniform-price demand function $D(p)$ that measures the number of units sold at the uniform price $p$, we measure the number or fraction $N(p, q)$ of customers purchasing the $q$-th unit at the marginal price $p$. This measurement convention is especially useful when considering increments in the purchase size that are not units of size 1: if an increment includes $\delta$ units, then the demand for that increment is $N(p, q)\delta$.

Similarly, the price elasticity of demand for the $q$-th increment at the marginal price $p$ is

$$\eta(p, q) \equiv -\frac{p}{N} \frac{\partial N}{\partial p}.$$  

In this definition, $\partial N/\partial p$ indicates the (partial) derivative of $N$ as the price is changed without changing the unit $q$ to which that marginal price applies. Both $N$ and $\partial N/\partial p$ are evaluated at $(p, q)$. In discrete terms, this price elasticity of the demand profile is just the percentage change in demand for the $q$-th unit divided by the percentage change in its price: thus $\partial N \equiv N(p + \partial p, q) - N(p, q)$ for a price change $\partial p$.

The bivariate function $N(p, q)$ is called the demand profile. It has a central role in tariff design. One reason is that the demand profile includes more information than the demand function does. To see this, observe that the total demand in response to a uniform price $p$ is

$$D(p) = N(p, \delta)\delta + N(p, 2\delta)\delta + \cdots,$$

$$= \sum_{k=1}^{\infty} N(p, k\delta)\delta,$$

where $\delta$ is again the size of an increment and the index $k$ indicates the $k$-th increment purchased. If quantities are finely divisible then we can interpret the increment $\delta$ as infinitesimally small even though the quantity $q = k\delta$ is not. In this case total demand is represented as the integral

$$D(p) = \int_0^{\infty} N(p, q) dq.$$  

57
This formula represents a sum over all values of the quantity \( q \) of the number \( N(p, q) \) purchasing the (infinitesimally) small increment at \( q \) that is of size \( dq \equiv \delta \).

These formulas express the fact that total demand is the sum of the numbers sold of first units, second units, et cetera. Thus, from the demand profile one can always reconstruct the demand function for a uniform price. Even with a nonuniform price schedule \( p(q) \) the same sort of formula is true: total demand is

\[
\sum_{k=1}^{\infty} N(p(k\delta), k\delta) \delta \quad \text{or} \quad \int_{0}^{\infty} N(p(q), q) dq.
\]

The revenue is similarly

\[
\sum_{k=1}^{\infty} N(p(k\delta), k\delta) \cdot p(k\delta) \delta \quad \text{or} \quad \int_{0}^{\infty} N(p(q), q) p(q) dq.
\]

The integral on the right represents the aggregate of the charges \( p(q) dq \) collected from the number \( N(p(q), q) \) of customers purchasing the increment of size \( dq \) at the price \( p(q) \) per unit for each quantity \( q \).

**Interpretations of the Demand Profile**

We use two interpretations of the demand profile repeatedly. Both derive from a basic feature: nonlinear pricing can be construed as pricing a product line comprising the incremental units of the product. Thus, the first, second, \ldots, units of the generic product are treated as separate products priced separately. The fact that each unit’s purchase requires buying its predecessors in the sequence is implicit; for example, a second unit requires prior or simultaneous purchase of a first unit.

The first interpretation of the demand profile \( N(p, q) \) is that it describes the distribution of purchase sizes in response to each uniform price \( p \). That is, if \( p \) is fixed and \( q \) is increased then a graph of \( N(p, q) \) depicts the declining number of customers purchasing each successive \( q \)-th unit. This is the same as the number of customers purchasing at least \( q \) units, so for each fixed price \( p \) the demand profile \( N(p, q) \) is the right-cumulative distribution function of the customers’ purchase sizes. In summary:

- For each price \( p \) the demand profile specifies the number or fraction \( N(p, q) \) of customers purchasing at least \( q \) units.

Because this number is observable, the demand profile can be measured directly from demand data. This requires, of course, that information about customers’ purchase sizes
is accumulated and recorded. It is in this sense that the demand profile is a disaggregated version of the demand function: the demand data is disaggregated according to size of purchase.

Figure 3.1 shows how the distribution of purchase sizes is predicted from ordinary models of customers’ benefits in which each customer is described by a parameter \( t \) indicating his type. Assuming a particular benefit function \( U(q,t) \) and tariff \( P(q) \), the diagram plots a customer’s net benefit from each purchase size \( q = 0, \ldots, 4 \) as a function of the type parameter \( t \). For simplicity, this function is assumed to be linear and to increase as the type parameter increases. As \( q \) increases, a customer’s gross benefit increases. But for each fixed type \( t \) there comes a point where the next increment in the gross benefit is less than the increment in the tariff, namely the marginal price for the next unit, whereupon he ceases to purchase additional increments. The diagram therefore shows for each type \( t \) the net benefit corresponding to the optimal purchase size, which is the largest quantity for which the marginal benefit of the last unit purchased exceeds the marginal price charged. Such a model predicts that the customers purchasing exactly \( q \) units are those with type parameters between \( t_{q-1} \) and \( t_q \) as shown in the figure. Those purchasing the \( q \)-th increment are all those purchasing at least \( q \) units, namely those described by type parameters exceeding \( t_{q-1} \). Conversely, the firm can use the observed distribution of purchase sizes to estimate the distribution of types in the population: again, the number actually purchasing at least \( q \) units is the estimated number with type parameters exceeding \( t_{q-1} \).

The second interpretation of the demand profile is that it describes the distribution of customers’ willingness to pay. If we fix \( q \) and increase the marginal price \( p \), then the demand profile measures the declining number of customers willing to pay this price for the \( q \)-th increment. Thus:

- The demand profile specifies for each \( q \)-th unit the number or fraction \( N(p,q) \) of customers willing to pay the price \( p \) for that unit.

This interpretation is useful in applications based on parameterized models of customers’ behavior. For instance, suppose market segments, volume bands, or customer types are indicated by a parameter \( t \), and a benefit-cost analysis indicates that type \( t \) obtains the dollar benefit \( U(q,t) \) from purchasing \( q \) units, and therefore a marginal benefit \( v(q,t) = \partial U(q,t)/\partial q \) from the \( q \)-th unit. Then the demand profile \( N(p,q) \) measures the estimated number of customers whose types \( t \) are such that \( v(q,t) \geq p \), indicating that they are willing to pay at least \( p \) for the \( q \)-th unit.
Figure 1: Purchase sizes $q$ selected by customers of different types $t$. Customers with larger type parameters obtain their maximum benefits from larger purchases.

The two interpretations of the demand profile are represented schematically in Figure 3.2. The number of customers purchasing at least $q^*$ units at the price $p^*$, and the number willing to pay at least $p^*$ for the $q^*$-th unit, are the same because they are both measured as the number of customers whose demand functions intersect the shaded region of price-quantity pairs $(p, q) \geq (p^*, q^*)$. Thus, $N(p^*, q^*)$ is the number of customers with types $t \geq t^*$, where $t^*$ is the type that is indifferent along both boundaries of the shaded region. In particular, the demand function of type $t^*$ passes through the point $(p^*, q^*)$.

In summary, the demand profile can be interpreted as the distribution of purchase...
3.1. Descriptions of Demand Behavior

Figure 2: Measurement of the demand profile from customers’ demand functions. At \( (p^*, q^*) \) the demand profile measures the number of customers purchasing more than \( q^* \) at price \( p^* \), or purchasing \( q^* \) at prices exceeding \( p^* \).

sizes \( q \) for each uniform price \( p \), say\(^3\)

\[
N(p, q) = \# \{ t \mid D(p, t) \geq q \},
\]

where \( D(p, t) \) is the demand function of customers of type \( t \). Or, it can be measured as the distribution of marginal valuations for each unit \( q \), say

\[
N(p, q) = \# \{ t \mid v(q, t) \geq p \}.
\]

\(^3\) The notation \# indicates the number of customers satisfying the stated condition, not the number of types per se. The two usages are equivalent if the number of customers per type is constant across types.
These two interpretations are used in different ways. The first is most useful for analysis of demand data; the second, for applications of benefit-cost models. In addition, for analyses of welfare effects, the second allows one to infer the distribution of customers’ net benefits.

The following two subsections present more technical renditions of the demand profile that can be skipped.

**The Implicit Type Model** *

To simplify exposition, later chapters in Parts I and II focus on models in which the heterogeneity among customers is described by a single type parameter that varies in the population — although §8.4 demonstrates that this restriction is immaterial for most practical purposes. We can anticipate that result here by reconsidering the basic meaning of a type parameter.

Type parameters are never observed directly; moreover, they have no fundamental units of measurement outside a particular parameterization of customers’ benefit or demand functions: any monotone transformation of a type parameter will again serve the same practical purposes. Further, their ultimate role is only to predict the customer’s purchase size, which is a one-dimensional quantity. Thus, there is a fundamental sense in which one can expect that a single type parameter should suffice for practical purposes.4

This conjecture can be made more precise as follows. Suppose the demand profile is measured in terms of fractions of the population rather than absolute numbers of customers. We can define an implicit type parameter by the property that a customer of type \( t \) is one who purchases \( q \) if the price is \( p \) provided \( p, q \) and \( t \) are related by the equality \( t = 1 - N(p, q) \). If we further assume that these types are uniformly distributed in the population, then we obtain a model that accounts exactly for the observable demand behavior.5 Thus, the exposition in §6 concentrates initially on the case of one-dimensional type parameters because this is largely sufficient for practical analysis of single-product pricing problems.

---

4 This applies only to tariffs for single products. With multiple products, multiple type parameters are useful for realistic description of customers’ demand behaviors. Typically one wants at least as many type parameters as there are quantity and quality dimensions to a customer’s bundle of purchases.

5 The demand function for each implicit type can be inferred as the solution of a differential equation derived from the definitions above.
More General Versions of the Demand Profile *

Use of the demand profile \( N(p, q) \) described above invokes several implicit assumptions. In this subsection we clarify what these assumptions are by describing the general formulation and how the demand profile is derived as a special case.

Our aim in a general formulation is to specify the demand for \( \theta \)-th increments induced by a tariff \( P \). A customer of type \( t \) prefers to purchase a \( \theta \)-th increment if his preferred purchase includes \( q \) or more increments. This says that there exists some purchase \( x \geq q \) such that the net benefit from \( x \) exceeds the net benefit from any purchase \( y < q \). In mathematical notation, therefore, the demand for \( \theta \)-th increments is

\[
N(P, q) \equiv \# \{ t \mid (\exists \ x \geq q)(\forall y \not\geq q) U(x, t) - P(x) \geq U(y, t) - P(y) \}.
\]

In principle, this demand depends on the entire tariff schedule \( P \). For most of the applications we address, however, the only relevant values of \( x \) and \( y \) are \( x = q \) and \( y = q - \delta \), where \( \delta \) is the size of an increment. That is, to know whether type \( t \) purchases the \( q \)-th increment it is sufficient to know whether the \( q \)-th increment provides an incremental benefit that exceeds its price. This defines the special form of the demand profile for increments of size \( \delta \):

\[
N(p(q, \delta), q; \delta) \equiv \# \{ t \mid [U(q, t) - U(q - \delta, t)]/\delta \geq p(q, \delta) \},
\]

where \( p(q, \delta) \equiv [P(q) - P(q - \delta)]/\delta \) is the marginal price per unit of the \( q \)-th increment of size \( \delta \). The basic demand profile defined previously is just the special case in which increments are arbitrarily small:

\[
N(p(p), q) \equiv \lim_{\delta \to 0} N(p(q, \delta), q; \delta), \quad \text{where} \quad p(q) \equiv \lim_{\delta \to 0} p(q, \delta).
\]

For this demand profile to be an accurate prediction of customers' demand behaviors, it is sufficient that the price schedule and customers' demand functions are nonincreasing, and that the price schedule intersects each demand function just once, from below. This assures that a customer's marginal benefit from an incremental purchase is positive for purchases less than the optimal purchase, and negative for purchases exceeding the optimal purchase; thus, buying the \( q \)-th increment is beneficial if and only if \( q \) is less than the optimal purchase. Assuring that this property is true is therefore an auxiliary

---

6 Read \( \exists \) as “there exists . . . such that” and \( \forall \) as “for all.”
requirement that must be imposed on the design of the optimal price schedule. To enforce this requirement, later chapters either impose assumptions sufficient to ensure that it holds automatically, or specify remedies that constrain the price schedule sufficiently that it becomes true.

Other versions of the demand profile are used for special purposes. For instance, suppose that the tariff imposes a minimal charge $P_\delta$ that provides a minimal purchase $q_\delta$. The number of customers willing to pay this minimal charge is

$$N_\delta(P_\delta, q_\delta) \equiv N(P_\delta / q_\delta, q_\delta; q_\delta),$$

where in this case the relevant increment is $\delta = q_\delta$ and the average price per unit of the minimal purchase is $P_\delta / q_\delta$.

A more general master profile is used occasionally to take account of both the total tariff and the marginal price schedule:

$$M(P(q), p(q, \delta), q; \delta) \equiv \# \{ t \mid U(q, t) \geq P(q) \& [U(q, t) - U(q - \delta, t)] / \delta \geq p(q, \delta) \},$$

and $M(P, p, q) \equiv \lim_{\delta \to 0} M(P, p, q; \delta)$. For instance, the demand for a minimal purchase (as above) is

$$N_\delta(P_\delta, q_\delta) \equiv M(P_\delta, 0, q_\delta; q_\delta).$$

Or, if a two-part tariff $P(q) = P_0 + pq$ is used then the demand for a $q$-th increment is

$$N_\delta(P_0, p, q) \equiv M(P_0 + pq, p, q),$$

of which a special case is the basic demand profile $N(p, q) \equiv N_0(0, p, q)$ as originally defined.

Other kinds of demand profiles adapted to various special purposes can be constructed. In each case, the demand profile summarizes the demand for an increment (of some positive size $\delta$, or the limiting case as $\delta \to 0$) based on assumptions about the relevant values of $x$ and $y$ in the general definition that are (or for simplicity, are assumed to be) relevant to the problem addressed. For the main topics addressed here, it suffices to consider only the basic demand profile $N(p, q) \equiv N_0(0, p, q)$ as originally defined: the next section describes how it is estimated from demand data.\footnote{The definitions of the various demand profiles can be extended straightforwardly to encompass several products; cf. §12.1 and §14.1. In this case, $x \geq q$ in the definition of $N$ means $x_i \geq q_i$ for each product $i$ represented in the bundle $q = (q_1, \ldots, q_n)$ of quantities of the products.}

---

64
Regarding notation: to ease exposition we do not distinguish between the demand profiles $N(p, q; \delta)$ and $N(p, q)$ when the size of the increment $\delta$ is presumably small or is obvious from the context.

### 3.2. Estimation of the Demand Profile

The demand profile is comparatively easy to estimate when the firm has ample demand data obtained from a variety of prices. We first describe how this can be done using the interpretation that the demand profile represents the distribution of purchase sizes at each uniform price. When the firm has not previously used other prices than the current one, however, estimation of the demand profile or its price elasticity is usually based on a parameterized model of customers’ benefits. Consequently, we also present a second procedure that relies on the interpretation that the demand profile represents the distribution of customers’ marginal valuations of incremental units.

### Direct Measurement of the Demand Profile

This first approach supposes that the firm has used several uniform prices in the past. Moreover, the firm has had the foresight to record for each price the distribution of customers’ purchase sizes over a standard billing period. For instance, during the 1980s most long-distance telephone rates declined steadily (in both real and nominal terms), producing thereby a wealth of data about customers’ usage in response to different rate levels — as well as some novel forms of nonlinear tariffs. Typically this data is disaggregated by customer class (commercial, residential), location (urban, rural), and product (regular MTS service, WATS lines). In addition, most firms are able to record a rich array of additional data that can aid demand estimation, such as line-of-business data for commercial customers and socio-demographic data for households.

Represent the several prices (or price levels, in real terms) that have been charged by the increasing sequence $p_j$, where $j$ is an index that distinguishes the different prices; for example, cents per minute per mile in the case of a telephone company. Similarly, represent the possible purchase sizes by an increasing sequence $q_k$, where $k$ is an index that distinguishes several volume bands of total usage per billing period; for example, minute-miles per month of long distance calls, in the case of a telephone company. Corresponding to each price $p_j$ the data provides a measure $n_{jk}$ of the number or

---

8 For illustrations of the construction of telephone customers’ usage distributions see Pavarini (1979), and Heyman, Lazorchak, Sibley, and Taylor (1987).
fraction of potential customers whose purchase size was in volume band \( k \) when the price was \( p_j \). The direct estimate of the demand profile is then

\[
N(p_j, q_k) = \sum_{t \geq k} n_{jt},
\]

at these prices and quantities. That is, if the data in the array \((n_{jk})\) are represented as a spreadsheet with rows indexed by prices and columns indexed by volume bands, then the demand profile is the new array obtained by replacing each element by the sum of the elements further along the same row.

Usually, however, this estimate is insufficient because it does not cover all the possible prices and quantities that might be relevant for rate design. A variety of methods are available to obtain a smooth estimate of the demand profile. For example, a so-called spline curve can be fitted to the direct empirical estimates (Press et al. (1986), §3); or the empirical estimates can be construed as data for use in a regression equation whose parameters are then estimated by the usual statistical techniques of regression analysis. If a regression model is used, then for each price \( p_j \) an estimate is made of \( N(p_j, q) \) construed as a function of the purchase size \( q \); or the regression can be applied to a bivariate function of both the price and the quantity. A hypothetical example is shown in Figure 3.3 using a linear function at each price. For each price \( p = 10, \ldots, 16 \) the data points represent the distribution of customers’ purchases. The regression lines are chosen to minimize the sum of squared errors.

In subsequent examples we describe some of the functions that are commonly estimated to obtain this sort of smooth approximation of the demand profile. Even so, it is a good practice to initiate, and later to confirm, rate studies with analyses based on the direct empirical estimates; only later might one examine refinements based on smoothed estimates of the demand profile.

Another useful step is to compare the two estimates of the demand profile based on its two interpretations. The second interpretation conforms exactly to the first if the alternative direct estimate

\[
N(p_j, q_k) = \sum_{h \geq j} n_{hk}
\]
closely approximates the first direct estimate. Discrepancies between these two estimates might plausibly be resolved in favor of the first, however. The reason is that the second relies on the presumption that customers’ behaviors maximize their net benefits, whereas


3.2. Estimation of the Demand Profile

![Diagram showing the number of customers purchasing a certain quantity or more at different prices.](image)

**Figure 3:** Estimation of the demand profile from the distribution of purchase sizes. The four prices yield different distributions of purchase sizes. In the figure, estimates of these distributions are obtained by linear approximations.

...the first does not invoke this assumption as strongly and is therefore relatively immune to idiosyncratic behavior by customers.

**Indirect Measurement of the Demand Profile**

When the extent of price variation has been small, a firm must rely on an indirect approximation of the demand profile derived from estimation of customers’ demand functions or benefit functions. Usually this is done by hypothesizing that each customer (within a relatively homogenous class, such as residential customers) is described sufficiently by a list $t$ of type parameters. These parameters define a measure $U(q, t)$ of the customer’s benefit, which is usually specified as an explicit function of the quantity $q$ in which the
parameters in the list $t$ enter as coefficients. For instance, a very simple model supposes that the function $U$ is quadratic of the form

$$U(q, t) = t_1q - \frac{1}{2}t_2q^2,$$

using the two type parameters $t = (t_1, t_2)$. From such a model one can derive the marginal benefit function $v(q, t)$ of type $t$. And by inverting the relation $p = v(q, t)$ to solve for $q$, this can be converted to the predicted demand function $D(p, t)$ of type $t$ in response to the uniform price $p$. For the example above, this yields

$$v(q, t) = t_1 - t_2q \quad \text{and} \quad D(p, t) = \frac{1}{t_2}[t_1 - p],$$

both of which are linear (in $q$ or $p$).

Converting this model into an estimate of the demand profile requires an auxiliary datum, which is the distribution of types in the population of potential customers. This estimate is usually specified in the form of a density function $f(t)$, or in a discrete version, as the proportion of each type $t$ in the population. An equivalent specification is a distribution function $F(t)$ indicating the number or fraction of potential customers with type parameters not exceeding $t$. The estimate of the demand profile is then obtained as the sum or integral

$$N(p, q) = \int_{T(p, q)} f(t) \, dt,$$

or

$$N(p, q) = \int_{T(p, q)} dF(t),$$

where $T(p, q)$ is the set of types $t$ for which $v(q, t) \geq p$ or $D(p, t) \geq q$. For example, for the quadratic model above,

$$T(p, q) = \{ t \mid t_1 - t_2q \geq p \},$$

which comprises all the types $(t_1, t_2)$ on one side of the line for which $t_1 - t_2q = p$. This line represents those customers who are indifferent about purchasing the $q$-th increment at the price $p$.

Given a model of customers’ benefits specified in terms of their type parameters, the remaining empirical problem is to estimate the distribution of types in the population of potential customers. This is done in two steps. The first is to specify a form of the
distribution function $F(t; \beta)$ that depends on a list $\beta$ of parameters. These parameters appear as coefficients in an explicit model of the type distribution, and similarly the demand profile $N(p, q; \beta)$ depends on these coefficients. The second step is to estimate the coefficients using data about the distribution of purchase sizes in response to various prices that have been offered.

\textbf{Example 3.1}: To illustrate the first step, suppose the distribution function is supposed to be a bivariate Normal distribution with parameters $(\mu, \sigma, \rho)$ representing the means, variances, and correlation of the type parameters $t = (t_1, t_2)$ in the population. Then, for the quadratic model above, $N(p, q; \beta)$ is the probability $\Phi(x)$ that a standard Normal random variable (one with mean 0 and variance 1) exceeds the number $x$ defined as

$$x \equiv [p - (\mu_1 - \mu_2 q)]/\Delta(q; \sigma, \rho), \quad \text{where} \quad \Delta(q; \sigma, \rho) \equiv \sqrt{\sigma_1^2 - 2\rho \sigma_1 \sigma_2 q + \sigma_2^2 q^2},$$

in which $\sigma_i^2$ is the variance of $t_i$, and $\rho$ is the correlation between $t_1$ and $t_2$. The distributional coefficients are therefore the five parameters $\beta \equiv (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ of the type distribution.

The second step is the statistical task of estimating these parameters from observed distributions of purchase sizes for various prices. From the first step, the predicted relationship among the number $N$ of customers purchasing the $q$-th increment at the price $p$ is $y = [p + \mu_2 q - \mu_1]/\Delta(q; \sigma, \rho)$ where $y \equiv \Phi^{-1}(N)$. The data consists of observations of triplets $(y; p, q)$ and the aim is to find the estimate of the parameters in the list $\beta$ that provides the best fit of the data to the predicted relationship. Any of the variety of standard procedures for statistical estimation can be used for this task.

As in the example, a nonlinear regression model is often used to estimate the parameters of the type distribution. In such a model the independent variables are the prices $p_j$ and the volume bands $q_k$, and the dependent variable is $N(p, q; \beta)$, for which the observed data points are the numbers $(\sum_{t \geq k} n_{jt})$ of customers purchasing at least $q_k$ at the price $p_j$. An ordinary least-squares estimate of $\beta$, for instance, chooses the estimate $\hat{\beta}$ that minimizes the sum of squared deviations from the predicted relationship:

$$\sum_j \sum_k \left[ N(p_j, q_k; \beta) - \sum_{t \geq k} n_{jt} \right]^2.$$

Statistical software programs to calculate such estimates are widely available. Other approaches to this sort of estimation include maximum likelihood estimation and probit
and logit techniques, as well as a variety of others included in standard statistical software. The estimation is considerably simplified if the demand function has a single type parameter that specifies the (constant) price elasticity; in this case one needs only to estimate the parameters of the distribution of this elasticity in the population.

A familiar part of standard statistical methodology controls for so-called nuisance parameters; thus, additional independent variables are included to account for characteristics of the customer (for example, line and size of business or number of household members) and incidental characteristics of the product (season of the year or time of day, type of service contract). Including such variables can improve the forecasting accuracy of the model, although for rate design one usually wants eventually to use the model with these nuisance parameters set at their averages for the population to which the price schedule is offered to all customers on the same terms. If separate demand profiles are estimated for various market segments, but these segments are offered the same tariff, then for rate design one uses the sum of the segments’ demand profiles. A fortunate aspect of the demand profile is that it enters linearly into the specification of the firm’s costs and revenues. Consequently, it is sufficient, and indeed optimal, to use a simple unbiased estimator of the demand profile since that will in turn yield an unbiased estimate of costs and revenues.

Additional sources of information can also be gleaned from socio-demographic data (for households) and line-of-business data (for commercial customers) by using the techniques of correlation analysis and its variants. For example, correlation analysis of purchase size distributions with regional socio-demographic averages (for example, size of household, income, number of appliances) can be used to extend observations from pilot programs, survey data, or panel studies to other regions.

3.3. Welfare Considerations

Many of the important applications of nonlinear pricing are in the major regulated industries, such as power, communications, and transport. In these industries it is especially important to recognize the distribution of net benefits among customers. Nonlinear pricing need not improve the welfare of all customers if it is applied indiscriminately. In particular, because the price schedule usually declines as the purchase size increases,

---

10 An exposition of the use of probit models for related kinds of estimation is provided by McFadden and Train (1991). Detailed applications of these models to electricity and telephone pricing are reported by Train et al. (1987, 1989).
customers making small purchases prefer the uniform price that raises the same revenue for the firm when that price is less than the nonlinear price schedule’s charges for initial units. In §5 we analyze the form of nonlinear pricing that is often adopted to avoid disadvantaging small customers. Called Pareto-improving nonlinear pricing, it refers to the increasingly common practice of offering at least two price schedules: each customer can choose whether to purchase at the uniform price (small customers prefer this) or the nonlinear price schedule that offers quantity discounts (large customers prefer this). This policy assures that no customer is disadvantaged by the adoption of a nonlinear tariff. When options of this sort are offered, small customers are unaffected by nonlinear pricing but large customers (and the firm) can benefit substantially.

There are two approaches to measuring the incidence of benefits among customers. One method relies on an explicitly parameterized model of customers’ types that specifies directly the net benefit each type obtains. Consequently, the firm’s profit contribution and consumers’ benefits can be aggregated by summing these amounts pertinent to each customer. The second, indirect method uses the demand profile to infer the number of customers purchasing each increment. This enables the firm’s profit contribution and the consumers’ surplus to be obtained by aggregating over increments.

**Surplus Measurement via Parameterized Models**

Suppose for simplicity that the firm’s total cost is simply the sum of the costs it incurs to serve individual customers. Then its profit contribution, or producer’s surplus, can be represented by a formula of the form:

\[
\text{Producer’s Surplus} \equiv \int_{a}^{b} \left[ P(q(t)) - C(q(t)) \right] dF(t).
\]

In this formula, \(q(t)\) is the purchase selected by type \(t\), for which the customer pays the tariff \(P(q(t))\) and the cost incurred by the firm is \(C(q(t))\). We use the interval \(a \leq t \leq b\) to represent the range of type parameters in the population.

The customer’s purchase satisfies the demand condition \(q = D(p(q), t)\), or equivalently \(p(q) = v(q, t)\). This purchase provides type \(t\) with the net benefit \(U(q(t), t) - P(q(t))\) and the aggregate obtained by all customers is

\[
\text{Consumers’ Surplus} \equiv \int_{a}^{b} \left[ U(q(t), t) - P(q(t)) \right] dF(t).
\]

Figure 3.4 depicts the division of the total surplus between the producer’s and consumer’s surplus for one customer. The seller’s profit on each unit is the marginal price collected,
Figure 4: Division of the total surplus between the producer’s and consumer’s surplus. The latter are measured as the areas between the price schedule and the schedules of the firm’s marginal cost and the customer’s demand.

less the marginal cost, so the producer’s surplus is the area between the schedules of marginal price and marginal cost. Similarly, the consumer’s surplus is the difference between the customer’s marginal valuation $v(q,t)$ and the marginal price $p(q)$ so the total on all units purchased is the area between the demand function $D(p,t)$ and the price schedule.

**Surplus Measurement via the Demand Profile**

The second approach measures the firm’s profit contribution using a different accounting convention. In this version, the firm observes that a profit margin $p(q) - c(q)$ is obtained from each of the $N(p(q), q)$ customers who buy the $q$-th increment. Summing these
§3.3. Welfare Considerations

amounts for all increments yields an alternative statement of the profit.

This technique is called integration by parts or in a discrete version, summation by parts. The effect of integrating by parts is represented in Figure 3.5 for the term representing the firm’s revenue. The left diagram represents the revenue as the tariff \( P(q(t)) \) charged for the purchase \( q(t) \) by type \( t \). This is then multiplied by the number of customers of this type and then summed over all types. The dashed lines represent the translation from types to their purchases. The right diagram represents this revenue as the marginal price charged for each unit times the number of customers purchasing this unit, as represented by the corresponding horizontal slice of the area representing the revenue. The number of \( q \)-th units purchased is the number of customers purchasing at least \( q \) units: if the translation between \( q \) and \( t \) is increasing, then this is the number of customers with types exceeding that type \( t(q) \) purchasing exactly \( q \) units; namely, the type for which \( v(q, t(q)) = p(q) \).

A similar procedure applies to the construction of the consumers’ surplus. One exploits the fact that, according to its second interpretation, the demand profile specifies the distribution of customers’ marginal valuations or reservation prices for each incremental unit. For a fixed value of \( q \), the number of customers barely willing to pay the marginal price \( p \) for the \( q \)-th unit is the number who would cease buying the \( q \)-th unit if the price were raised slightly, which is measured by the resulting decrease in \( N(p, q) \). Interpreting the demand profile as the right-cumulative distribution function of reservation prices, the associated number or density of customers with the reservation price \( p \) for the \( q \)-th unit is therefore \( -\partial N(p, q)/\partial p \).

Consequently, the total consumers’ surplus from purchases of the \( q \)-th unit at the marginal price \( p(q) \) is

\[
\int_{p(q)}^{\infty} (p - p(q))[-\partial N(p, q)/\partial p] \, dp \equiv \int_{p(q)}^{\infty} [p - p(q)] \, d[-N(p, q)] \\
= \int_{p(q)}^{\infty} N(p, q) \, dp.
\]

The second line restates the first (using again the method of integration by parts) to express this consumers’ surplus as the area under the demand profile to the right of the

---

11 This is not the same as the number purchasing exactly \( q \) units, which is \( dN(p(q), q)/dq \). This is calculated as a total derivative that measures the rate at which \( N \) declines as \( q \) increases and \( p \) decreases due to the change in \( q \). This is necessary to account for how the predicted number is affected by the slope of the price schedule.
Figure 5: Two versions of the firm’s revenue. The representation in the right panel results from integration by parts: it accumulates the revenues obtained from successive increments in the purchase size.

marginal price $p(q)$ charged for the $q$-th increment. In words, the first formula credits $p - p(q)$ to consumers’ surplus for each of the customers willing to pay more than the price $p(q)$ charged. The second uses the alternative accounting scheme in which $1$ is credited for each of the $N(p, q)$ customers willing to pay more than the higher price $p$. Thus, customers’ net benefit from purchases of the $q$-th unit is the sum of the number willing to pay the next higher price $p(q) + 1$, the number willing to pay $p(q) + 2$, et cetera, each weighted by their contribution $dp = 1$ per customer to the aggregate surplus.
3.4. Cautions and Caveats

The total consumers’ surplus from purchases of all increments is therefore
\[ \int_0^\infty \int_{p(q)}^{\infty} N(p, q) \, dp \, dq. \]

The welfare measures used in later chapters can be computed from this formula and the preceding formula for the producer’s surplus.

Another aspect of welfare analysis recurs repeatedly in later chapters. At times the analysis addresses optimal pricing by a profit-maximizing monopolist, at others it considers monopolistic competition and oligopolistic competition, and it also examines the problem of efficient pricing to meet the revenue requirement of a regulated firm. Although these appear at first to be distinct problems, they are essentially similar because they are all special cases of the formulation of Ramsey pricing developed in §5.1. The theme of Ramsey pricing is to choose the tariff to maximize the total of consumers’ surplus subject to meeting a revenue requirement for the firm. A special case is pricing by a profit-maximizing monopoly if the revenue requirement is so severe that this is the only scheme that will raise sufficient revenue. Further, if the revenue requirement is not so severe, Ramsey pricing suggests that the regulated firm should price as though it were one of several firms in an oligopolistic market, which has the effect of reducing its profits down to the level of the revenue requirement. These ideas are set forth in more detail in §5.1 and §12.3.

Thus, the analysis of profit and efficiency motivations for rate design can be combined into a unified framework. It is a general principle that pricing must be done efficiently if it is to raise the maximum revenue. There is always a further dimension of customer satisfaction and the distribution of benefits among customers. Our main analysis of this dimension is in §5.2, where we explain the design of so-called Pareto-improving tariffs assuring that no customer is disadvantaged by nonlinear pricing as compared to a revenue-equivalent uniform price. However, in some cases it is necessary to impose further constraints on the distribution of benefits among customers, and typically these constraints incur losses in allocative efficiency.

3.4. Cautions and Caveats

There are numerous amendments that are bypassed in this exposition, but it is worth mentioning some briefly.

1. Customers’ benefits and demand behaviors are assumed to be exogenous; that is, their behaviors are unaffected by the introduction of nonlinear pricing. Although
this may be a valid assumption in the short run, it is usually false in the long run. For example, the introduction of appreciable quantity discounts for large purchases may induce some customers to alter end-uses and their investments in appliances and production technologies. Although these adaptations typically take time, if only because of the durability of capital equipment, eventually they alter demand behavior. For this reason, measures of demand responses to quantity discounts based on existing responses to uniform prices can severely underestimate the eventual changes in demand. Rate design can anticipate these secondary responses by benefit-cost analyses of appliances and production equipment in major end uses, and consideration of adoption delays due to behavioral factors.

2. Customers’ benefits are assumed to be denominated in money terms. A more general approach allows income effects, risk aversion, impatience or discounting of delayed benefits, and other behavioral parameters. These are omitted here for simplicity, because a sufficient exposition would entail vastly more technical detail and notational complexity. Also, they are rarely of prime importance for rate design in the industries that use nonlinear pricing intensively.

3. The exposition of parameterized models here and in §6 assumes that the firm knows or can estimate the distribution of types in the population. In fact, this distribution is usually variable and at any one time the firm can usually only estimate the underlying probabilistic process by which customers’ types are created or evolve over time. This simplification is adopted because nonlinear pricing is usually applied to markets with large stable populations of customers, for which the population distribution (that is, the histogram of types in the population) differs little from the long-term average distribution. Further, nonlinear pricing schedules are usually offered over relatively long time spans during which the average distribution is the primary concern.

4. Customers are assumed to obtain benefits directly from consumption of the product. In fact, many customers are other firms who are intermediaries in production or distribution. If they operate in competitive markets then their benefits are passed on to their retail customers via lower prices. But if they operate in imperfectly competitive markets allowing some monopoly power then they may be able to retain some of the gains as profits. We bypass this complicated subject until it is taken up briefly in §5. Further details are in Brown and Sibley (1986, Chapter 6).

5. The firm is assumed to be a monopoly. In fact, several amendments are necessary
if the firm operates in an imperfectly competitive market. Estimates of the demand profile and customers’ benefits are always conditioned on the prices prevailing for other products, especially those that are close substitutes or complements, as well as general economic conditions. This consideration is particularly important when competing firms also use nonlinear pricing, and it is severe when each firm’s price schedule is a response to others’ schedules. A single firm that offers several products encounters similar problems described in Part IV.

6. In all chapters where we address the problem of cost recovery by a regulated firm, we assume for expositional purposes that it is sufficient to consider only the case that the tariff is designed to maximize the total of customers’ net benefits subject to a net revenue requirement for the firm. In fact, one can also consider the possibility that some types of customers are given greater weight in the measurement of aggregate welfare. We ignore this possibility here because it complicates the exposition without adding significant new insights into the application of nonlinear pricing. But all the methods described in these chapters can be extended to address such considerations if necessary.

The gist of these provisos is that nonlinear pricing is a vast subject to address in all its myriad versions. Our exposition is intended mainly to convey the basic principles rather than to include all the variations in a single comprehensive theory.

The following optional section lists some basic assumptions of a mathematical character that are used in subsequent chapters.

3.5. Standard Assumptions *

No single set of assumptions provides a uniform standard for all the applications of nonlinear pricing. This section lists some basic assumptions that prevail throughout subsequent chapters and that are largely sufficient for most practical applications. Additional assumptions are imposed in later chapters addressing particular topics. The exposition generally emphasizes the formulation of nonlinear pricing problems and the main ideas involved in their analysis. The key analytical characterizations are conveyed mainly in the form of first-order necessary conditions for an optimal solution. Mathematically sophisticated readers will notice that as a result the treatment of sufficiency conditions is abbreviated, and sometimes absent altogether. My partial remedy for this deficiency is the exposition in §8.1 and 2.
The sufficiency assumptions invoked in the journal literature are often much stronger than necessary, and possibly their restrictiveness and daunting complexity have impeded applications. On the other hand, the recent work by Milgrom and Shannon (1991) provides the exact necessary and sufficient condition in terms of the property called quasi-supermodularity. I have not revised the exposition to cast the construction in terms of this condition, but it appears that in the future this will be the right framework for the development of the theory of nonlinear pricing.

The Firm’s Cost Function

Except in §15, we assume throughout that the firm’s cost depends only on the distribution of purchase sizes among customers. Serving a customer \( i \) incurs a specific cost \( C(q_i) \) that depends on the quantity \( q_i \) that \( i \) purchases, but not directly on the identity \( i \) of the customer, nor on any type parameters that describe the customer.\(^{12}\) In particular, any two customers purchasing the same quantity impose the same costs on the firm.

The firm may incur an additional cost that depends on the aggregate quantity supplied to all customers, but for expositional simplicity we usually assume that the firm’s marginal cost of aggregate supply is constant: §4.2 indicates how more general costs structures can be included. The cost function \( C \) is generally assumed to be nonnegative, increasing, and convex except possibly for a fixed cost of access or hookup. (However, many of the results apply equally to the case of decreasing marginal cost if care is taken to avoid local optima.) The marginal cost \( c(q) \equiv C’(q) \) is therefore nondecreasing. We bypass the role of increasing returns to scale by assuming that for a firm in a regulated industry a revenue requirement is specified sufficiently large to recover the full costs of operations and investments in capacity.

With a few exceptions, we do not distinguish supplies by the firm and purchases by customers according to date, location, contingencies, or other conditions of delivery. Thus, cost is interpreted as an average over the billing period and an expectation over possible contingencies. The exceptions, mostly in Part III, pertain to cases in which the conditions of delivery are interpreted as quality attributes of the product.

The simplest version of these assumptions is used frequently for examples: the firm’s marginal cost is constant. Often we simplify further by assuming this marginal cost is zero.

\(^{12}\) A distinguishing feature of the applications of nonlinear pricing to principle-agent problems and related contexts is that the seller’s cost depends on the customer’s type.
§3.5. Standard Assumptions *

An unfortunate deficiency of the exposition is that insufficient attention is given to the particular structure of the costs of capacity, production, and distribution in the main industries that use nonlinear pricing. Also, little account is provided about the role of general features such as economies of scale or scope. For expositions that include more detail about such aspects I recommend Mitchell and Vogelsang (1991) regarding the telecommunications industry, and Joskow and Schmalansee (1983) regarding the electric power industry.

The Demand Profile

We show in §4 that an optimal tariff can be constructed from the demand profile. The relevant assumptions on the demand side, therefore, impose structure on the demand profile. We describe here only the assumptions pertinent to the construction of a nonlinear tariff for a single product with a single quantity dimension: supplementary assumptions are made in Parts III and IV.

Much of the exposition assumes that the demand profile depends only on the marginal price, and not the total tariff. This is a restrictive assumption — and generally false because it ignores the fact that some customers will cease purchasing as the tariff is raised; that is, it ignores the role of market penetration. Consequently, in §4.4, §6.7, and §7.4 we address explicitly the choice of the fixed access fee that accompanies the tariff, and optimize it too in a fashion that recognizes the effect on market penetration. It suffices initially to ignore fixed fees because (as shown in §6.7) when the firm incurs no setup cost from serving a customer, an optimal tariff imposes no fixed fee.

The demand profile \( N(p, q) \) is nonnegative and decreasing in the quantity variable \( q \) by definition, and we generally assume further that it is also decreasing in the price variable \( p \). This corresponds to the familiar intuition that a higher price restrains demand. We also assume that \( N(p, q) = 0 \) if the price \( p \) or the quantity \( q \) is sufficiently large; that is, potential demand is bounded. To apply calculus methods, the demand profile should also be twice differentiable. To ensure that the necessary condition for an optimum is also sufficient to identify a global optimum, the profit contribution (defined in various ways in different chapters) must have a single local maximum; that is, it is unimodal or quasi-concave. This is violated by some models used in practice, so one must be careful to discard inferior local optima.

\[13\] This simplifies the exposition considerably. §8.5 shows that extensions to more general formulations are possible but complicated.
Simple versions of these assumptions are used for many of the examples in later chapters. For instance, the simplest model has a linear demand profile of the form
\[ N(p, q) = 1 - ap - bq, \]
and by choosing the units of measurement appropriately, the coefficients \( a \) and \( b \) can be taken to be 1. An example of a demand profile that violates an assumption is
\[ N(p, q) = 1 - p^aq^b, \]
because it remains positive for every \( q \) if the price is zero.

### Customers’ Benefit and Demand Functions

Customers’ benefit and demand functions are assumed to be very regular. As a function of the quantity \( q \), each benefit function \( U(q, t) \) is nonnegative, increasing, bounded, concave, and twice differentiable. As a function of the customer’s list \( t \) of type parameters, it is assumed to be increasing in each component of \( t \). This is partly a convention, since type parameters can often be redefined appropriately by rescaling or by some other transformation. The substantive assumption is that by some redefinition of the type parameters the benefit function can be made monotone.

The customer’s marginal valuation function \( v(q, t) \equiv \partial U(q, t)/\partial q \) is also assumed to be increasing in the type parameters. This is a strong assumption since it says in effect that the customers’ demand functions are strictly ordered in terms of their type parameters, independently of the quantity purchased. This is a strongly sufficient assumption to ensure that the resulting demand profile predicts correctly how customers will respond to the price schedule. It is far from necessary, however, as we explain in §8.1 and §8.4.

Various benefit functions and marginal valuation functions are used in the numerical examples. The quadratic benefit function mentioned in Section 2 is a simple version with a linear marginal valuation function. The benefit function with a demand function having a constant price elasticity \( 1/[1 - bt] \) is \( U(q, t) = aq^{bt} \), and by choice of units of measurement the coefficients \( a \) and \( b \) can be taken to be 1.

### The Type Distribution

To obtain the requisite properties of the demand profile, the types’ distribution function must also satisfy some regularity conditions. If there is a single type parameter, and the distribution function \( F(t) \) has a density function \( f(t) \equiv F'(t) \), then an amply sufficient condition is that \( F \) has an increasing hazard rate, meaning that the ratio \( h(t) \equiv f(t)/(1 - F(t)) \) increases as \( t \) increases, or is constant. This condition is not very restrictive, since it is satisfied by nearly all the distribution functions commonly used in empirical work. These include the Normal, exponential, and uniform distributions that

\[ 80 \]
§3.5. Standard Assumptions *

we use often in examples — as in §6.6. Analogous assumptions for distribution functions of multiple type parameters are cumbersome to state: §13 relies on the assumptions used in the basic work by Mirrlees (1976, 1986). The net effect of these assumptions is to avoid bunching of customers at particular purchase sizes.

In §6 we establish that only the ratio of the type elasticities of the marginal benefit function and the right-cumulative distribution function affect the construction of an optimal tariff. Consequently, §8.1 replaces the assumptions that the marginal valuation and the hazard function are increasing functions of the type parameter with the weaker assumption that the ratio of these type elasticities is decreasing.

The Tariff and the Price Schedule

The tariff is generally required to be nonnegative and increasing. Practical considerations often require that it is also subadditive. This means that the charge for several small purchases is no less than the charge for the purchase that is the sum of the smaller amounts. This property prevents a customer (or arbitrageur) from circumventing the tariff by dividing a large purchase into several smaller ones. A strongly sufficient condition for this property is that the tariff is concave, or equivalently the marginal price schedule has no increasing segments. In §4 and §6 we show how an optimal price schedule that does not initially satisfy this condition can be altered so that it is nonincreasing as required.

A further implicit constraint is that each customer’s predicted purchase is an optimal response to the tariff offered. This requires that a customer’s purchase satisfies both the first and second-order necessary conditions as well as global optimality. The usual form of a customer’s first-order necessary condition is that the marginal valuation of an incremental unit equals the marginal price, and the second-order condition states further that increasing the quantity purchased would decrease rather than increase the net benefit obtained. The latter requires essentially that the tariff is less concave than the customer’s benefit function. Alternatively, it states that the customer’s demand function intersects the price schedule from above — and only once, to ensure global optimality. Versions of this constraint imposed directly via assumptions on the benefit or demand functions are called single-crossing conditions in the technical literature. This constraint is satisfied automatically by many commonly-used parameterized models, but §8 shows how to extend the analysis to more general models.

The net effect of the assumptions placed on customers’ benefit functions and on the type distribution is to ensure that the demand profile constructed from them has
the required properties. In practice, empirically estimated demand profiles usually have the right properties. It is possible in principle that the demand profile could be non-monotone, but such cases are far removed from realistic predictions of customers’ behavior. Nevertheless, parameterized models pose a hazard in that poorly formulated versions are capable of anomalous predictions.

3.6. Summary

This chapter introduces the basic apparatus used subsequently to describe customers’ demand behaviors in response to nonlinear pricing. With ordinary uniform pricing the basic datum required for rate design is an estimate of the aggregate demand function. Nonlinear pricing requires disaggregated data that indicates demands for the different increments in the purchase size. This requirement reflects the primary feature of nonlinear pricing that it is a kind of product differentiation in which successive increments are interpreted as distinct products. Disaggregated data also preserve relevant information about the heterogeneity of customers’ demand behaviors.

The cogent summary of demand data is the demand profile. It can be interpreted as specifying either how the distribution of purchase sizes varies as the price changes, or how the distribution of customers’ valuations of an incremental unit varies as the quantity changes. The first interpretation is convenient for estimating the firm’s profit contribution from a nonlinear tariff, and the second, the consumers’ surplus.

The characterizations developed in this chapter for the single-product contexts studied in Parts I and II are generalized in Part III to study products with multiple quality attributes, and in Part IV, to multiproduct tariffs.
Chapter 4

TARIFF DESIGN

A nonlinear tariff represents a special kind of product line. For each increment purchased during the billing period a customer is charged a corresponding price specified in the schedule of marginal prices. Thus, the price schedule differentiates among increments. This chapter uses the product-line interpretation to construct the optimal price schedule from customers’ demands for increments.

We assume in this chapter that the firm is a profit-maximizing monopoly seller of a single product. Section 1 demonstrates that the optimal schedule of marginal prices can be constructed by choosing each increment’s price to maximize that increment’s profit contribution. For this purpose, the demand profile $N(p(q), q)$ is interpreted as representing the demand for $q$-th increments when the marginal price is $p(q)$. Section 2 adds some technical refinements. Section 3 sketches the considerations involved in the construction of a fixed access fee to accompany the schedule of marginal prices; however, no fixed fee is charged if the firm incurs no fixed cost in serving an individual customer. Section 4 reconsiders these results in terms of the bundling interpretation of product differentiation. Section 5 extends the analysis to the design of multipart tariffs with block-declining price schedules.

The exposition in this chapter uses elementary methods – except for some optional material in Section 2 indicated with an asterisk (*). It presents the main ideas in tariff design without technical details. Advanced topics receive abbreviated treatment, and in particular some aspects summarized here are not precise unless further assumptions are specified. Subsequent chapters examine these topics in more detail, but they also rely on more complicated analysis.

4.1. The Price Schedule of a Monopolist

For a single product, a monopolist would ordinarily choose the price to maximize the profit contribution, which is the profit margin times the demand at the chosen price. The data required for this calculation are the firm’s cost schedule and estimates of demand
or demand elasticities at each price. Similar data are required to determine prices for a product line of differentiated quantities but the demand estimates must be disaggregated according to purchase sizes.

We assume here that the available data provide estimates of demand for each of several prices denoted by $p$ and several purchase sizes denoted by $q$. These estimates can be represented as a tabular array in which the rows correspond to the different prices, the columns correspond to the different purchase sizes, and an entry in the table indicates the number or fraction $n(p, q)$ of customers who at the price $p$ will purchase $q$ units. This is the form in which demand data is ordinarily accumulated if appropriate care is taken to record customers’ purchases: for each price $p$ that has been offered the firm observes the distribution of purchase sizes among customers. A first trial of nonlinear pricing typically begins with data from only a few prices used in the past; in this case, estimates must be based on inferences about customers’ demand elasticities, as described in §3. Care must also be taken to recognize the role of the billing period and the dimensions in which purchase sizes are measured. For example, purchase sizes must be interpreted consistently in terms of the number of units purchased within the assigned billing period; thus the rate, say units per month, is the relevant measure.

The key idea in the analysis recognizes that the tariff can be interpreted as imposing a different charge for each successive increment in the purchase size. Thus, a tariff represented as a schedule in which $P(q)$ is the total amount charged for a purchase of size $q$ can also be represented as a schedule in which a price $p(q)$ per unit is charged for the $q$-th increment in the purchase size. For example, if the possible purchase sizes are the integral amounts $q = 1, 2, \ldots$ then $p(q) = P(q) - P(q - 1)$ is the price charged for the $q$-th unit. If increments are of size $\delta$ then $p(q) = [P(q) - P(q - \delta)]/\delta$ is the price per unit charged for the $q$-th increment, namely the last increment to reach $q$ units. In practice, the same price is usually charged for a range of increments, as in a block tariff, but here we allow initially that a different price is charged for each increment.

This idea is implemented by further recognizing that a customer buying the $q$-th increment must also buy all lesser increments. This feature is essentially a consequence of the exclusion of resale markets. The demand for each increment is the number of customers purchasing amounts that require that increment. This is the same as saying that the demand for the $q$-th increment is the demand for all purchase sizes at least as large as $q$. Using the tabular array $n(p, q)$, the demand for the $q$-th increment at the
price \( p \) is therefore the number

\[
N(p, q) = \sum_{x \geq q} n(p, x).
\]

of customers purchasing \( q \) or more.

The tabular array \( N(p, q) \) is the demand profile defined and described in §3. Each row of \( N \) is a list of partial sums from the same row of \( n \). For the purpose of constructing optimal tariffs, the demand profile can be measured as either the number of customers or the fraction of potential customers demanding the \( q \)-th increment at the price \( p \).

It is important to realize that the demand profile differs from the aggregate demand obtained from a single uniform price \( p \), which is

\[
D(p) = \sum_q n(p, q)q
= \sum_{q=\delta, 2\delta, \ldots} N(p, q)\delta,
\]

where \( \delta \) is the increment in the purchase size and \( p\delta \) is charged for each increment. The demand profile is a disaggregated version of the total demand function that preserves the information about customers’ purchases sizes in response to each uniform price.

Similarly, when a nonincreasing price schedule \( p(q) \) is offered the number of customers purchasing the \( q \)-th increment is predicted to be \( N(p(q), q) \). Consequently, the aggregate demand is

\[
Q = \sum_q N(p(q), q)\delta,
\]

which counts the sum of the numbers of customers purchasing each increment \( q = \delta, 2\delta, \ldots \).

An alternative representation of the demand data that preserves even more information is the collection of demand functions \( D_i(p) \), one for each customer \( i \) in the population. That is, \( D_i(p) \) indicates the purchase size chosen by customer \( i \) at the uniform price \( p \) for each unit. If the data is available in this form, then the demand profile can be constructed by interpreting \( N(p, q) \) as the number of customers \( i \) for

\[1\] Later examples assume that the demand profile measures fractions of the number of potential customers in the population. The units of revenue measures, such as the firm’s profit or consumers’ surplus, are therefore dollars per billing period per potential customer in the population.
whom \( D_i(p) \geq q \). The shorthand notation we use for this is

\[
N(p, q) = \# \{ i \mid D_i(p) \geq q \}.
\]

The demand profile is used in the following construction because it embodies the minimal information needed to find the optimal tariff. It is important to realize, however, that its use involves an implicit assumption. To ensure that \( N(p(q), q) \) will in fact be the realized demand for the \( q \)-th increment requires that a customer who would purchase at least \( q \) units at the uniform price \( p = p(q) \) will also purchase at least \( q \) units when offered the entire schedule of prices for increments. Typically the price schedule is nonincreasing and so is the maximum price that each customer is willing to pay for the \( q \)-th increment; reliance on the demand profile to summarize the demand data imposes the further requirement that the schedule of prices for increments intersects each customer’s demand function at most once, and from below. In practice, this qualification is usually met without difficulty, but in Section 2 we present more details about its role in the analysis.

To complete the formulation, we assume initially that the firm’s cost schedule reflects a constant cost \( c \) per unit, or \( c\delta \) for an increment of size \( \delta \). In particular, the variable cost of supplying a purchase \( q \) is \( cq \), and if \( m(q) \) customers select the purchase size \( q \) then the total variable cost is \( C(Q) = cQ \) where \( Q = \sum_q m(q)q \) is the total quantity supplied.

The total profit contribution expected from a price schedule \( p(q) \) for increments can therefore be expressed as the sum of the profit contributions from the various market segments differentiated according to the increments in the purchase size:

\[
Pft = \sum_q N(p(q), q) \cdot [p(q) - c]\delta.
\]

In this case, to find the optimal price schedule it suffices to find the optimal price for each market segment separately. That is, the optimal price for the \( q \)-th increment is the one that maximizes that increment’s profit contribution:

\[
R(p(q), q) = N(p(q), q) \cdot [p(q) - c].
\]

The following simple example illustrates the calculation.

\[\text{Example 4.1} : \text{Suppose that the possible prices and purchase sizes are the ones shown in Table 4.1, which tabulates the demand profile. Each column corresponds to a}\]

86
### Table 4.1

**Demand Profile for Example 4.1**

**Optimal Tariff for Marginal Cost** \( c = \$1 \)

<table>
<thead>
<tr>
<th>( q )</th>
<th>( p ) (in units)</th>
<th>( D(p) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>$2/\text{unit}</td>
<td>90</td>
<td>75</td>
</tr>
<tr>
<td>$3</td>
<td>80</td>
<td>65</td>
</tr>
<tr>
<td>$4</td>
<td>65</td>
<td>50</td>
</tr>
<tr>
<td>$5</td>
<td>45</td>
<td>30</td>
</tr>
</tbody>
</table>

| \( p(q) \) : | \$4 | \$4 | \$3 | \$3 | \$2/\text{unit} | \$4 |
| \( P(q) \) : | \$4 | \$8 | \$11 | \$14 | \$16 |
| \( R(p(q), q) \) : | \$195 | \$150 | \$90 | \$40 | \$5 |
| Total Profit : | \$480 | \$450 |
| ‘CS’(q) : | \$45 | \$30 | \$40 | \$5 | \$0 | \$120 | \$85 |
| ‘TS’(q) : | | | | | | \$600 | \$535 |

Market segment for a product line in which the purchase sizes or “volume bands” are 1, 2, 3, 4, and 5 units with \( \delta = 1 \). The entry in the first column and the first row, for instance, shows that at least one unit is purchased by 90 customers if the price is \$2 per unit. (The number who purchase exactly one unit is 15 since 75 purchase at least two units.) At the bottom of the first column is shown the optimal price \$4 charged for the first unit, assuming that the marginal cost of supply is \( c = \$1 \). This optimal price is found by calculating, for each price \( p = \$2, \ldots, \$5 \), the profit contribution. The profit contribution from the price \$2, for instance, is the demand of 90 units at this price times the profit margin \( \$2 - \$1 = \$1 \), yielding the total contribution \$90, which is less than the contribution of \( 65 \times (\$4 - \$1) = \$195 \) obtained from the optimal price of \$4 per unit. Similarly, the profit contribution from the third increment (\( q = 3 \)) is \$55, \$90, \$90, or \$40 from the four possible prices, and the maximum of these is \$90, obtained with either the price \$3 or the next higher price \$4, although we have indicated that the lower price is chosen. The entries in the table that correspond to optimal choices of the prices are underlined. The optimal prices shown in the table imply a tariff \( P(q) \) that charges \$4, \$8, \$11, \$14, and \$16 for the purchase sizes 1, 2, 3, 4, and 5, respectively.

When using numerical data as in the table, it is useful to check the demand profile.
for consistency. In addition to the requirement that the demand profile is nonincreasing along each row and down each column, there is the further requirement that the optimal schedule of marginal prices predicts a consistent pattern of purchases. For instance, on the assumption that customers respond to lower marginal prices with larger purchases, it would be inconsistent that a decreasing marginal price schedule predicts, say, a demand for fourth units that exceeds the demand for third units. In the example, this consistency check is satisfied because the demand for each successive unit is less than the previous one. That is, the predicted demands for first, ..., fifth units are 65, 50, 45, 20, and 5, respectively, when the optimal price schedule is offered. Consistency can be violated when the units of measurement used in the tabulation exceed the actual increments chosen by customers.

Also shown in the table is the aggregate demand function $\bar{D}(p)$ that indicates the total demand if the firm charges a uniform price $p$ for all units. Note that $\bar{D}(p) = \sum_{q=1,...,5} N(p, q)$. The optimal uniform price is the one that maximizes the aggregate profit contribution $\bar{D}(p) \cdot [p - c]$. For the example shown in the table, the optimal uniform price is $4, which yields a profit contribution of $450 from sales of 150 units. In contrast, the optimal tariff yields a higher profit contribution of $480 from sales of 185 units, corresponding to an average price of $3.60 per unit. Also shown in the table are minimal estimates of consumers’ and total surplus, denoted CS and TS, based on the demand profile’s indication of the numbers of customers willing to pay higher prices than those charged; for example, 45 customers are willing to pay a dollar more than the $4 charged for the first unit.

The advantage of nonlinear pricing can be seen in the table. Offering price breaks for large purchase sizes stimulates demand that would otherwise be choked off by the higher uniform price. This particular tariff has the further advantage that it increases the firm’s profit without disadvantaging any customer, since the optimal price schedule charges no more for each unit than does the optimal uniform price. In §5 we show that this feature can be generalized: whenever the uniform price exceeds marginal cost there exists a nonuniform price schedule that raises the same net revenue for the firm, increases the net benefits for some customers, and does not reduce the net benefit of any customer.

The gains from nonlinear pricing evident in this example stem from heterogeneity among customers. Segmenting the market into volume bands enables the seller to offer lower prices for the larger purchase sizes selected by some customers. The discounts are
4.1. The Price Schedule of a Monopolist

equally available to all customers, but only those customers demanding larger purchases take advantage of the opportunity.

The demand profile summarizes the heterogeneity among customers at the coarsest level of aggregation that still allows analysis of nonlinear tariffs. As described in §3, for each specified price it indicates the distribution of purchase sizes demanded by customers; and for each specified increment it indicates the distribution of customers’ reservation prices. In applications it is often useful to represent this information in terms of the price elasticities of demands for different units. For instance, a price \( p^o \) yields a higher profit contribution than another price \( p \) if

\[
N(p^o, q)[p^o - c] > N(p, q)[p - c].
\]

This can be stated equivalently when \( dN \equiv N(p^o, q) - N(p, q) > 0 \), and therefore \( dp \equiv p^o - p < 0 \), by the condition that

\[
\frac{p^o - c}{p^o} > \left[ \frac{dN/N}{-dp/p} \right]^{-1},
\]

or the reverse inequality if the signs of \( dN \) and \( dp \) are reversed. The optimal price \( p(q) \) therefore provides a percentage profit margin that is approximately equal to the reciprocal of the price elasticity \( \eta(p, q) = [dN/N]/[-dp/p] \) of the demand profile for the \( q \)-th unit:

\[
\frac{p(q) - c}{p(q)} \approx \frac{1}{\eta(p(q), q)}.
\]

The price elasticity measures the percentage increase in demand per percentage decrease in the price. Thus, this condition says that the percentage profit margin times the rate at which demand decreases per unit decline in the percentage margin should be 1, indicating that the profit from existing customers that is lost by decreasing the price is compensated by the profit on new demand it brings from other customers.

Characterizations of this kind are exact if we interpret the price \( p \) as a continuous variable and the demand profile is differentiable. The price that maximizes the profit contribution \( N(p, q)[p - c] \) in this case satisfies the necessary condition

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

This condition states that the firm’s marginal profit from a small change in the price charged for the \( q \)-th unit is nil. In particular, a $1 increase in the price obtains an
additional revenue of $1 from each of the $N$ customers who purchase this unit, but it also loses the profit margin $p(q) - c$ from each of the $|\partial N/\partial p|$ customers who respond to this price increase by refraining from purchasing this unit. The price elasticity of the demand profile in this case is

$$\eta(p, q) \equiv -\frac{p}{N(p, q)} \frac{\partial N}{\partial p}(p, q),$$

so the necessary condition can be stated equivalently as

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

That is, for the optimal marginal price $p(q)$, the percentage profit margin on $q$-th units is precisely the reciprocal of the price elasticity of the demand profile. This "inverse elasticity rule" is common to all monopoly pricing situations but here it is applied separately to each market for incremental units.

Typically the induced price-elasticity of the demand profile increases as $q$ increases, so the optimal price schedule $p(q)$ is decreasing and the optimal tariff $P(q)$ is a concave function of the purchase size. This can be seen in extreme form for the last unit $q^*$ purchased by any customer: $N(p(q^*), q^*) = 0$ for this value of $q$, but $\partial N/\partial p < 0$, so $p(q^*) = c$. That is, the last unit is sold at marginal cost. In some cases, however, the price schedule computed in this way can have increasing segments or folds; we mention in Section 2 the amendments required when this happens.

An alternative statement of the optimality condition requires that the price schedule satisfies $\frac{d}{d_q} R = \frac{\partial}{\partial_q} R$ all along the locus $(p(q), q)$ of the price schedule. That is, as $q$ increases.

---

2 This is quite different than saying that the price elasticity of customers’ demands increases. For the parameterized models studied in §6, the price-elasticity of the demand profile is the product of the type-elasticity of the type distribution and the ratio of the price and type elasticities of the demand functions of customers purchasing exactly the amount $q$. Even if each customer’s demand function has a constant price-elasticity, the price-elasticity of the demand profile varies with the ratio of the type-elasticities of the type-distribution and customers’ demand functions, and indeed must eventually decline to zero for customers with the highest types. Higher price-elasticities of demand for customers purchasing large amounts tends to increase the price-elasticity of the demand profile, but it is far from necessary.

3 This can be misleading in contexts where the price drops slowly until a steep decline at the end. Also, this property depends on there being a last unit of finite magnitude that is sold. Although this is the realistic case in practice, we later exhibit examples in which $q^* = \infty$ and the price remains above marginal cost for all units $q < \infty$; see the “bad examples” in Section 2 and Example 8.1. A convenient test of whether a model is well formulated is to check whether it implies exhaustion of demand at some finite quantity.
increases the decline in the profit contribution \( R \) from the \( q \)-th unit can be attributed entirely to shifts in demand and marginal cost, indicating that there is no further possibility for profit improvement from alteration of the price schedule. This condition can be applied to test the optimality of a price schedule using ordinary accounting data.

Examples

\begin{itemize}
  \item \textbf{Example 4.2} : Figure 4.1 depicts the maximization of the profit contribution for an example in which the demand profile is \( N(p, q) = 1 - p - q \) and the marginal cost is \( c = 0 \). For each of several values of the \( q \)-th unit, the profit contribution is maximized at the marginal price \( p(q) \) shown in the figure.\(^4\) The feature that the optimal marginal price is larger for smaller values of \( q \) is typical, but as mentioned above it is not universal.
  \item \textbf{Example 4.3} : A more elaborate example is derived from an explicit model of customers’ demand functions of the kind studied in §6. Each potential customer is described by two parameters \( t \) and \( s \) such that his net benefit is \( U(q, x; t, s) = P(q) - p^* x \) if he spends \( P(q) \) to purchase \( q \) units of the firm’s product and spends \( p^* x \) on a composite \( x \) representing all other commodities. The benefit function \( U \) is quadratic of the form

\[
U(q, x; t, s) = qt + xs - \frac{1}{2}[q^2 + 2ax + x^2],
\]

where the parameter \( a \) measures the degree to which the firm’s product is a substitute for other commodities. The customer’s demand for the firm’s product is therefore

\[
D(p, p^*; t, s) = [(t - p) - a[s - p^*])]/[1 - a^2],
\]

assuming that no nonnegativity restriction is imposed on the composite aggregate \( x \).

Consequently, the customer of type \( (t, s) \) purchases at least \( q \) units if \( t - a s \geq q[1 - a^2] + p - ap^* \). Suppose that among customers in the population, \( t \) and \( s \) have a bivariate Normal distribution with means 1, standard deviations 1, and correlation \( r \). Then the demand profile, measured as a fraction of the population, is the probability that a standard Normal random variable \( \xi \) with mean 0 and standard deviation 1 exceeds the value

\[
z(p, q) \equiv (q[1 - a^2] + [p - 1] - a[p^* - 1])/\sqrt{1 - 2ra + a^2}.
\]

\(^4\) This curve is a step function if a multipart tariff is used; that is, a single marginal price applies over each of several intervals of \( q \).
4. TARIFF DESIGN

Figure 1: Maximization of the profit contribution from the $q$-th unit via optimal selection of the marginal price $p(q)$. Assumes $N(p, q) = 1 - p - q$ and $c = 0$.

That is, $N(p, q) = \Pr \{ \xi \geq z(p, q) \}$. For this model of customers’ demand behaviors, Figure 4.2 shows the firm’s optimal marginal price schedule $p(q)$ for several values of the parameters. Generally, increasing $p^*$ raises the price schedule and increasing $r$ lowers it. The main effect of increasing the substitution parameter $a$ is to lower the price schedule if $p^* < 1$ and to raise it if $p^* > 1$. Observe that the last-listed price schedule has an increasing segment around $q = 4$: this is due to price competition with other commodities. As mentioned, this anomaly is addressed in Section 2.

4.2. Extensions and Qualifications

There are several extensions and qualifications to be considered in using the construction
variable marginal cost. If marginal cost varies with the quantity supplied to the customer then the proper marginal cost to use in the calculation of the marginal price for the $q$-th unit is the marginal cost $c(q)$ of producing and delivering this unit. Alternatively, if the firm’s total cost varies with the aggregate of the quantities supplied to all customers, then the marginal cost to be used is the marginal cost anticipated for the last unit of the aggregate supply to be produced. More generally, it may be that the firm’s total cost $C(Q)$ of supplying a list $Q = (q_1, \ldots, q_n)$ of quantities $q_i$ to customers $i = 1, \ldots, n$ is a function of both the
aggregate supply \( Q = \sum_i q_i \) and the individual quantities:

\[
C(Q) = C_1(Q) + \sum_i C_2(q_i)
\]

In this case the relevant marginal cost for an individual purchase \( q \) is

\[
c(q) = C_1^r(Q) + C_2^r(q),
\]

where \( C_1^r(Q) \) is the marginal cost of the aggregate supply.\(^5\) To apply this construction, one must proceed iteratively until the anticipated marginal cost of the aggregate used to construct the tariff agrees with the marginal cost predicted from the aggregate of the purchases induced by the tariff design.

If marginal cost varies over time due to changing demand and supply conditions, then its average is the relevant measure. If there are capacity limitations then a further premium is added to reflect the rationing of scarce supplies: the premium measures the benefits foregone by other customers who might have been served instead.\(^6\)

An alternative formulation supposes that the firm has a fixed supply or capacity \( Q \) available and therefore aggregate demand is restricted by the feasibility condition

\[
\int_0^\infty N(p(q), q) \, dq \leq Q.
\]

In this case the marginal cost \( c \) is augmented by a nonnegative Lagrange multiplier \( \gamma \) chosen sufficiently large to keep aggregate demand within the limit of aggregate supply.

\[\Box \quad \text{Example 4.4} \quad \text{: Suppose the demand profile is } N(p, q) = 1 - pq^a \text{ where } 0 < a \leq 1. \text{ Then the optimal price schedule is } p(q) = \frac{1}{2}[c + \gamma + q^{-a}] \text{ when the Lagrange multiplier is } \gamma, \text{ and the resulting demand for the } q \text{-th unit is } N(p(q), q) = \frac{1}{2}[1 - (c + \gamma)q^a]. \text{ Consequently, the optimal multiplier that keeps demand within a supply limit } Q \text{ is }
\]

\[
\gamma = \max \left\{ 0, \left( \frac{a}{2 \left[ 1 + a \right] Q} \right)^{1/a} - c \right\}.
\]

\(^5\) This is true even when each individual’s purchase has an imperceptible effect on the aggregate, as in models where demand is described by a smooth distribution function over purchase sizes and \( Q \) represents the average purchase.

\(^6\) Average marginal cost is relevant here because we are considering only a single product. If service is differentiated between peak and offpeak periods then marginal costs are assigned separately to each product, as in later chapters on multiple products.
Note that the multiplier is zero if the marginal cost is already large enough to restrain demand sufficiently.

This reflects the general principle that supply constraints are enforced by inflating marginal cost to include an imputed price of scarce capacity.

**Direct Derivation of the Necessary Condition** *

As a technical aside for readers interested in mathematical aspects, we note that the necessary condition for the optimal price schedule can be derived directly by maximizing the total profit contribution from the tariff $P(q)$. Assuming total costs are simply the sum of the costs incurred for each customer, the profit contribution is the producer’s surplus

$$PS \equiv \int_0^\infty [P(q) - C(q)] \cdot \nu(q) \, dq,$$

where the density of customers purchasing the quantity $q$ is $\nu(q) \equiv -dN/dq$ evaluated at $(p(q), q)$. This density takes account of the density $-\partial N/\partial q$ of customers purchasing $q$ at the uniform price $p = p(q)$, and also the shift in demand due to the slope $p'(q)$ of the marginal price schedule. In this form the problem involves the calculus of variations and the relevant optimality condition is the Euler condition, which here takes the form:

$$\frac{d}{dq} \left\{ N + \frac{\partial N}{\partial p} \cdot [p(q) - c(q)] \right\} = 0.$$ 

This condition appears weaker than the one derived previously, but the transversality conditions from the calculus of variations require also that the quantity in curly brackets is zero at the largest quantity $q^*$ purchased, and therefore it is zero for all $q \leq q^*$. This is the form of the necessary condition for an optimal price schedule derived directly in Section 1.

A similar conclusion is derived by first reformulating the profit contribution using integration by parts:

$$PS \equiv \int_0^\infty [P(q) - C(q)] \cdot \nu(q) \, dq = \int_0^\infty N(p(q), q) \cdot [p(q) - c(q)] \, dq,$$

apart from any fixed fee and fixed cost. Using the reformulated version on the far right, the optimal marginal price $p(q)$ for the $q$-th unit is selected pointwise to maximize the profit contribution from all purchases of that unit. This justifies the construction in Section 1.
§4. TARIFF DESIGN

Sufficiency of the Necessary Condition

The preceding analyses characterize the price schedule in terms of the first-order necessary condition for an optimum. In general, this condition can have multiple solutions. Extraneous solutions, such as local minima, can be excluded by imposing also the second-order necessary condition. A useful form of this condition is that the price elasticity \( \eta(p(q), q) \) of the demand profile is an increasing function of the price at the solution selected. More generally, the first-order necessary condition provides a unique solution for the optimal price schedule if the price elasticity of the demand profile is everywhere an increasing function of the price. This property is often satisfied in practice because it reflects the realistic fact that at higher prices customers have more opportunities to substitute competing products and services: the greater sensitivity to price is represented by higher price elasticities at higher prices.

Decreasing Price Schedules and the Ironing Procedure

Many applications require that the tariff is a concave function of the purchase size in order to preclude arbitrage by customers. This is equivalent to the requirement that the price schedule is nonincreasing — allowing quantity discounts but never imposing quantity premia. This requirement may affect the optimal price schedule if the marginal cost is a strongly increasing function of the purchase size, or the demand profile becomes insensitive to the price as the quantity increases. In particular, the price schedule is increasing at \( q \) if \( \partial^2 N / \partial q \partial p \) is too large at \( (p(q), q) \). Many models used in practice assure that the price schedule is decreasing because the price elasticity of the demand profile increases at a rate that is larger for larger values of \( q \). In other cases, however, the schedule of marginal prices can increase over some intervals of the purchase size, or even have folds or gaps. In such cases, the price schedule can be modified according to the following procedure.

Suppose that the solution of the necessary condition for optimality,

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

results in a price schedule \( p(q) \) that is increasing over some interval \( a < q < b \). Two such situations are depicted in Figure 4.3. In the left panel the price schedule is well defined but not monotone, whereas in the right panel the optimality condition allows multiple solutions at some values of \( q \) where the price schedule folds backward.\(^7\) In

\(^7\) There can be further complications if the necessary condition allows solutions in different connected components; cf. §8.1.
such situations, the optimal constrained price schedule must be constant, say \( p(q) = p \), over a wider interval \( A \leq q \leq B \) such that \( A \leq a < b \leq B \), \( p(A) = p \), and \( p(B) = p \).

Over this interval, the optimality condition must be satisfied on average; that is,

\[
\int_A^B \left\{ N(p, q) + \frac{\partial N}{\partial p}(p, q) \cdot [p - c(q)] \right\} \, dq = 0.
\]

Observe that these conditions provide three equations to determine the three values \( A \), \( B \), and \( p \) that need to be specified. If the initial price schedule is increasing over several intervals then it may be necessary to repeat this procedure several times in order ultimately to obtain a nonincreasing price schedule. That is, first one applies the procedure separately to each shortest segment where the schedule is increasing; if the new schedule still has increasing segments, then the process is repeated; and so on until no increasing segment remains. Flattening the price schedule this way is called “ironing.”

Difficulties of a related kind are depicted in Figure 4.4. In the left panel the price schedule has a segment that declines more steeply than customers’ demand schedules and therefore intersects their demand schedules from above; the right panel illustrates that this can also occur where the price schedule folds downward. In both cases the price schedule has a gap because customers will not select purchase sizes wherever the price schedule declines so steeply. In §8 we describe a variant of the ironing procedure that can be used determine the endpoints of the gap.

A specific example of a price schedule with an increasing segment is shown in Figure 4.5, which displays the “optimal” price schedule for Example 4.3 in the case that the parameters are \( a = 0.6 \), \( r = 0.6 \), and \( p^* = 0.5 \). In this case the products are close substitutes, the customers’ valuations of the products are highly correlated, and the competing product has a relatively low price. The horizontal dashed line shows the optimal flattened segment of the price schedule obtained by applying the ironing procedure. This segment occurs at a price slightly less than the price \( p^* \) of the competing commodity, as one expects.

An example of a price schedule with a steeply declining segment is presented in §12.1, Figure 12.3.

**Predictive Power of the Demand Profile**

The exposition above assumes that the demand profile is an adequate predictor of customers’ purchase behavior in response to a nonlinear price schedule. As we have noted, when arbitrage by customers is possible, an increasing segment of the price schedule may
Figure 3: Construction of a nonincreasing price schedule via the ironing procedure. The horizontal segment is selected so that the optimality condition is satisfied on average.
4.2. Extensions and Qualifications

Figure 5: An instance of Example 4.3 in which the ironing procedure is applied to ensure that the price schedule is nonincreasing.

need to be eliminated to preserve the predictive power of the demand profile. More generally, whenever the price schedule intersects a customer’s demand function at more than one point, or intersects at some point from above rather than below, then the predictive power of the demand profile fails and a more complicated analysis is required. To exclude these problematic cases, theoretical analyses of nonlinear pricing use assumptions sufficient to assure that the optimal price schedule intersects each customer’s demand function once, from below. These same assumptions also justify the use of the demand profile to summarize customers’ demand behavior. The full analysis of more general formulations is deferred to §6 and §8, where we present a formulation capable of handling cases with more complex features.

One must also check that a specified demand profile is realistic. The following are
4. TARIFF DESIGN

Typical “bad” examples.

1. If the demand profile is \( N(p, q) = 1 - k p^a q^b \) and the marginal cost is zero, then a mechanical application of the optimality condition indicates that the optimal marginal price schedule is \( p(q) = \beta / q^{b/a} \), where \( \beta = 1/k[1 + a] \). The predicted result, however, is that the same fraction \( 1 - k \beta^a \) of the customers purchase every increment.

2. A demand profile with some similar properties is \( N(p, q) = 1 - p/[1 - q] \), provided \( p + q < 1 \). If marginal cost is zero, then the optimal price schedule is \( p(q) = \frac{1}{2} \max \{0, 1 - q\} \). However, the predicted outcome is again that the same fraction \( 1/2 \) of the customers buy all increments \( q \leq 1 \); thus, the actual purchase size is 1 for all of these customers, and each pays the total charge \( P(1) = 1/4 \). This is another instance of bunching of customers at a single purchase size, but in this case the purchase is finite. If the marginal cost \( c \) is positive then there is no bunching and the price schedule is \( p(q) = \frac{1}{2} \max \{c, 1 + c - q\} \).

These examples indicate that it is a good practice to check that the overall implications of a demand profile specification are realistic. Sufficient conditions to preclude such pathologies are presented in §8.

4.3. The Bundling Interpretation

An alternative view construes nonlinear pricing as an instance of bundling. Products are said to be bundled if the charge for a purchase of several products in combination is less than the sum of the charges for the components. Bundling applies to products that are diverse (such as the options on a new automobile), but if the “products” are units of the same generic commodity then the effect is the same as nonlinear pricing. That is, for a bundle of two units a customer is charged less than twice the charge for a single unit.

The bundling approach is especially useful in understanding why nonlinear pricing can be applied in situations beyond the strictures of parameterized models. Figure 4.6 depicts the two-dimensional space of customers’ valuations \((v_1, v_2)\) for a first unit and a second unit, along with a scattering of points representing the distribution of these valuations in the population. For a pair of marginal prices \((p_1, p_2)\) for first and second units, the solid lines in the figure separate the space into three regions identifying the customers who purchase 0, 1, or 2 units. Calculation of approximately optimal prices follows the methodology in Section 1; that is, the demand profile is calculated from the distribution of points in the figure. The important observation is that the points can
be scattered arbitrarily; also, random fluctuations in their locations are immaterial as long as the overall distribution is stable. In contrast, a parameterized model in which differences among customers are described by $m$ type parameters implies that the points lie on an $m$-dimensional locus.

The figure also indicates why the calculation is only approximate. The demand profile approach can make two errors: customers whose valuations lie in the triangle $A$ are believed to purchase a second unit but not a first, and in fact they buy neither; and customers whose valuations lie in the triangle $B$ are similar but they buy both units. These errors are usually small and not serious if the size $\delta$ of a unit is small relative to customers’ purchases, for then the triangles $A$ and $B$ are also small.\footnote{Nevertheless, whenever the calculation of a price schedule indicates a precipitous}

Figure 6: Bundling analysis of nonlinear pricing.
Note: $\circ$ indicates valuations of one type of customer.
assume that units are infinitesimal and that the tariff increases smoothly as the purchase size increases. The motivation for this assumption is that analysis based on marginal calculations is much simpler, and for practical applications the errors are usually small; in addition, the basic concepts of nonlinear pricing are revealed more clearly.

There is a further lesson to be learned from the figure. Customers in the triangles $A$ and $B$ are ones for whom the price schedule is above their demand functions at a quantity of one unit and below at a quantity of two units. Thus, at two units the price schedule does not intersect their demand functions from below. In §6 we require explicitly that the price schedule does intersect customers’ demand functions once and from below. This requirement is satisfied automatically for many parameterized models. Generally it need not be satisfied, however, and when it is violated significantly the analysis must be based on detailed consideration of how customers select their purchase sizes. In the figure, for instance, the $45^\circ$ boundary line between the two triangles $A$ and $B$ indicates customers’ optimal choices of whether to purchase two units or none.

To make this explicit we show in Figure 4.7 how the demand profile is constructed from customers’ responses to a uniform price $p$. The count of valuations in the region where a single unit is purchased is $n(p, 1) = 14$, and where two units are purchased, $n(p, 2) = 10$. Recall that $N(p, 1)$ is the number of customers buying one or more units at this price, and $N(p, 2)$ is the number buying at least two units; so, $N(p, 1) = 24$ and $N(p, 2) = 10$.

### 4.4. Fixed Costs and Fixed Fees

In important cases the firm incurs a fixed cost for each customer served. The most common source of such a cost is capacity reserved for an individual customer. This cost is usually recovered by an auxiliary hookup or access fee, or by a minimum purchase requirement. The “demand charges” used in the electric power industry are fixed in relation to energy usage but they are based on the customer’s maximum power demand. Fixed costs of installation (for example, connection equipment) and administration (for example, metering and billing) for individual customers are also significant in some applications. These costs are often recovered via a fixed fee appended to the tariff. Their role in nonlinear pricing is usually minor.

---

*drop in price between successive units, it is well to modify the schedule to take account of the numbers of customers whose valuations are in the triangles $A$ and $B$ – as we elaborate in Section 4.*
4.4. Fixed Costs and Fixed Fees

$\$/unit 1

$p$

$\$/unit 2

Purchase Size

$n(p,1)$

$n(p,2)$

2 units

1 unit

$p$

Figure 7: Construction of the demand profile from the distribution of customers’ valuations. Note: $n(p, 1) = 14$ and $n(p, 2) = 10$ customers.

An important property of an optimal nonlinear tariff is that it imposes no fixed fee if the firm incurs no fixed cost in serving a customer. We sketch briefly why this is so. Envision the fixed fee $P_0$ and the minimal purchase $q_*$ as related by the condition that $P_0$ is the amount that makes a customer purchasing $q_*$ at the uniform price $p_* = p(q_*)$ indifferent whether to purchase at all. The minimal charge can be specified as $P(q_*) = P_0 + p_* q_*$. Moreover, the fixed fee can be interpreted as

$$P_0 = \int_0^{q_*} [\beta(q, q_*) - p_*] dq .$$

where $\beta(q, q_*)$ is a surcharge such that the same (number of) customers purchase each unit $q \leq q_*$, namely $N(\beta(q, q_*) , q) = N(p_*, q_*)$. The optimal choice of $q_*$ maximizes
the total profit contribution

\[ N(p_s, q_s) \cdot [P(q_s) - C(q_s)] + \int_{q_s}^{\infty} N(p(q), q) \cdot [p(q) - c(q)] \, dq, \]

which includes both the minimal charges collected and the marginal charges for purchases exceeding \( q_s \). The necessary condition for an optimal choice of \( q_s \) derived from this criterion requires that

\[ P(q_s) = C(q_s) + \left[ q_s / K(q_s) \right] \cdot \int_0^{q_s} \frac{\partial \hat{p}}{\partial q_s}(q, q_s) \, dq, \]

where \( K(q) = -[q/N] \frac{d}{dq} N \) is the total elasticity of the demand for the \( q \)-th increment. If the fixed cost is nil, namely \( C(0) = 0 \), then the solution of this condition is \( q_s = 0 \), indicating that the fixed fee is also nil. Similarly, if the fixed cost is small then so too is the fixed fee.

The essential lesson is that nonlinear pricing emphasizes primarily the design of the schedule of marginal prices. If fixed costs are nil then fixed fees are not used because they would restrict the market penetration without providing benefits to other customers. This is not exactly true of multipart tariffs, which do rely on fixed fees, but only because of the limitations imposed on replacing fixed fees with marginal prices. It is important to recognize, nevertheless, that this conclusion depends on an assumption that a customer bypasses service altogether if the net benefit is negative. If customers subscribe in any case, or at least a fixed fee does not decrease market penetration, then a fixed fee is generally optimal.

If the firm incurs a fixed cost serving each customer then a positive fixed fee may be necessary to recover full costs. In this case the fixed fee may exceed the fixed cost; moreover, the tariff charged for the least purchase size generally exceeds the firm’s cost. This case is presented as part of the general analysis in \$6 and \$8.

4.5. Multipart Tariffs

Characterizing the optimal price schedule in terms of the price elasticity of the demand profile depends on the possibility of varying the marginal price for each increment of the customer’s purchase. In practice, however, a different price for each increment is differentiation too fine to justify the transaction costs incurred by customers and the firm. As illustrated in \$2 for *Time, Newsweek*, and EDF, most firms’ tariffs specify prices
that are constant over wide ranges. Tariffs that are piecewise-linear in this way are called multipart tariffs.

Multipart tariffs take many forms described in §1. The simplest is a two-part tariff comprising a fixed fee plus a uniform price for every unit purchased. An $n$-part tariff is usually presented as a fixed fee plus $n - 1$ different “block declining” marginal prices that apply in different intervals or volume bands. In the usual case that the successive marginal prices decrease, the same net effect is obtained by offering $n - 1$ two-part tariffs from which each customer can choose depending on the purchase anticipated. A set of optional two-part tariffs provides a menu with the same basic consequences for customers — although a customer uncertain about usage risks selecting a two-part tariff that will be more expensive than the corresponding $n$-part tariff. An important instance is the pricing of product lines of services or leased machines adapted to different rates of usage by the customer. For example, a line of small, medium, and large machines, say copiers, offered at successively higher monthly rentals and successively smaller charges per copy, constitutes a menu of three two-part tariffs or a single four-part tariff. In the case of machines, considerations of cost and design may effectively limit the tariff to a few parts. More generally, however, the profit and total surplus foregone by using an optimal $n$-part tariff rather than an optimal tariff with continuously varying prices is of order $1/n^2$ for large values of $n$; thus, cost considerations need not be large for a tariff with only a few parts to be optimal.

The construction of a multipart tariff follows a procedure similar to the one described above. The main distinction is that the demand profile must be formulated to take account of the wider range of increments over which each marginal price applies. In addition, there is a practical proviso explained below.

Suppose that the $i$-th marginal price $p_i$ is to apply to each $q$-th unit in the range $q_i \leq q \leq r_i$. We indicate this volume band by the notation $[q_i, r_i]$. An appropriate specification of the demand profile in terms of volume bands is

$$
\bar{\mathcal{N}}(p, [q_i, r_i]) = \bar{\mathcal{N}}(p, i) = \sum_{q_i \leq q \leq r_i} \mathcal{N}(p, q)\delta ,
$$

or if $q$ varies continuously then

$$
\bar{\mathcal{N}}(p, i) = \int_{q_i}^{r_i} \mathcal{N}(p, q) \, dq .
$$

This is the number of customers purchasing increments in the range between $q_i$ and $r_i$ at the uniform price $p$. From each of these increments the firm obtains the profit
contribution \( [p - c] \delta \) if it charges the marginal price \( p \) and incurs the marginal cost \( c \) per unit. As a first approximation, the marginal price \( p_i \) to be charged for increments in this interval maximizes the profit contribution

\[
\mathcal{N}(p_i, i) \cdot [p_i - c].
\]

The following example illustrates this calculation.

\textbf{Example 4.5:} We adapt Example 4.1 to the case that the same price \( p_1 \) must apply to the first and second units purchased, and the second price \( p_2 \) applies to all additional units. Table 4.2 shows the calculation of the price schedule, which specifies marginal prices \( p_1 = \$4 \) and \( p_2 = \$3 \), again assuming that \( c = \$1 \).

\begin{table}[h]
\centering
\caption{Demand Profile for Example 4.5}
\begin{tabular}{lccc}
\hline
\textit{c = 1} & \textit{\( \mathcal{N}(p, i) \)} & & \\
\hline
\textit{p} & \textit{q = [0, 2]} & \textit{[3, 5]} & \textit{Total} \\
\hline
$2/\text{unit}$ & 165 & 90 & \\
$3$ & 145 & 65 & \\
$4$ & 115 & 35 & \\
$5$ & 75 & 10 & \\
\hline
\end{tabular}
\end{table}

The proviso mentioned above, and the reason that the price schedule derived for the example is only approximately optimal, can be seen as follows. In the example, if the price were \$5 for the first volume band and \$2 for the second then it would appear that fewer customers (75) demand increments in the first band than in the second (90). This is because the large price drop between the second and third units induces some extra customers to purchase a second unit in order to get the low price on the third; and others who would have bought a third unit at \$2 will not if they must pay \$5 for a second unit. Thus, the actual number who would purchase second and third units is between 75 and 90, but this number cannot be determined solely from the demand profile. The same error may occur for the price schedule calculated in the example. In
general, explicit account must be taken of customers’ motives to purchase a bundle of increments to obtain price breaks on the last increment(s) in the bundle, as explained in the description of bundling in Section 3. For now we ignore this consideration but “optimal” will remain in quotes to acknowledge that incomplete account is taken of customers’ demand behaviors at boundaries between volume bands.

An n-part tariff includes specification also of the intervals over which the marginal prices are to apply, as well as the fixed fee if any. The general construction therefore requires that the sum of the profit contributions over all the intervals, plus the sum of the fixed fees received, is maximized. We describe an approximate formulation that ignores customers’ behaviors at boundaries between segments but allows simple computations. An exact formulation is presented in §6.4.

The Approximate Formulation

For this formulation we assume that the number of customers purchasing a q-th unit in the interval \([q_i, r_i]\) at the uniform price \(p_i\) is the same when a multipart tariff is offered. As mentioned, this ignores demand behaviors at the boundaries: for \(q\) near \(q_i\) some customers are included who should not be, and for \(q\) near \(r_i\) some customers are excluded who should not be. Also, we take \(r_i = q_i + 1\) so that there are no gaps in the price schedule. These errors are typically small if the tariff has many segments.

For this formulation, the profit contribution is

\[
N_s(P(q_1), q_1) \cdot [P(q_1) - C(q_1)] + \sum_{i=1}^{n-1} \hat{N}(p_i, [q_i, q_{i+1}]) \cdot [p_i - c].
\]

The first term represents one way to specify a fixed fee: the minimum purchase size is \(q_1\) for which the customer pays \(P(q_1)\) and \(N_s(P, q)\) indicates the predicted number of customers willing to purchase this minimal quantity, namely the number for whom the gross benefit from \(q\) units exceeds the tariff \(P\). One can construe \(P_0 = P(q_1) - p_0 q_1\) as the imputed fixed fee if, as is frequently the case, this fee still requires the customer to pay for an initial purchase of \(q_1\) units at the price \(P_0\). If the tariff is implemented as a menu of two-part tariffs with fixed fees \(P_i\) and marginal prices \(p_i\), then the i-th fixed fee can be calculated from the previous one via the relationships

\[
P_{i-1} + p_{i-1} q_i = P(q_i) = P_i + p_i q_i,
\]

\[9\] The source of this difficulty is that the demand profile is too aggregated to allow inferences about customers’ behaviors at boundaries between segments of multipart tariffs. The disaggregated models studied in Part II allow exact characterizations, as in §6.4.

107
which yield the recursive formula
\[ P_i = P_{i-1} + [p_{i-1} - p_i] \cdot q_i. \]

That is, the increment in the fixed fee accounts for the lower marginal prices payable for all units \( q \leq q_i \). Such a menu assures that the \( i \)-th two-part tariff minimizes the amount paid for a quantity in the interval \( q_i \leq q \leq q_{i+1} \).

This profit contribution is to be maximized by choosing both the prices \( p_i \) and the breakpoints \( q_i \) for each interval \( i = 1, \ldots, n - 1 \). One uses \( q_n = \infty \) so that no upper bound is placed on the size of a purchase. Usually the prices must be constrained to be nonincreasing to avoid arbitrage by customers, as described in §1.3, and of course the breakpoints must be increasing.

\textbullet \textbf{Example 4.6} : For this example we assume that the demand profile is \( N(p, q) = 1 - p - q \). As mentioned, for the computations here we ignore customers’ behaviors at the boundaries between intervals and simply use the demand profile for volume bands directly. With this proviso,

\[ \tilde{N}(p_i, [q_i, q_{i+1}]) = \begin{cases} 
[q_{i+1} - q_i][1 - p_i - q_i] & \text{if } i < n - 1, \\
\frac{1}{2}[1 - p_i - q_i]^2 & \text{if } i = n - 1,
\end{cases} \]

where \( q_i = \frac{1}{2}[q_i + q_{i+1}] \) is the midpoint of the volume band. Consequently, an “optimal” \( n \)-part tariff uses the marginal prices

\[ p_i = \begin{cases} 
\frac{1}{2}[1 + c - q_i] & \text{if } i < n - 1, \\
\frac{2}{3}c + \frac{1}{3}[1 - q_i] & \text{if } i = n - 1,
\end{cases} \]

where for \( i > 1 \) the breakpoints between the volume bands satisfy \( q_i = 1 + c - [p_{i-1} + p_i] \). Given \( q_1 \) these conditions provide a system of simultaneous equations that determine all the volume bands and the price charged for each one.

To give the flavor of these results, we suppose that the last price \( p_{n-1} \) satisfies the same formula as the others; also, we use \( q_1 = [1 - c]/n \), which turns out to be correct when there is no fixed cost, fixed fee, nor minimal purchase. In this case, the breakpoints are \( q_i = [1 - c]/n \) for \( i = 1, \ldots, n \) and the price for the \( i \)-th band is \( p_i = \frac{1}{2}[1 + c - (1 - c)(i + .5)/n] \). If a menu of two-part tariffs is used, then the fixed
fee associated with the \( i \)-th price \( p_i \) is \( P_i = \frac{1}{4}[1 - c]^2 i[i + 1]/n^2 \). Figure 4.8 shows the resulting multipart tariff for \( n = 3 \) and \( n = 5 \) when the marginal cost is \( c = 0 \). Note that the breakpoints are evenly spaced and the marginal prices decline steadily between each pair of adjacent intervals, whereas the associated fixed fees increase quadratically. In terms of sales, the same number of customers selects purchase sizes in each volume band, but the number of units sold per interval declines steadily and therefore, since the marginal prices decline too, revenues from these units decline quadratically. As the number \( n \) of intervals increases, the firm’s profit increases to the maximum profit \( \frac{1}{12}[1 - c]^3 \) obtained from the nonlinear tariff \( P(q) = \frac{1}{2}[1 + c]q - \frac{1}{4}q^2 \).

Example 6.3 recomputes these results taking explicit account of customers’ demand behaviors at the boundaries between intervals, and obtains an exact answer that is only slightly different if \( n \) is large. In particular, if \( c = 0 \) then the \( i \)-th truly optimal marginal
price is \( p_i^* = \frac{n}{n - 0.75} p_i \), where \( p_i \) is the value calculated in the example above; in addition, the boundaries between segments differ slightly. This relationship is indicative of the degree to which the “optimal” price schedule approximates the truly optimal one. Generally, customers’ behaviors at boundaries between segments of multipart tariffs are immaterial if there are many segments so that the marginal prices do not differ greatly between adjacent segments.

### 4.6. Summary

This chapter construes nonlinear pricing as a particular kind of product differentiation. The product line comprises the various incremental units, each offered at its own price, plus possibly a minimal purchase. From this perspective, the construction of an optimal price schedule reduces to the calculation of the optimal price for the minimal purchase, if any, and for each increment in the purchase size. This viewpoint is also illustrated by construing nonlinear pricing as a special case of bundling.\(^\text{10}\)

The demand for increments is measured by the demand profile \( N(p, q) \), which for each price \( p \) specifies the demand for the \( q \)-th increment. The demand profile can be estimated from data obtained from uniform prices, since for each price it represents the right-cumulative distribution function of purchase sizes among the customers in the population. Having estimated the demand profile, a firm can calculate an optimal price schedule \( p(q) \) by maximizing the profit contribution from sales of \( q \)-th units. This amounts to a simple arithmetical task when the demand profile is given in tabular form, as in Table 4.1, or as in Table 4.2 in the case of a block declining price schedule. Variable marginal costs per customer and an aggregate marginal cost component can be included. Fixed costs can be recovered via a fixed fee, but the optimal fixed fee is nil if fixed costs are nil.

The analysis assumes that the price schedule is nonincreasing, but if not then it can be flattened appropriately via the ironing procedure so that the marginal contribution to consumers’ and producer’s surplus averages to zero over an interval of units for which the marginal price is uniform. Other technical amendments also apply; for instance, the marginal price schedule must intersect each customer’s demand function once, from below, if the demand profile is to be a fully accurate predictor of customers’ demand behaviors. Further, the analysis via bundling indicates that if increments are large then it

\(^{10}\) As shown in §13.5, multiproduct tariffs involve bundling of units of each product as well as bundling among products.
is necessary to consider also how customers choose between larger and smaller purchase sizes when intermediate purchase sizes are not optimal.

The ironing procedure addresses several complications that can occur in general models of nonlinear pricing described in §8. Nonmonotonicities, folds, steep segments, and multiple components of the price schedule are rare in models used in practice, but they cannot be excluded entirely without imposing unrealistically strong assumptions. Most of the exposition in subsequent chapters therefore focuses on regular cases in which these complications do not arise, but amendments to handle irregular cases are mentioned occasionally. Most amendments share the common feature that a segment of the price schedule is altered (to be flat, vertical, or a gap) over an interval (of units or types) for which the optimality condition is required only to be satisfied on average.
Chapter 5

RAMSEY PRICING

Many public enterprises and regulated privately-owned utilities are allowed revenues sufficient only to recover their total costs. Besides operating and administrative costs, these include costs of capital invested in capacity. For such firms, one procedure that produces an efficient tariff design is Ramsey pricing. The construction in §4 of a price schedule for a profit-maximizing monopolist is an application of tariff design based on Ramsey pricing; monopoly pricing is the special case in which the firm’s capital and administrative costs are so large that profit maximization is required. More generally, Ramsey pricing allows the firm sufficient use of monopoly power to meet its revenue requirement.

The guiding principle of Ramsey pricing is to construct the tariff to maximize an aggregate of customers’ benefits, subject to the constraint that the firm’s revenues recover its total costs. The most common aggregate used in applications is the simple unweighted sum of the money values of customers’ net benefits, namely consumers’ surplus. Additional constraints are also included in some applications. The most important requires that no customer is worse off with Ramsey pricing than the uniform price schedule that provides the same net revenue for the firm. This constraint is imposed to ensure customers’ and regulators’ acceptance of Ramsey pricing as an improvement.

1 This reflects a narrow interpretation of monopoly. Often a firm with a monopoly position in the market is also subject to the threat of entry by other firms, and to forestall entry it must keep its prices low. For analyses of pricing to deter entry, see the survey by Wilson (1992b) and for studies of the special case of contestable markets see Baumol, Panzar, and Willig (1982) and Maskin and Tirole (1988).

2 This is the main ingredient of Ramsey pricing, named after Frank P. Ramsey (1903 - 1930), who first proposed and analyzed an application in a 1927 article. It is sometimes called Ramsey-Boiteux pricing to recognize its further development by Marcel Boiteux (1956). It is widely used by regulatory agencies; for example, in 1983 the United States Interstate Commerce Commission adopted Ramsey pricing as the basic principle for setting railroad rates. Baumol (1987) provides a brief exposition that includes a short history of the subject. For an analysis of Ramsey pricing in relation to statutory prohibitions against undue price discrimination in public utility rates, see Henderson and Burns (1989).
over an existing uniform price schedule. Its practical effect is usually to put a cap on the
prices charged for small purchases; typically it has no effect on the quantity discounts
offered for large purchases.

Section 1 introduces the Ramsey formulation of pricing by a regulated monopoly. In this
formulation, the tariff is designed to maximize consumers’ surplus, namely the
aggregate of customers’ net benefits, subject to the firm’s revenue requirement. The
main concepts of nonlinear pricing carry over to this formulation except that the price
elasticity of the demand profile is artificially inflated, thereby lowering the price schedule.
Subsequent sections elaborate the modifications required to ensure that no customer is
affected adversely by the firm’s adoption of a nonlinear tariff. Section 2 introduces the
concept of a Pareto-improving tariff and derives the conditions for an optimal tariff
constrained by this requirement. An application to the telephone industry is described
in Section 3.

5.1. The Price Schedule of a Regulated Firm

In this section we demonstrate that the net effect of Ramsey pricing is simply to reduce the
percentage profit margin on each unit sold until the utility’s revenue equals its total cost.
The key requirement is that this reduction should be the same fraction of the monopoly
percentage profit margin on every unit. This uniformity of the percentage profit margin is
called the Ramsey pricing rule. The percentage profit margin that is common to all units
is called the “Ramsey number” and is usually denoted by $\alpha$. For a profit-maximizing
monopoly, $\alpha = 1$, and for a regulated firm with no binding revenue requirement, $\alpha = 0$.
The Ramsey number is typically intermediate between these extremes.

Thus, if the construction of the optimal tariff for a profit-maximizing monopolist
produces a percentage profit margin $m(q) = [p(q) - c(q)]/p(q)$ on the $q$-th unit, then
a regulated utility would use a price schedule with the profit margin $\alpha m(q)$ for some
fixed fraction $\alpha$ that is independent of $q$ — were it true that customers’ demands
remained unchanged. In fact, however, customers respond to the lower prices offered
by a regulated utility. The actual rule is therefore that the utility’s price schedule should
satisfy the condition that

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

where again $\eta$ is the price elasticity of the demand profile, for a value of $\alpha$ that recovers
the utility’s total costs. This Ramsey pricing rule is often expressed by the condition that
the product of the percentage profit margin and the price elasticity of the demand profile
should be the same (that is, equal to the Ramsey number) for every increment. Thus, an
empirical test of the optimality of an existing or proposed price schedule is obtained by
tabulating and comparing these products to see if there are any large differences among
them.

We derive the Ramsey pricing rule by taking advantage of the dual interpretation
of the demand profile. The possibility that the firm charges each customer a fixed fee
or insists on a minimal purchase is omitted since the analysis parallels the treatment in
§4.4; a complete exposition is included in §6.

Construction of Ramsey Price Schedules
Recall from §3 that the demand profile has two interpretations. For a fixed uniform price
$p$, the demand profile $N(p, q)$ indicates the right-cumulative distribution of customers’
purchase sizes. Based on this interpretation, $N(p(q), q)$ customers buy a $q$-th unit at
the marginal price $p(q)$, and aggregate demand is

$$Q = \int_0^\infty N(p(q), q) \, dq.$$

From each one of the $q$-th units sold the firm’s profit contribution is the difference
$p(q) - c(q)$ between the price and the marginal cost, and the number of such units sold is
$N(p(q), q)$. The firm’s profit contribution from all units sold according to the price
schedule $p(q)$ is therefore

$$PS = \int_0^\infty N(p(q), q) \cdot [p(q) - c(q)] \, dq.$$

The firm’s profit contribution is also called the producer’s surplus to parallel the nomen-
clature for consumers’ surplus. Assuming the firm requires a net revenue $R_*$ to recover
its total costs, we impose the constraint that $PS \geq R_*$.

According to the second interpretation, for each fixed $q$-th unit the demand profile
$N(p, q)$ indicates the right-cumulative distribution of customers’ marginal valuations of
this unit. The aggregate of customers’ net benefits from purchasing the $q$-th unit at the
price $p(q)$ is therefore

$$CS(q) = \int_{p(q)}^\infty [p - p(q)] \, d[1 - N(p, q)] = \int_{p(q)}^\infty N(p, q) \, dp.$$

In this formula the first summation or integration uses the variable $p$ to parameterize
customers’ marginal valuations or reservation prices for a $q$-th unit and uses $1 - N(p, q)$
to represent the usual left-cumulative distribution of customers’ marginal valuations. The integral on the right results from integration by parts, using the property $N(p, \infty) = 0$. The aggregate over all incremental units is therefore

$$\text{CS} = \int_0^\infty \int_{p(q)}^\infty N(p, q) \, dp \, dq,$$

which is the total consumers’ surplus.

The method of Ramsey pricing selects the price schedule to maximize this aggregate measure of customers’ net benefits within the limit allowed by the revenue constraint. This principle can be interpreted as seeking an efficient allocation of purchases to customers subject to two provisos reflecting political considerations affecting regulatory policy. The first is dubbed the “every tub on its own bottom” constraint: the utility’s full costs are recovered from its customers, rather than subsidized by other means such as general tax revenues. The second proviso is that among the many efficient allocations, the one selected maximizes consumers’ surplus. The usual rationale for this selection is that each customer is affected by utilities in several industries and by an assortment of public programs, as well as taxes, subsidies, and welfare programs with substantial distributive effects. Absent a coherent scheme to coordinate all these programs to achieve distributive objectives, if each program maximizes total benefits then in aggregate the greatest potential benefit is available for redistribution via taxation and other programs with explicit welfare or distributional objectives.

For historical reasons, the optimization is usually stated equivalently as the maximization of the total surplus $\text{TS} \equiv \text{CS} + \text{PS}$ subject to the revenue constraint. A Lagrange multiplier $\lambda$ is used to include the revenue constraint explicitly in this objective. It is a nonnegative number chosen large enough to meet the revenue requirement, but zero if the requirement is exceeded. Including the Lagrangian term, the problem posed is to maximize

$$\mathcal{L}_* \equiv \text{TS} + \lambda [\text{PS} - R_*], \quad \text{or equivalently} \quad \mathcal{L} \equiv \text{CS} + [1 + \lambda]\text{PS},$$

where in the latter the dependence on the revenue requirement is summarized entirely by the multiplier $\lambda$. In the present context, the multiplier $\lambda$ measures the reduction in total surplus caused by a unit increase in the firm’s revenue requirement. This reduction is usually positive because the firm obtains a greater net revenue only by exploiting further its monopoly power, which reduces the consumers’ surplus that can be achieved.
In the extreme case of a profit-maximizing monopoly, only the firm’s profit matters, which corresponds to $\lambda = \infty$.

Writing out the Lagrangian augmented objective function in full, we obtain:

$$\mathcal{L} \equiv \int_0^\infty \left\{ \int_{p(q)}^\infty N(p, q) \, dp + [1 + \lambda] \cdot N(p(q), q) \cdot [p(q) - c(q)] \right\} \, dq.$$ 

Consequently, for each $q$-th unit the marginal price $p(q)$ maximizes the expression in curly brackets. This yields the necessary condition that characterizes an optimal price schedule:

$$-N(p(q), q) + [1 + \lambda] \left\{ N(p(q), q) + \frac{\partial N}{\partial p}(p(q), q) \cdot [p(q) - c(q)] \right\} = 0.$$ 

The interpretation of this condition is straightforward: the factor in curly brackets is familiar from §4 as the firm’s marginal profit contribution from the $q$-th increment, and the first term merely states that a $1$ price increase reduces by $1$ the consumer’s surplus from each of the $N(p(q), q)$ customers who purchase that increment. The price increase also induces $|\partial N/\partial p|$ customers to refrain from purchasing the $q$-th increment: this affects the firm’s profit contribution if $p(q) > c(q)$ but these customers’ surpluses are essentially unaffected since they were indifferent whether to add this last increment to their purchases.

Using the Ramsey number $\alpha \equiv \lambda/[1+\lambda]$, this optimality condition can be expressed equivalently as

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

which is the standard form we use hereafter.

The optimality condition can also be expressed in terms of the percentage profit margin,

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

where $\eta(p, q)$ is the price elasticity of the demand profile as in §4.1. This form of the optimality condition reveals that the essential effect of Ramsey pricing is to reduce the monopoly percentage profit margin, uniformly for all units, so that only the required revenue is obtained by the firm. This effect can be given several interpretations. The direct interpretation recognizes that the price for each unit includes an ad valorem or value-added tax to meet the firm’s revenue requirement. This tax is stated as a percentage markup inversely proportional to the price elasticity of the demand for that unit. Units
with lower price elasticities are taxed more because their demands are curtailed less by the tax. In particular, the tax imposes a welfare loss due to the resulting departure from the fully efficient demands that would result from marginal cost pricing, and this welfare loss (as measured in terms of consumers’ surplus) is roughly proportional to the price elasticity. The resulting pricing rule uses the firm’s monopoly power efficiently to meet the revenue requirement.

Two other interpretations are sometimes used. One is that regulation enforces behavior by the firm that is based on a price elasticity of the demand profile that is artificially inflated by a factor $1/\alpha$, as though price competition from other firms’ products that are imperfect substitutes were more severe than it actually is. A second interpretation developed in §12 is that regulation forces the firm to behave as if it were one of $1/\alpha$ firms offering products that are perfect substitutes. Either interpretation construes the net effect of regulation as the imposition of competition when in fact there is none.

This form of the optimality condition generalizes the condition derived in §4, since $\alpha = 1$ is the case of a profit-maximizing monopoly. It also includes the case of fully efficient marginal cost pricing: when the revenue constraint is not binding, $\lambda = 0$ and $\alpha = 0$ and therefore $p(q) = c(q)$. Thus, Ramsey pricing is a generalization of both monopoly pricing and marginal-cost pricing, and between these extremes it includes a spectrum of possibilities that correspond to the various revenue requirements the firm might have. Profit margins are higher if the Ramsey number is higher, in order to recover the firm’s greater fixed costs. Ironing procedure, for Ramsey price schedules

Ironing of the price schedule has a pronounced effect on Ramsey pricing when the firm’s marginal cost is increasing and its revenue requirement is small. The extreme case where $\alpha = 0$, for instance, corresponds to completely efficient pricing, as in a perfectly competitive market: each marginal price equals the corresponding marginal cost, and therefore is increasing. The price schedule must be constrained to be subadditive (or more stringently, nonincreasing), however, if a customer can substitute several small purchases for a large one. The net result of ironing, therefore, is to charge a single uniform price equal to the marginal cost of the last unit in the largest size purchased. This may entail charging the marginal cost for a single unit, relying on each customer to make multiple unit purchases to make up a larger purchase.

**Welfare Aspects of Ramsey Pricing**

Ramsey pricing has rarely been embraced by regulatory agencies without amendments.
In this subsection we mention two sources of their concern. We describe some efficiency considerations first and then introduce the distributional considerations that motivate the analysis in Section 2.

**Efficiency Aspects of Ramsey Pricing**

Although some use of monopoly power may be required if a regulated firm is to recover its full costs, the kinds of monopoly power that the firm is allowed to exercise is a matter of choice; moreover, monopoly power can be abused by inefficient use.

Nonlinear tariffs derived from the principles of Ramsey pricing suppose that the firm is allowed to charge different prices for different increments. If marginal cost is constant then this is a kind of price discrimination created solely by the design of the tariff, since typically increments sold to one customer are generically the same, and the same as those sold to other customers. Indeed, it may seem puzzling that the general principles of Ramsey pricing, which are based on a criterion of efficiency, lead inexorably to nonlinear pricing and therefore to differentiated prices for increments that are physically indistinguishable in production or consumption. As mentioned above, the answer to this puzzle is that use of its monopoly power to differentiate prices is the efficient way for the firm to meet its revenue requirement. In technical terms this answer is manifest in the fact that the aggregate price elasticity (of the demand profile) is different for different increments; thus, the aggregate welfare loss incurred in meeting the revenue requirement is minimized by pricing increments differently depending on their price elasticities. This conclusion follows inevitably from the restriction that the same tariff is offered to all customers, so only the aggregate price elasticities matter, and the objective of maximizing an aggregate of customers’ net benefits, such as consumers’ surplus.

Apprehension about the use of price discrimination, however, has deeper roots. Quite apart from distributional aspects, concerns about price discrimination stem from the possibility that it can promote productive and allocative inefficiencies, especially through quality degradation. A typical example is an airline that uses nonlinear pricing in the form of nonrefundable discounts for advance purchases: high prices for tickets sold shortly before departure can mean that some seats go unfilled, and customers may incur costs adhering to rigid itineraries. Similarly, a publisher that decreases its price over time incurs storage costs and delays benefits for customers. It is conceivable as well that the qualities of utility services such as power and communications could be degraded artificially to facilitate price discrimination. Regulatory agencies are therefore cautious in allowing tariffs that involve price discrimination, in part to ensure that they
are not sustained by quality degradation or other sources of productive or allocative inefficiencies.

The formulation of Ramsey pricing used here bypasses these considerations by assuming implicitly that the product quality and the production technology are fixed. It does not address the issue of whether productively inefficient uses of monopoly power should be allowed in order to meet the firm’s revenue requirement; that is, the formulation simply assumes that quality specifications and the production and distribution systems are operated efficiently.³

**Distributional Effects of Ramsey Pricing**

A second source of concern is that Ramsey pricing can reduce the net benefits of customers making small purchases, even though it provides compensating gains to others, when maximization of an aggregate of customers’ net benefits is used as the objective. To illustrate, we present two examples that demonstrate possible consequences of unfettered use of Ramsey pricing.

**Example 5.1**: For this example, suppose that the demand profile is \( N(p, q) = 1 - q/[1 - p] \), provided \( p + q \leq 1 \). Using the condition above, the optimal price schedule is found to be

\[
p(q) = 1 + \frac{q}{2} \frac{1 - \alpha}{\alpha} - \sqrt{\left( \frac{q}{2} \frac{1 - \alpha}{\alpha} \right)^2 + q \frac{1 - c}{\alpha}}.
\]

Figure 5.1 shows this price schedule for several values of the Ramsey number \( \alpha \), assuming the marginal cost is \( c = 0 \). Figure 5.2 shows how the resulting profit contribution increases as the Ramsey number increases. Notice that for an initial increment the price schedule charges \( p(0) = 1 \) for any Ramsey number that is positive. Even for a profit-maximizing monopoly, the uniform price is only half as large. Thus, the Ramsey price schedule charges more for initial units and less for the last units than a uniform price. Customers making small purchases therefore encounter tariffs that exceed the charges imposed by a uniform price.

**Example 5.2**: A more extreme example has the demand profile \( N(p, q) = 2 + \log(1 + p)/\log(q) \), provided \( q < 1 \). The optimal price schedules are depicted in Figure 5.3 for the same set of Ramsey numbers as the previous example, and the marginal cost \( c = 0 \). The marginal price for an initial increment is \( p(0) = \infty \), because the price elasticity of the

³ For analyses of these topics see Laffont and Tirole (1992) and Spulber (1989).
Figure 1: Example 5.1: The marginal price schedule for Ramsey numbers $\alpha = 0.2(0.2)1$. 

Profit Contribution ($/Pot.Customer)
5.1. The Price Schedule of a Regulated Firm

demand profile is zero for an initial increment. The dependence of the profit contribution on the Ramsey number is shown in Figure 5.4.

In both examples the prices for initial increments are essentially the monopoly prices for these increments, regardless of the Ramsey number. In the second example, moreover, the infinitely high prices are an indirect way of imposing a fixed fee.

For customers making small purchases, the Ramsey price schedule is disadvantageous compared to the uniform price that raises the same net revenue for the firm. In the first example, for instance, a uniform price $\bar{p}$ provides the net revenue $\tilde{D}(p)[\bar{p} - c]$, where the total demand is $\tilde{D}(p) = \frac{1}{2}[1 - \bar{p}]$. For each uniform price $\bar{p} < 1$ there is a Ramsey number $\alpha(\bar{p})$ such that the nonuniform Ramsey price schedule yields the same net revenue to the firm. If offered a choice between these two schedules, customers making small purchases prefer the revenue-equivalent uniform price, whereas customers making large purchases prefer the Ramsey price schedule, due to its substantial quantity discounts for large purchases.

A compromise that avoids this conflict among customers puts a cap on the marginal prices allowed. Choosing the price cap to be equal to the revenue-equivalent uniform price assures that no customer is disadvantaged by the adoption of nonlinear pricing. If this is done then the resulting schedule is

$$p^\circ(q) = \min \{ \bar{p}, p(q) \} = \begin{cases} \bar{p} & \text{if } q \leq \bar{q}, \\ p(q) & \text{if } q > \bar{q}. \end{cases}$$

This compromise obtains the efficiency advantages of quantity discounts for large customers without disadvantaging customers purchasing small amounts, but the Ramsey number is larger. The nonuniform part $p(q)$ is computed using a larger Ramsey number chosen so that the firm’s revenue requirement is still satisfied.

Practical implementations are often based on a menu of options among which each customer can choose. For instance, as in AT&T’s tariffs described in §2.4, one option could be the uniform price $\bar{p}$ applicable to all units, and a second option could be the nonuniform price schedule $p(q)$ applicable to units $q > \bar{q}$ after payment of a fixed fee $\bar{p}\bar{q}$ that provides free service for units $q \leq \bar{q}$. Customers purchasing more than $\bar{q}$ would then prefer the second option.

In Section 2 some further comments on the distributional effects of Ramsey pricing provide a cautionary reminder: when customers are actually intermediate producers, the welfare consequences of a nonlinear tariff depend on the extent that benefits from quantity discounts are passed on to retail consumers.
§5. RAMSEY PRICING

Figure 3: Example 5.2: The marginal price schedule for Ramsey numbers $\alpha = 0.2(0.2)1$.  

Profit Contribution ($/Pot.Customer)

$0  $5  $10  $15  $20  $25

$0  0.2  0.4  0.6  0.8  1

Ramsey Number

122

Ramsey Number

q-th Unit

Price $p(q)$ ($/unit$)

$0  $2  $4  $6  $8  $10

$0  0.2  0.4  0.6  0.8  1

q-th Unit
5.2. Pareto-Improving Tariffs

As described above, for a public utility the implementation of a nonlinear tariff is often constrained by the distributional consequences for customers. Changing from a uniform price to a multipart tariff or a fully nonlinear tariff can benefit some customers but affect others adversely. For instance, introducing a two-part tariff with a fixed fee plus a lower marginal price benefits customers with large demands but precludes others from making any purchases at all. Utilities and regulatory agencies therefore find it desirable to consider modifications that do not disadvantage any customer. Tariff amendments that have this property, and that also do not reduce the firm’s revenue, are said to be Pareto-improving.4

Examples of Pareto-Improving Tariffs

There are many ways that an existing nonoptimal tariff can be modified in a Pareto-improving way, of which we describe here only a few. Initially we assume that customers’ demands are independent, in the sense that the tariff offered to one customer does not affect another customer’s response; this assumption is reexamined later.

Introduction of a Two-Part Tariff

When the existing tariff is linear with a uniform price for all units purchased, a Pareto-improving change can be made by introducing an appropriately designed two-part tariff

---

4 This name derives from the economist Vilfredo Pareto (1848-1923) who studied the criterion of unanimous consent for changes in economic policy.
as an option available to customers. It is important to realize that electing the two-part tariff must be optional, not required. The new tariff can be construed as a menu with two options: each customer can select either the existing linear tariff, say $P_1(q) = P_1 + p_1q$ where $P_1 = 0$ and $p_1$ is the present uniform price, or the new two-part tariff, say $P_2(q) = P_2 + p_2q$ where $P_2$ is a positive fixed fee and the marginal price is $p_2 < p_1$. It is clear that a fully informed customer cannot be disadvantaged by this menu, since each customer retains the option to purchase under the old tariff, and indeed those customers intending to purchase amounts substantially less than 

$$q^* = \frac{P_2}{p_1 - p_2}$$

prefer to remain with the old tariff. It remains therefore only to ensure that the firm’s revenue is not reduced. A key fact is that there does exist such a choice with $P_2 > 0$ and $p_2 < p_1$ whenever $p_1$ exceeds the firm’s marginal cost. Together, the two optional tariffs provide a block-declining price schedule in which the second block offers the lower price $p_2$ for incremental units in excess of $q^*$. Because $p_1$ exceeds marginal cost, there is some choice of $q^*$ beyond which the firm’s revenue is increased by offering a lower marginal price $p_2$.

This result evidently extends also to the case that the existing pricing policy is a two-part tariff, namely $P_1 > 0$. Indeed, it can also be beneficial to introduce an option that has a lower fixed fee and a higher marginal price. In general, if the optimal tariff is concave, so that it is equivalent to a menu of optional two-part tariffs, then an existing menu of several two-part tariffs that is not fully optimal can be Pareto-improved by introducing an additional option.

Introduction of a Quantity Discount

An alternative approach introduces a quantity discount for the largest customers to stimulate sales. Whenever the existing tariff has a marginal price (for the last unit purchased by the largest customer) exceeding marginal cost, an appropriately designed quantity discount is Pareto-improving. To see this, it is simplest to consider the addition of a quantity discount that brings entirely new sales to the firm. Suppose that under the existing tariff the maximum purchase by any customer is $\bar{q}$, where by assumption the marginal price charged for this unit exceeds the marginal cost that the firm incurs to provide it. Suppose now that the firm revises the existing tariff by offering units after the $\bar{q}$-th at some price between marginal cost and what the existing tariff charges. With this revised tariff, each customer will purchase at least as much as before, and at the same

5 See Willig (1978) for a proof of this fact, which depends on the assumption of independent demands among customers.
prices, except that some additional units will be purchased at lower prices exceeding the firm’s marginal cost. Thus, no customer is disadvantaged and some benefit (at least the largest customer will be encouraged to purchase more), and the firm’s revenues are increased.

Recall that an optimal nonlinear tariff has the property that the last unit is sold at marginal cost; thus, further quantity discounts cannot Pareto-improve an optimal tariff. The fact that a nonoptimal tariff can be improved by quantity discounts for the largest customers to increase their purchases is a familiar one for many public utilities: often they perceive opportunities to increase revenues while benefiting their industrial customers by offering favorable terms for very large purchases — and ultimately the increased revenue can enable lower prices or better service for all customers.

Introduction of an Optimal Nonlinear Tariff Segment

Another tactic allows customers the option of purchasing units according to the price schedule for an optimal nonlinear tariff. Suppose that \( P(q) \) and \( p(q) \) are the current tariff and its price schedule, which may reflect uniform pricing, a two-part tariff, or any more complicated menu of options; and let \( P^*(q) \) and \( p^*(q) \) represent an optimal nonlinear tariff and its price schedule. Typically the current price schedule is lower for small purchases and the optimal price schedule is lower for large purchases. Thus, there is again some purchase size \( q^* \) such that the current prices are less expensive for purchases less than \( q^* \), and the optimal prices are less expensive for incremental units in excess of \( q^* \). In this case again no customer is disadvantaged, larger customers benefit, and the firm’s revenues increase — or if not needed for cost recovery then they can be used to reduce all prices.

As in the other cases, the net effect of this Pareto-improving modification is to offer a new menu of options in which the marginal price for the \( q \)-th unit is effectively the lesser of the current price \( p(q) \) and the optimal price \( p^*(q) \). For this reason, some regulatory agencies (such as the United States Federal Communications Commission) explicitly require that new optional tariffs be offered without withdrawal of existing tariffs.

The modified tariff is depicted in Figure 5.5, which assumes that for each unit a customer pays the lesser of the uniform price \( p \) and the optimal monopoly price schedule \( p^*(q) \). If customers are indexed by a type parameter \( t \) then, as in the figure, all types greater than a critical value \( t^* \) increase their purchases and their net benefits increase.
Moreover, the firm’s profit increases because the optimal price schedule \( p^*(q) \) is designed to obtain the maximum profit contribution from units exceeding \( q(t^*) \), which is the last unit purchased by type \( t^* \).

**Optimal Constrained Nonlinear Pricing**

A general formulation that encompasses the three versions cited above is based on explicit recognition of the requirement that the tariff design must not disadvantage any customer. This requirement adds numerous constraints to the problem of maximizing total surplus in addition to the previous constraint that the firm obtains sufficient revenue to recover its costs. In terms of the formulation in §6, if the existing tariff \( P(q) \) enables a customer of type \( t \) to obtain the net benefit \( U(q(t), t) - P(q(t)) \), then the optimal tariff \( P^o(q) \)
must satisfy the constraint
\[ U(q^*(t), t) - P^0(q^0(t)) \geq U(q(t), t) - P(q(t)), \]
for each type \( t \), where \( q(t) \) and \( q^0(t) \) are the optimal purchases of type \( t \) under the two tariffs. If the existing tariff is concave then these constraints imply that the marginal prices have the form
\[ p^0(q) = \min \{ p(q), p^*(q) \}. \]

Here, \( p^*(q) \) is the marginal price schedule for an unconstrained optimal nonlinear tariff that is chosen so that the required revenue is obtained by the firm. The unconstrained price schedule \( p^*(q) \) provides quantity discounts for units in various intervals. In effect, for each \( q \)-th unit a customer has a choice between the old marginal price \( p(q) \) and the new one \( p^*(q) \). When there is a single such interval, it provides quantity discounts for all units in excess of some breakpoint \( q^* \) at which \( p(q^*) = p^*(q^*) \). Thus, customers purchasing fewer than \( q^* \) units select the old tariff and are unaffected by the new tariff. Customers purchasing more than \( q^* \) units benefit from the quantity discounts provided.

A typical situation is shown in Figure 5.6, which assumes that the existing tariff charges a uniform price \( p \). The continuity of the price schedule is a general property: discontinuities of the sort shown in Figure 5.7 do not occur in an optimal price schedule. In each case there is an interval \((q_1, q_2)\) of purchases not selected by any customer. Such a tariff can always be improved by using a price schedule that is intermediate between the two shown, and that benefits both the customers and the firm. The reasons are essentially those already mentioned about why a uniform tariff can be improved.

The following example illustrates the analysis in the case that the existing tariff uses a uniform price.

\[ \text{Example 5.3 :} \] Suppose that the demand profile is \( N(p, q) = 1 - p - q \), resulting from a population with types \( t \) that are uniformly distributed between zero and one, and demand functions \( D(p, t) = t - p \) derived from benefit functions \( U(q, t) = tq - \frac{1}{2}q^2 \). Assume that the firm’s marginal cost is zero for simplicity. For this example, a uniform price \( p \) allows the firm to recover an amount \( R(p) = \frac{1}{2}p(1 - p)^2 \) in revenues, resulting in a total consumers’ surplus of \( CS(p) = \frac{1}{6}(1 - p)^3 \). Hereafter, we assume that \( R(p) \) is the amount of the firm’s revenue requirement and we examine the consequences of meeting this revenue requirement via either unconstrained or constrained nonlinear pricing.

In the case of unconstrained nonlinear pricing, the Ramsey pricing rule implies that the optimal tariff has the property that the percentage profit margin on the \( q \)-th unit
Figure 6: Pareto-improving price schedule.
should be $\alpha/\eta(p(q), q)$, where $\eta$ is the price elasticity of the demand profile. This rule implies that the optimal unconstrained nonlinear price schedule is

$$p(q) = \beta[1 - q],$$

where $\beta = \alpha/[1 + \alpha]$. The revenue realized from this price schedule is $R[\beta] = \frac{1}{2}\beta[1 - \beta]$ and the consumers’ surplus is $CS[\beta] = \frac{1}{8}[1 - \beta]^2$, which by construction is the maximum attainable consumers’ surplus given that the revenue constraint $R[\beta] = R(p)$ is used to determine the appropriate slope $\beta$ of the marginal price schedule.

We first note that the maximum revenue obtainable from any uniform price is 2/27, via $p = 1/3$, whereas an unconstrained nonlinear price schedule can obtain the revenue 2/24 by using $\beta = 1/2$, which is higher by 12.5%. Thus, for revenue requirements exceeding 2/24 per potential customer in the population, no system of nonlinear pricing is sufficient; and for revenue requirements exceeding 2/27 no uniform price is sufficient. In particular, there are some revenue requirements that preclude uniform pricing. In these cases, of course, $q^* = 0$ since it is infeasible to assure that no customer is disadvantaged and still raise the required revenue.

Next we note for revenue requirements obtainable by either uniform or nonlinear pricing that the distributional effects on customers are quite different. The least type served under uniform pricing is $t = p$, whereas with nonlinear pricing it is $t = \beta$. Further, for those types served the net benefit is $\frac{1}{2}[t - p]^2$ under uniform pricing but it is $\frac{1}{2}[t - \beta]^2/[1 - \beta]$ under nonlinear pricing. In particular, if $0 < p < 1/3$ then the revenue constraint $R[\beta] = R(p)$ requires $\beta > p$; consequently, nonlinear pricing serves fewer customers and provides smaller net benefits to those customers with low types — although of course nonlinear pricing provides greater net benefits to high types that could more than compensate.

Now consider a nonlinear tariff that is optimal subject to the further requirement that no customers are disadvantaged as compared to uniform pricing that raises the same revenue for the firm. In this case the optimal price schedule takes the form specified above:

$$p^\circ(q) = \min\{p, p^*(q)\},$$

where now

$$p^*(q) = \gamma[1 - q], \quad \text{and} \quad \gamma = 1/3,$$

independently of which uniform price $p$ is considered. Note that this choice of $\gamma$ corresponds to the multiplier $\lambda = 1$ and the associated Ramsey number $\alpha = 1/2$ that
would yield \( \beta = 1/3 \). With this price schedule, types \( t \leq 1 - 2p \) purchasing quantities less than \( q^* = 1 - 3p \) are charged the uniform price \( p \), whereas those purchasing more pay the uniform price for the first \( q^* \) units and thereafter obtain quantity discounts according to the price schedule \( p^* \) that applies to units in excess of \( q^* \). The optimal price schedule \( p^o \) assures that no customer is disadvantaged compared to uniform pricing, and those with high types obtain greater net benefits, and still the required revenue \( R(p) \) is obtained for the firm.

Table 5.1 compares uniform pricing, unconstrained nonlinear pricing, and constrained nonlinear pricing in terms of the consumers’ surplus that is realized for a fixed revenue requirement. Observe that unconstrained nonlinear pricing realizes the maximum consumers’ surplus, but also constrained nonlinear pricing realizes more consumers’ surplus than can be obtained from uniform pricing, as well as assuring that no customer is disadvantaged. Over much of the range of revenue requirements, constrained nonlinear pricing yields at least half as much gain in consumers’ surplus as does unconstrained nonlinear pricing, and more if the revenue requirement is high. As can be seen in Figure 5.8, the differences in consumers’ surplus are small compared to the effects of variations in the revenue requirement.

\[ \diamond \quad \text{Example 5.4:} \quad \text{A similar pattern can be seen in Figure 5.9, which is based on Example 4.3 in the case that the substitution parameter is } a = 0.4, \text{ the correlation between the type parameters is } r = 0.4, \text{ the price of a substitute commodity is } p^* = 0.50, \text{ and the firm’s marginal cost is } c = 0. \text{ The figure shows the revenues from uniform pricing (the dashed curve) and constrained nonlinear pricing for several Ramsey numbers.} \]

\[ \diamond \]

\[ \text{As in the previous example, revenue is relatively insensitive to the choice of the Ramsey number if the price cap } p \text{ is substantially less than a monopolist’s optimal uniform price. This occurs whenever a larger Ramsey number affects revenue mainly via higher prices for initial units, but these higher prices are curtailed by the price cap } p. \text{ In the figure, variations around the revenue-equivalent Ramsey number produce variations in the revenue that are virtually imperceptible. There is in fact some variation but it is quite} \]

\[\text{In this example the Ramsey number } \alpha \text{ associated with the optimal price schedule } p^* \text{ is independent of the uniform price } p \text{ and its associated revenue } R(p). \text{ This feature is peculiar to a limited class of models. The fact that it does not hold generally is illustrated for the next example in Figure 5.10. My experience with examples indicates, however, that the optimal Ramsey number is often insensitive to the uniform price used as the price cap for the associated nonlinear schedule.} \]

\[\text{Revenue is measured per potential customer, defined as those 80.86% with positive demands at a zero price, by normalizing so that } N(0, 0) = 1.\]
Table 5.1

Comparison of Three Pricing Policies

<table>
<thead>
<tr>
<th>Revenue Requirement $R(p)$:</th>
<th>0.0049</th>
<th>0.0226</th>
<th>0.0405</th>
<th>0.0542</th>
<th>0.0640</th>
<th>0.0703</th>
<th>0.0735</th>
</tr>
</thead>
</table>

Uniform Pricing

<table>
<thead>
<tr>
<th>Price $p$:</th>
<th>0.01</th>
<th>0.05</th>
<th>0.10</th>
<th>0.15</th>
<th>0.20</th>
<th>0.25</th>
<th>0.30</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS:</td>
<td>0.1617</td>
<td>0.1429</td>
<td>0.1215</td>
<td>0.1024</td>
<td>0.0853</td>
<td>0.0703</td>
<td>0.0572</td>
</tr>
</tbody>
</table>

Constrained Nonlinear Pricing

<table>
<thead>
<tr>
<th>Breakpoint $q^*$:</th>
<th>0.97</th>
<th>0.85</th>
<th>0.70</th>
<th>0.55</th>
<th>0.40</th>
<th>0.25</th>
<th>0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS:</td>
<td>0.1617</td>
<td>0.1430</td>
<td>0.1222</td>
<td>0.1046</td>
<td>0.0907</td>
<td>0.0807</td>
<td>0.0752</td>
</tr>
</tbody>
</table>

Unconstrained Nonlinear Pricing

<table>
<thead>
<tr>
<th>Slope $\beta$:</th>
<th>0.0149</th>
<th>0.0730</th>
<th>0.1415</th>
<th>0.2043</th>
<th>0.2592</th>
<th>0.3024</th>
<th>0.3282</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS:</td>
<td>0.1617</td>
<td>0.1432</td>
<td>0.1228</td>
<td>0.1055</td>
<td>0.0915</td>
<td>0.0811</td>
<td>0.0752</td>
</tr>
</tbody>
</table>

small compared to the scale of total revenue: Figure 5.10 shows the difference between the revenues from constrained nonlinear pricing and uniform pricing for several Ramsey numbers. Each circled point identifies the price cap (indicated by a mark along the abscissa) that if used with the associated Ramsey number produces the same revenue from constrained nonlinear pricing as from uniform pricing. Observe that the scale of the difference in revenues is multiplied by a factor of 10,000 to produce perceptible variation in the vertical dimension. Figure 5.11 shows the resulting relationship between the Ramsey number and the price cap. Each point on the locus represents a Ramsey number and a price such that uniform pricing with that price yields the same revenue as constrained nonlinear pricing with that Ramsey number and price cap. The curve is nearly flat for small price caps because there are few customers for whom the quantity discounts offered by the declining portion of the price schedule is applicable.

The chief implication of these and similar examples is that often there are appreciable gains from nonlinear pricing if the firm’s revenue requirement is large; moreover, the costs of assuring that no customer is disadvantaged are relatively small. It should be
Figure 8: Comparison of three pricing policies. Consumers’ surplus is more sensitive to the revenue requirement than to constraints on the form of the price schedule.

remembered, however, that there are other cases in which uniform pricing is the optimal means of meeting the firm’s revenue requirement: this is the case in the first example above if the types are distributed in the population according to an exponential distribution, rather than a uniform distribution as assumed. In these cases, there are no further gains from offering quantity discounts for large purchases.

In Section 3 we illustrate further with an application to the design of a Pareto-improving tariff for telephone service. In this application customers are classified into several groups based on their levels of usage under the existing tariff. The design problem is therefore posed as a constrained maximization problem of the sort usually solved by mathematical programming. That is, in addition to the usual criterion that the tariff
Figure 9: Example 5.4: The dependence of revenue on the price cap $p$ and the Ramsey number $\alpha$. The dashed curve shows the revenue from uniform pricing.
maximizes an appropriately weighted average of consumers’ surplus and the firm’s net revenue, explicit constraints are added to represent the requirement that customers in each existing volume band are as well off under the new tariff as under the old one. Often this approach results in very nearly the answer derived above: if the optimal tariff is constrained to be Pareto-improving, it charges marginal prices that are the lesser of the old prices and the unconstrained optimal prices so constructed as to meet the revenue requirement. The sole exception is the price $p(q^*)$ at the margin between the old and new tariffs, since usually a customer buying a last unit at this price under the old tariff now has an incentive to buy additional units under the new tariff. The example in Section 3 has this property.
Dependent Demands

The above construction of Pareto-improving tariffs depends on the assumption that customers’ demands are independent. In important cases, however, customers’ demands are jointly dependent in ways that preclude Pareto improvements. In fact, quantity discounts for large customers might conceivably reduce the utility’s revenue.

In the extreme case that illustrates this possibility, the customers are firms in a highly competitive industry and they use the service sold by the utility as an input into their own production operations. Further, customers are heterogeneous because there are some large firms, whose monopoly on a superior low-cost technology allows positive profits, as well as some small firms (the “competitive fringe”) whose marginal cost determines the price of the industry’s product. To be precise, assume that both types of firms use a fixed quantity of the utility’s service for each unit of output. In this case, if the utility offers an additional quantity discount advantageous only to the large firms, then the industry’s market price, and therefore also the demand for the industry’s product, remains unchanged, since the small firms’ marginal cost remains unchanged; however, the additional cost advantage given the large firms further increases their market share. The net effect for the utility is that the increment in the large firms’ purchases results in an exactly offsetting decrement in the small firms’ purchases. Thus, the utility’s revenues decline, because the price paid by the small firms is higher than the marginal price paid by the large firms.\(^8\)

This is not a complete argument that no Pareto-improving tariff is possible because one must still show that there is no way to design an alternative tariff that redistributes some of the large firms’ gains to the small firms and the utility so that none are disadvantaged.\(^9\) It shows nevertheless that the simple constructions described above need

\(^8\) As Ordover and Panzar (1980) note, this is essentially the “secondary line injury” to competing small firms that is prohibited in the United States by the Robinson-Patman Act of 1936. That is, a quantity discount for large firms can reduce the market share of small firms. In the present example the utility’s revenue declines too, so such a discount might not be offered voluntarily, but the Act was intended to address cases in which a large customer has enough market power to bargain for price concessions. Such bargaining power requires, of course, that a large customer can credibly threaten to operate without the utility’s service. In the case of electric power (or telecommunications), the feasibility of cogeneration (or bypass) adds credibility to such threats, but the Act also provides as a defense that the seller can meet an equally low price of a competitor. Cogeneration presumably qualifies as a competitor.

\(^9\) A complete proof, using some further assumptions, is provided by Ordover and Panzar (1980). To show that a Pareto-improving tariff is not generally possible it suffices
not suffice.

5.3. Long-Distance Telephone Tariffs

We describe an application of nonlinear pricing to the design of long-distance telephone rates to ensure that no class of customers is disadvantaged compared to a prevailing uniform price.\textsuperscript{10}

In the United States since 1984, interexchange (that is, long-distance) carriers have been charged two fees (\$/minute) for connections between their networks and their customers: a line charge to recover costs of the local loop that are not sensitive to traffic, and a traffic-sensitive charge for switching and trunk-line services in the local-exchange carrier’s network. In 1986 these fees averaged 7.56 \$/min, for peak-period originating switched access for measured services. Generally they were passed on to customers as part of their uniform charge (\$/min) for long-distance service. For large customers these charges are appreciable and there has been a resulting incentive for such customers to install direct connections that bypass the local loop and switching. In June 1986, NYTel Company filed with the Federal Communications Commission a request for a nonlinear tariff to be billed directly to customers, partly to forestall bypass by its larger customers. The nonlinear tariff can be construed as a block-declining tariff or as the lower envelope of a set of six two-part tariffs that will be described below.

Table 5.2 shows the current distribution of NYTel’s customers’ average usage under the prevailing uniform charge, divided into usage bands. Also shown are the current elasticities of demand for customers in each usage band based on econometric studies, and an alternative schedule of roughly estimated elasticities based on long-run prospects for bypass: the higher elasticities for high-usage customers indicates that they are more price-sensitive due to the alternative of using bypass. In the following, we assume that demand by each usage band \(i\) has the form of the constant-elasticity demand function

to consider the case that the utility’s current tariff uses a uniform price that maximizes its profit and, because the customers’ industry is competitive, therefore also maximizes the combined profits of the utility and the customers’ industry. In this case, it is evidently impossible to increase the profits of all three parties, since they are already fully maximized. In effect, if the utility’s uniform price is already an optimal way of taxing the profits of the large firms, then there is no advantage from quantity discounts. It should be noted, however, that this argument is not sufficient for a fully enlightened analysis in a regulatory context, since it does not take into account the consequences for final consumers of the industry’s product, who would benefit from reductions in the small firms’ marginal cost.

\textsuperscript{10} This section is based on Heyman, Lazorchak, Sibley, and Taylor (1987).
§5.3. Long-Distance Telephone Tariffs

Table 5.2
Distribution of Usage and Elasticities

<table>
<thead>
<tr>
<th>Type</th>
<th>Band</th>
<th>Usage % of Accounts</th>
<th>Monthly Minutes</th>
<th>Monthly Elasticities</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0 - 60</td>
<td>74.03</td>
<td>14.55</td>
<td>.16</td>
</tr>
<tr>
<td>2</td>
<td>61 - 1000</td>
<td>25.47</td>
<td>160.21</td>
<td>.16</td>
</tr>
<tr>
<td>3</td>
<td>1001 - 2000</td>
<td>0.26</td>
<td>1364.46</td>
<td>.22</td>
</tr>
<tr>
<td>4</td>
<td>2001 - 7000</td>
<td>0.17</td>
<td>3547.77</td>
<td>.22</td>
</tr>
<tr>
<td>5</td>
<td>7001 - 20000</td>
<td>0.05</td>
<td>11026.07</td>
<td>.31</td>
</tr>
<tr>
<td>6</td>
<td>20000+</td>
<td>0.02</td>
<td>67425.60</td>
<td>.31</td>
</tr>
</tbody>
</table>

$q_i = t_i/p_i^{e_i}$, where $p_i$ is the marginal price, $e_i$ is the elasticity, and $t_i$ indicates a type parameter associated with that class of customer. From the uniform price $p = 7.56 \,$c, the monthly minutes of use, and the elasticities shown in the table, the parameter $t_i$ can be calculated. As for costs, the marginal cost of service is assumed to be 1 \,$c/minute based on a similar rate filing by New England Telephone.

Table 5.3
Predicted Consequences of NYTel’s Proposed Tariff (Alternative Elasticities)

<table>
<thead>
<tr>
<th>Customer Type</th>
<th>Proposed Tariff</th>
<th>Change in Consumer Surplus</th>
<th>Change in Producer’s Surplus</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.0961</td>
<td>$-0.29$</td>
<td>$0.25$</td>
</tr>
<tr>
<td>2</td>
<td>.0713</td>
<td>-0.80</td>
<td>0.89</td>
</tr>
<tr>
<td>3</td>
<td>.0484</td>
<td>16.85</td>
<td>0.36</td>
</tr>
<tr>
<td>4</td>
<td>.0352</td>
<td>119.61</td>
<td>-50.93</td>
</tr>
<tr>
<td>5</td>
<td>.0302</td>
<td>582.87</td>
<td>-214.14</td>
</tr>
<tr>
<td>6</td>
<td>.0269</td>
<td>5,061.42</td>
<td>-1,134.39</td>
</tr>
<tr>
<td>Average</td>
<td>$1.135$</td>
<td>$-0.007$</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.3 shows the block-declining (or “tapered”) tariff proposed by NYTel and the
predicted consequences based on the assumed model, as compared to the current uniform charge, based on the alternative set of elasticities. The changes in surplus are recorded in terms of $ per month per customer. Note that the proposed tariff is nearly revenue neutral but it provides substantial benefits to the larger customers. It is economically efficient in that it increases total surplus, but it is not distributionally neutral since some customer types are disadvantaged.

Table 5.4

Optimized Tariff: I
(Alternative Elasticities)

<table>
<thead>
<tr>
<th>Type</th>
<th>Fixed Fee $/mon.</th>
<th>Usage Charge $/min.</th>
<th>Change in Consumers’ Surplus</th>
<th>Change in Producer’s Surplus</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$0</td>
<td>$.0756</td>
<td>$0</td>
<td>$0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>.0756</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0.52</td>
<td>.0752</td>
<td>0</td>
<td>0.23</td>
</tr>
<tr>
<td>4</td>
<td>29.18</td>
<td>.0674</td>
<td>0.82</td>
<td>12.06</td>
</tr>
<tr>
<td>5</td>
<td>342.17</td>
<td>.0446</td>
<td>66.00</td>
<td>170.61</td>
</tr>
<tr>
<td>6</td>
<td>3,495.58</td>
<td>.0238</td>
<td>2,350.00</td>
<td>1,956.60</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>$0.50</td>
<td>$0.50</td>
<td></td>
</tr>
</tbody>
</table>

An alternative methodology is to impose constraints on the design of the tariff to ensure that no customer type is disadvantaged. Table 5.4 shows the result of such a calculation done with the objective of maximizing the sum of consumers’ and producer’s surplus. Note that these results show that it is possible with such a set of optional two-part tariffs to disadvantage no customer, and yet increase profits (that is, producer’s surplus). A similar calculation for the current elasticities (rather than the alternative ones) is shown in Table 5.5: in this case the average change in profit is +10 ¢ per customer per month, and it is positive for every type.

These results are indicative of a developing methodology for designing rates for public utilities. The aim is to construct the rates as nonlinear tariffs so as to obtain the greatest gains in efficiency compatible with assuring that no major subgroup of customers is disadvantaged by the change to the new tariff. This is done mainly by adding new options to the current menu of tariffs. This constraint is useful in ensuring that the new
<table>
<thead>
<tr>
<th>Type</th>
<th>Fixed Fee $/mon.</th>
<th>Usage Charge $/min.</th>
<th>Change in Consumers’ Surplus</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$0</td>
<td>.0756</td>
<td>$0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>.0756</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>.0756</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0.49</td>
<td>.0754</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>73.27</td>
<td>.0690</td>
<td>1.04</td>
</tr>
<tr>
<td>6</td>
<td>2,989.17</td>
<td>.0313</td>
<td>381.00</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>$0.077</td>
<td></td>
</tr>
</tbody>
</table>

A regulated firm can be construed as setting prices efficiently to meet a revenue requirement. The formulation in terms of Ramsey pricing interprets efficiency in terms of maximizing consumers’ surplus, using only the minimal monopoly power required to raise the required revenue. In this case the firm proceeds in exactly the same fashion as a profit-maximizing firm except that profit margins are reduced proportionately and uniformly so that no more than the required revenue is obtained.

Applications of Ramsey pricing are often constrained by the requirement that no customer is disadvantaged by the changes proposed. In such cases, a Pareto-improving tariff can be constructed by allowing customers the option of purchasing each unit from the old price schedule or the new one. The new tariff can be of various kinds, such as a two-part tariff or the old tariff amended by inclusion of additional quantity discounts, such as a segment of an optimal price schedule. All of these can benefit some or all customers without reducing the firm’s revenue, or they can be used to increase the firm’s profits.

When the old tariff uses a uniform price, a nonlinear tariff that is optimally designed to meet the revenue requirement typically includes the old tariff plus quantity discounts for large purchases. In examples, the cost of assuring that no customer is disadvantaged is
small compared to the effect of the revenue requirement. This feature is largely explained
by the fact that the gains from nonlinear pricing are concentrated among customers
making large purchases, so the costs of providing assurances to customers making small
purchases are relatively small.

Nonlinear tariffs can also meet revenue requirements that cannot be attained with
uniform pricing. This feature can be important in providing services that otherwise could
not be offered profitably. It is especially important in industries affected by network ex-
ternalities. Telecommunications provides the standard example: each customer’s benefits
depend on how many others subscribe to service and thereby become available to ex-
change calls. Starting a new service therefore requires attracting a critical mass of initial
customers large enough to create mutual benefits among themselves and thereafter to
attract additional subscribers. Moreover, these benefits must be sufficient to allow prices
that will recover at least the hookup, access, and fixed operating costs of the system.
Pricing based on, say, a fixed fee or a two-part tariff has the disadvantage that it requires
a larger critical mass than does nonlinear pricing. For example, Oren, Smith, and Wilson
(1982a) report on an application to the design of a special system for subscribers with
impaired hearing, for which nonlinear pricing enables a critical mass that is 22% of the
size required by a fixed fee, and 41% of the size required by a two-part tariff. Once
the system is fully established, the optimal profit-maximizing nonlinear tariff generates
more total surplus and more profits for the firm but less consumers’ surplus for sub-
scribers; the average charge per subscriber is the nearly the same for all three regimes.\footnote{When there are positive externalities of this and other kinds, an optimal nonlinear
tariff charges a fixed fee that is less than the firm’s cost of providing access. Due to the
benefits created for other subscribers, it is advantageous to attract additional customers
by subsidizing access. This is a main source of the smaller critical mass, and after the
system is fully established it also results in a market penetration that is larger.}
Nevertheless, without nonlinear pricing the larger critical mass might mean that the sys-
tem would never be established. Thus, to the advantages of nonlinear pricing cited in
this chapter one can add that its use might be necessary to provide services that would

140
otherwise be unprofitable or dependent on initial subsidies to reach a critical mass.

Part II

DISAGGREGATED DEMAND MODELS
Chapter 6

SINGLE-PARAMETER DISAGGREGATED MODELS

The basic principles of nonlinear pricing described in Part I rely on the demand profile to summarize the minimal data required to construct an optimal price schedule. This chapter presents a parallel exposition using explicit models of customers’ preferences or demand behaviors. In these models, individual customers or market segments are characterized by parameters indicating their types, and their benefits or demands are estimated directly as functions of these parameters. Disaggregated models of this sort are frequently used in econometric studies to represent how customers’ demands are affected by various characteristics, such as income and socio-demographic category for residential customers, or line of business and production rate for commercial customers.

The analyses in previous chapters account only for the aggregate demand for each increment in the purchase size. In this chapter the analysis is conducted at the same disaggregated level as the data in the model. Thus, the quantities selected and the prices paid by individual customers or market segments are specified explicitly. To simplify the presentation, however, the exposition is confined to the case that the differences among customers are described by a single one-dimensional type parameter with a known distribution in the population. In §8.4 we show that using a single parameter entails no substantial loss of generality when the tariff applies to a single product.

The presentation is simplified initially by assuming in Section 1 that the population of customers is finite. Section 2 adopts the other extreme hypothesis: the population of customers is so large that it can be modeled as a continuum in which each customer is described by a one-dimensional type parameter with a continuous distribution in the population. The special case of a two-part tariff is examined briefly in Section 3, followed by a detailed analysis of general multipart tariffs in Section 4. The extension to fully nonlinear tariffs is presented in Section 5, followed by several examples in Section 6. Lastly, Section 7 characterizes optimal fixed fees; in particular, optimal nonlinear tariffs omit fixed fees if the firm incurs no fixed cost in serving a customer. This is not true
§6.1. A Model with Discrete Types

for optimal multipart tariffs, but the difference is insignificant if even a few options are offered.

Section 1 assumes that the firm designs the tariff to maximize its profit contribution. Later sections include Ramsey pricing in which the tariff is designed to maximize consumers’ or total surplus subject to a constraint that the firm’s profit contribution is sufficient to meet a revenue requirement. Only necessary conditions for an optimal tariff are considered in this chapter: sufficiency conditions are deferred to §8.

6.1. A Model with Discrete Types

The basic element of the formulation is a model of the heterogeneity of customers’ demand behaviors. For this purpose, each customer or market segment is classified as one of several types, and the model specifies the demand behavior of each type. The possible types are indicated by an index \( i = 1, \ldots, m \). We assume that the firm knows the number or fraction \( f_i \) of the customers classified into each type \( i \), and that the demand of each type is predicted exactly.

A customer of type \( i \) is described by a utility function \( U_i(q) \) indicating the gross benefit from a purchase of size \( q \). If the customer purchases this quantity from the tariff \( P(q) \) then the net benefit is \( U_i(q) - P(q) \). There are no income effects so benefits are measured directly in dollar amounts, and \( U_i(0) \equiv 0 \).

The firm’s costs generally depend on the list \( Q = (q_1, \ldots, q_m) \) of all customers’ purchases according to a function \( C(Q) \). For example, one often assumes that \( C(Q) = C(\sum_i f_i q_i) \), or \( C(Q) = \sum_i f_i C(q_i) \), or some combination. We assume the latter form here for simplicity.

This formulation implies that the firm’s profit contribution is

\[
\text{Profit Contribution} = \sum_{i=1}^{m} f_i \cdot [P(q_i) - C(q_i)]
\]

when each customer of type \( i \) purchases the quantity \( q_i \) in response to the tariff \( P \). Providing incentives for customers to make these purchases requires that the tariff satisfies two kinds of constraints. The first kind assures that a customer prefers to make the designated purchase rather than none at all:

\[
U_i(q_i) - P(q_i) \geq U_i(0) - P(0) \equiv 0.
\]
This is called the participation constraint for type \(i\). Note that it excludes charging a positive amount unless the customer makes a positive purchase. The second kind assures that a customer prefers to make the designated purchase rather than one assigned to another type:

\[
U_i(q_i) - P(q_i) \geq U_i(q_j) - P(q_j), \quad \text{for each } j \neq i.
\]

These are called the incentive compatibility constraints for type \(i\). In the following we use \(P_i \equiv P(q_i)\) to denote the total charge imposed by the tariff for the purchase of size \(q_i\).

Suppose first that all the variables \(P_i\) and \(q_i\) are restricted to a finite set of possible values. In this case the optimal tariff is found by searching among the possible combinations of these values satisfying the participation and incentive-compatibility constraints to find a combination that attains the maximum feasible value of the profit contribution. Having found such an optimal combination, the tariff is specified by assigning the total charge \(P_i\) to the purchase size \(q_i\). For other quantities not among those assigned to any type, the usual specification is that the tariff \(P(q)\) is the same as the charge for the next larger assigned purchase; this is sufficient to deter purchases in the interval \(q_{i-1} < q < q_i\).

Finding an optimal combination of charges and purchase sizes as described above is a tedious task, and it is time-consuming even on a fast computer. The preferred method, therefore, is based on a specification that allows the possible purchase sizes to be any real numbers, and similarly the possible charges can be any real numbers. Hereafter, therefore, the purchase sizes are constrained only to be nonnegative real numbers \((q_i \geq 0)\) and similarly for the tariff charges. In this case, of course, the model must also specify the gross benefit function \(U_i(q)\) for each type in a way that is tractable for analysis. We generally assume convenient regularity conditions: \(U_i\) is increasing, concave, and differentiable. Given real domains for the purchase sizes and tariff charges, maximizing the profit contribution subject to the participation and incentive-compatibility constraints is a standard problem of nonlinear constrained optimization. For some applications, reliance on standard software to solve such problems is the best practical approach.

**Characterization of an Optimal Tariff**

Our interest here is to establish the key properties of the solution that can be derived from mathematical principles. The necessary conditions for an optimum can be expressed in
6.1. A Model with Discrete Types

terms of a Lagrange multiplier, say \( \lambda_{ij} \), associated with the \( j \)-th incentive-compatibility constraint for type \( i \). Such a multiplier is a nonnegative number that is positive only if the constraint is actually binding at the optimum. The key properties that can be derived are the following.

1. If type \( i \) is such that both \( P_i \) and \( q_i \) are positive, so that type \( i \) is an active customer, then

\[
v_i(q_i) - c(q_i) = \sum_{j \neq i} \lambda_{ji} \frac{f_j}{f_i} [v_j(q_i) - v_i(q_i)],
\]

where \( v_i(q) = U'_i(q) \) is the \( i \)-th type’s marginal benefit or willingness-to-pay for a \( q \)-th unit, and \( c(q) = C'(q) \) is the firm’s marginal cost.

2. If type \( i \)’s net benefit is positive, so that its participation constraint is not binding, then

\[
f_i = \sum_{j \neq i} [\lambda_{ij} f_i - \lambda_{ji} f_j].
\]

Recall that a customer of type \( i \) is predicted ultimately to choose a purchase size \( q_i \) such that \( v_i(q_i) = p(q_i) \), where \( p(q_i) \) is the tariff’s marginal price for the \( q_i \)-th unit. Consequently, property 1 states that the firm’s profit margin \( p(q_i) - c(q_i) \) on the \( q_i \)-th unit should be given by a weighted combination of the differences between the willingness to pay for this unit by other types \( j \neq i \) and by type \( i \). These weights are positive at the optimum only for those other types who are indifferent between their assigned purchases and type \( i \)’s purchase \( q_i \). Property 2 constrains what these weights can be: they must satisfy a consistency condition.

Without further assumptions, the construction of a solution that satisfies properties 1 and 2 can be difficult and usually one must rely on numerical analysis via computer programs for nonlinear constrained optimization. Here, however, we impose a further assumption of the sort used in most theoretical studies of nonlinear pricing in order to obtain a simple characterization of the optimal tariff. The purpose of this assumption is to obtain the special case that for each active type \( i \) the only binding incentive-compatibility constraint is the one for type \( j = i - 1 \); that is, only the next higher type is indifferent between its purchase and \( i \)’s purchase. Of course this special case depends on having initially ordered the customers’ types in such a fashion that this is possible. For example, one assumption that is commonly used to obtain this special case is that higher types have higher demands at every price, as we illustrate in later sections. Even with much
weaker assumptions, nevertheless, this special case often obtains, simply because the optimal tariff often has the property that each type’s purchase size is “envied” by at most one other type.¹

Given such an assumption, we can exploit the fact that the multiplier \( \lambda_{ij} \) is positive only for \( j = i - 1 \). In this case, if we define \( \hat{\lambda}_i = f_i \lambda_{i,i-1} \) then property 2 can be cast in the form

\[
\hat{\lambda}_i - \hat{\lambda}_{i+1} = f_i, \quad \text{and} \quad \hat{\lambda}_{m+1} = 0,
\]

from which it follows that

\[
\hat{\lambda}_i = \hat{F}_i, \quad \text{where} \quad \hat{F}_i = \sum_{j \geq i} f_j.
\]

Thus, \( \hat{F}_i \) is the number of customers of types \( i, i + 1, \ldots, m \) (and \( \hat{F}_{m+1} = 0 \)), and the weight \( \hat{\lambda}_i \) must be equal to this number. Using this fact in property 1 provides the key characterization:

\[
v_i(q_i) = c(q_i) + \frac{\hat{F}_{i+1}}{f_i} [v_{i+1}(q_i) - v_i(q_i)].
\]

This equation determines the purchase size \( q_i \). To induce a customer of type \( i \) to select the indicated quantity \( q_i \), the marginal price for the \( q_i \)-th unit must be \( p_i = v_i(q_i) \). And, from the incentive-compatibility condition we obtain further that the tariff is constructed recursively by the formula

\[
P_i = P_{i-1} + U_i(q_i) - U_i(q_{i-1}),
\]

starting from \( P_{i^*} = U_{i^*}(q_{i^*}) \), where \( i^* \) is the least type willing to purchase at the marginal price for its assigned quantity. If the resulting tariff is concave, then one implementation offers a menu of \( m \) two-part tariffs that associates the marginal price \( p_i \) with the fixed fee \( P_i - p_i q_i \).

A useful version of the optimality condition restates it as the requirement that

\[
[v_i(q_i) - c(q_i)]\hat{F}_i = [v_{i+1}(q_i) - c(q_i)]\hat{F}_{i+1}.
\]

In this form it states that the firm’s profit contribution from the sale of the \( q_i \)-th unit is the same whether it is sold at the marginal price \( v_i(q_i) \) to those types \( j \geq i \) or at the marginal price \( v_{i+1}(q_i) \) to those types \( j > i \). Figure 6.1 illustrates this calculation.

by showing for each type $i$ the profit contribution $[v_i(q) - c(q)] \hat{F}_i$ from the $q$-th unit when it is assigned to type $i$. The purchase size $q_i$ assigned to type $i$ is the largest for which the profit contribution is greater than would be obtained from assigning that unit to the next higher type. Subject to this condition, the total tariff is constructed so that each type is indifferent between its assigned purchase $q_i$ or the purchase assigned to the next lower type. This is shown in Figure 6.2, which graphs each type’s locus of pairs $(q, P)$ of purchase sizes and tariff payments that yield the net benefit $U_i(q_i) - P_i$ from the assignment $(q_i, P_i)$.

This characterization is essentially equivalent to the one derived previously in §4 using the demand profile $N$, as can be seen by identifying $\hat{F}_{i+1}$ with the number $N(p_i, q_i)$ of customers willing to purchase the $q_i$-th unit at the marginal price $p_i$. The exposition in §4 is basically an intuitively sensible way of presenting the conclusions derived here.

**Extension to a Continuum of Types**

This analysis can be extended straightforwardly to the case that customers’ types are described by a continuum of real-valued parameters $t$. In this case, taking the limit as the difference between adjacent types shrinks to zero yields the analogous characterization of each type $t$’s purchase $q(t)$:

$$v(q(t), t) = c(q(t)) + \frac{F(t)}{f(t)} \cdot \frac{\partial v}{\partial t}(q(t), t).$$

In this version, $f(t)$ is interpreted as the density of customers of type $t$, $F(t)$ is the number of customers of types higher than $t$, and $v(q, t)$ is the willingness to pay for the $q$-th unit by type $t$. Given the schedule $q(t)$ of predicted purchases by the various types, the marginal price schedule is $p(q) = v(q, t(q))$, where $t(q)$ is a type purchasing $q$. Figure 6.3 and Figure 6.4 illustrate the calculation of the optimal price schedule for an example in which marginal cost is nil. As shown in the first figure, for many of the models used in practice the hazard rate $f(t)/F(t)$ is increasing but $v(t)/v(q, t)$ is decreasing in $t$ and increasing in $q$. The intersections shown in the first figure identify the three types assigned the three quantities $q_i$ shown, and then these are used in the second figure to identify the corresponding marginal price $p(t)$ at which each type purchases its marginal unit $q(t)$ as the type’s marginal value $v(q(t), t)$ of that unit.
Figure 1: Each type is assigned all increments for which the profit contribution exceeds that from assigning it to the next higher type.
Figure 3: Construction of the assignment of types to quantities purchased, for three quantities \( q_1 < q_2 < q_3 \) and their assigned types \( t_1 < t_2 < t_3 \).
Alternatively, if \( D(p, t) \) is the demand function of type \( t \) then one can use the property that \( q = D(p, t) \) when \( p = v(q, t) \) to derive:

\[
\frac{\partial v(q, t)}{\partial t} = -\left[ \frac{\partial D}{\partial p}(p, t) \right] \cdot \frac{\partial D}{\partial t}(p, t) = -\frac{D_t}{D_p}.
\]

Consequently, in terms of the demand function the condition that characterizes the optimal price schedule is

\[
[\hat{p}(t) - c]D_p(\hat{p}(t), t)f(t) + D_t(\hat{p}(t), t)\hat{F}(t) = 0.
\]

In this version, each type \( t \) is assigned a marginal price \( \hat{p}(t) \) and then the inferred purchase is \( q(t) = D(\hat{p}(t), t) \), which is the quantity at which the marginal cost \( c \) is evaluated. The price schedule is inferred from the relation \( p(q(t)) = \hat{p}(t) \). A somewhat more intuitive rendition of the optimality condition is:

\[
[\hat{p}(t) - c] \cdot [\partial_p D(\hat{p}(t), t)] \cdot f(t)dt + dp \cdot [\partial_t D(\hat{p}(t), t)] \cdot \hat{F}(t) = 0.
\]

In this form it states that the loss from raising type \( t \)'s marginal price, which reduces the profit contribution by reducing type \( t \)'s demand by \( \partial_p D \) for each of the \( f(t)dt \) customers of that type, is compensated by the increment \( dp \) in the profit margin obtained from higher types. Note that type \( t + dt \) is newly assigned the price \( \hat{p}(t) \) previously assigned to type \( t \), which is why the indicated change in demand for higher types is \( \partial t D \).

In subsequent sections we develop this characterization in greater detail, including the extension to Ramsey pricing when the firm must meet a revenue requirement. After specifying details of the formulation, we derive analogous characterizations of two-part and multipart tariffs. Then we reconstruct the characterization for nonlinear tariffs directly from the basic formulation.

### 6.2. Models with One-Dimensional Types

In this section we formulate a general model with a one-dimensional continuum of customers’ types or market segments. This model will be used to justify the construction in Section 1 as well as the demand-profile formulation in §4.

As in the discrete case, the basic elements of the formulation include a model of each type’s demand behavior and a distribution of customers’ types in the population. Unlike the demand-profile formulation, this formulation requires two separate estimates
obtained from demand data. However, the degrees of freedom in the estimation procedure remain essentially the same: the distribution of types merely counts the number of customers with each estimated demand function, and indeed could be uniform without any loss of generality.

The type index $t$ is assumed to be a real number in an interval of possible types. The distribution function $F(t)$ indicates the number (viz., the measure or proportion) of customers having type indices not greater than $t$. The preferences of a customer of type $t$ are specified by a utility function $U(q,t)$ indicating the gross benefit obtained from purchase of a quantity $q$. If the customer purchases this quantity at a cost $P(q)$ then the net benefit is $U(q,t) - P(q)$. Again, there are no income effects, benefits are measured directly in dollar amounts, and $U(0,t) = 0$. Figure 6.5 portrays schematically a customer of type $t$’s selection of an optimal purchase $q(t)$: the selected quantity maximizes the net benefit available from the locus of quantities and associated charges along the tariff. In the figure, the concavity of the tariff reflects a decreasing schedule of marginal prices.

Each type can also be described by its demand function $D(p,t)$, indicating the optimal purchase in response to the uniform price $p$. If $v(q,t) \equiv U_q(q,t)$ is type $t$’s marginal valuation of the $q$-th unit, then the demand function satisfies $v(D(p,t),t) = p$, provided this is evaluated where the price schedule intersects the demand function from below.

We use the following technical assumptions:²

A1: The type distribution function has a corresponding density function $f(t) = F'(t)$ that is positive on an interval $a \leq t \leq b$. The number of customers having types greater than $t$ is denoted by $\bar{F}(t) \equiv 1 - F(t)$. We generally assume that the hazard rate $f(t)/\bar{F}(t)$ is increasing.

A2: The utility function is smoothly differentiable and increasing in each argument $q$ and $t$, and concave in $q$; that is, in terms of its partial derivatives, $U_q > 0$, $U_{qq} < 0$, and $U_t > 0$ if $q > 0$. Moreover, the marginal benefit from an incremental unit increases as the customer’s type increases; that is, $U_{qt} > 0$ or equivalently $v_t > 0$. This says that higher types demand more at each price.

A3: The cost function is smoothly differentiable, nondecreasing, and convex; in particular, the marginal cost $c(q)$ is nondecreasing.

² An additional assumption A4 is included in §6.1 to obtain fully sufficient conditions. These assumptions can be recast in terms of the property of quasi-supermodularity as in Milgrom and Shannon (1991) to obtain necessary and sufficient conditions.
Figure 5: A customer’s purchase selection: the net benefit \( U(q, t) - P(q) \) is maximized where the marginal benefit \( v(q, t) \) intersects the price schedule \( p(q) \).

The first assumption A1 says essentially that there are many customers in the population and their types are not bunched around any one type. The firm is assumed to know the distribution of types, so the population and its economic environment is implicitly assumed to be sufficiently stable that the firm can learn from experience. The second is most easily seen as specifying that each type \( t \) has a demand function \( D(p, t) \) that is a decreasing function of the marginal price \( p \) for incremental units. Moreover, a customer with a larger type index has uniformly higher demand, so if \( t' > t \) then \( D(p, t') > D(p, t) \) at every price \( p \). This excludes the possibility that two types’ demand functions intersect. Thus, the types signify market segments that are ordered by the sizes of their purchases, independently of the price charged. Figure 6.6 depicts an example for
which several demand functions are shown, as well as an indication of the price schedule \( p(q) \) for the \( q \)-th unit. The restrictions imposed by the assumptions are shown in the figure; most evident is the feature that a higher type has a uniformly higher demand function.

Note that a demand locus \( q = D(p, t) \) shown as a function of the price \( p \) can also be interpreted as the marginal benefit \( p = v(q, t) \) when interpreted as a function of the purchase size \( q \). According to assumption A2, \( v \) is a decreasing function of \( q \) and an increasing function of \( t \).

**Connection to the Demand-Profile Formulation**

It is useful to realize that the elements of this formulation are essentially summarized by the demand profile \( N(p, q) \) used in §4. To see this, observe that an alternative type
index is the customer’s rank \( r = F(t) \), which has a uniform distribution on the unit interval. Assumption A2 implies that \( r \) indicates the customer’s rank order in terms of his purchase size for any specified price. This assumption also implies that the demand profile, constructed according to either definition

\[
N(p, q) = \# \{ t \mid v(q, t) \geq p \} = \# \{ t \mid D(p, t) \geq q \},
\]

is a decreasing function of both the price \( p \) and the purchase size \( q \). Conversely, given the demand profile, the equation \( 1 - r = N(p, q) \) can be solved for either \( p \) or \( q \) as a function of the other two parameters. If we solve for \( q \) then we obtain the demand function \( q = D(p, r) \) of the \( r \)-th ranked customer, and if we solve for \( p \) then we obtain the marginal utility \( p = v(q, r) \) indicating the customer’s willingness to pay for a \( q \)-th unit. Moreover, the customer of rank \( r \) has a utility function that can be calculated as the sum

\[
U(q, r) = \int_0^q v(x, r) \, dx
\]

of the marginal valuations of the units purchased. Alternatively, \( \int_p^\infty D(\pi, r) \, d\pi \) represents his consumer’s surplus at the uniform price \( p \), provided there are no income effects.

The condition that identifies an optimal price schedule also has an exact analogy. To make the connection between the conditions derived in Section 1 and in \S4 for the case of a profit-maximizing firm, assume again that the type index has a distribution function \( F(t) \) that need not be uniform. Using the demand profile representation, we previously derived the condition

\[
N(p, q) + N_p(p, q) \cdot [p - c] = 0
\]

to characterize the price \( p = p(q) \) on the optimal price schedule for the marginal price of the \( q \)-th unit. Alternatively, as above we can use the identity \( 1 - r = N(p, q) \) to solve for the rank \( r(p, q) \) of the type \( \tau(p, q) \) whose purchase is \( q \) when the marginal price is \( p \). The demand profile is therefore

\[
N(p, q) = 1 - r(p, q) = \tilde{F}(\tau(p, q)),
\]

where by definition \( \tau(p, q) \) satisfies \( p = v(q, \tau(p, q)) \). This implies that

\[
N_p(p, q) = -f(\tau(p, q))t_p(p, q) = -f(\tau(p, q))/v_t(q, \tau(p, q)).
\]
Substituting these relationships into the optimality condition in terms of the demand profile yields the alternative condition

$$ \tilde{F}(t) - [f(t)/v_t(q, t)][v(q, t) - c] = 0, $$

which is the same as the condition derived in Section 1 to characterize the optimal assignment of the purchase size $q(t)$ to type $t$. This first-order necessary condition for a local optimum has a unique solution for the global optimum if the left side is a decreasing function of $q$ (when searching for the optimal assignment $q(t)$) or of $t$ (if the problem is interpreted as searching for the optimal assignment $t(q)$ of purchases to types): these monotonicity conditions translate into the monotonicity conditions imposed as sufficiency conditions in §4 and §8.1.

Overall, this construction indicates that the specification in terms of customers’ one-dimensional types is essentially equivalent to the formulation in terms of the demand profile used previously, provided customers’ demand functions do not cross so that the types represent well-defined market segments.

**Customers’ Second-Order Conditions**

There is a further proviso, however. The customer’s second-order necessary condition for a local maximum requires that

$$ v_q(q, t) - p'(q) \leq 0 \text{ if } t \text{ purchases } q. $$

This says that the price schedule intersects the demand function from below. In terms of the demand profile this translates into a fairly complicated requirement:

$$ [N_{pq}/N_q - N_{pp}/N_p] \cdot [p - c] \leq 1, $$

at $p = p(q)$. But in a model with explicit types this translates into the simpler requirement that the assignment $q(t)$ of types to purchase sizes must be nondecreasing. We defer the analysis of this requirement to §8.1.

### 6.3. Two-Part Tariffs

A special case of a nonlinear tariff is a two-part tariff, so called because it comprises a uniform price for each unit purchased, plus a fixed fee payable if any positive amount is purchased. A two-part tariff is not ordinarily the optimal form of a nonlinear tariff, but if the costs of monitoring customers’ cumulative purchases are substantial then it can be advantageous.
The motive for offering a two-part tariff is often simple expediency to raise revenue by assessing an additional uniform charge against each customer. The effects can be inefficient, however, if the fixed fee excludes some customers from purchasing. That is, unlike optimal nonlinear tariffs, which in the absence of fixed costs never impose fixed fees that exclude customers, the main effect of a two-part tariff is to reduce the market penetration in exchange for higher fixed fees collected from the market remaining. A two-part tariff can be relatively efficient, as compared to using a uniform price above marginal cost to meet the same revenue requirement, because it enables the price to be set closer to marginal cost. Nevertheless, a full comparison must also take into account the reduced market penetration.

We first demonstrate the construction of an optimal two-part tariff for the case of a profit-maximizing monopolist. The firm’s profit contribution is obtained from two sources: the profit margin \( p - c \) on each unit sold, and the fee \( P \) collected from each customer electing to subscribe to the tariff. If \( t_* \) is the lowest type among the subscribers, so that \( F(t_*) \) is the number of subscribers, then these two sources provide in total the profit contribution or producer’s surplus

\[
PS = P \cdot F(t_*) + (p - c) \cdot \int_{t_*}^{\infty} D(p, t) \, dF(t).
\]

That is, each type \( t \geq t_* \) responds with the demand \( D(p, t) \) to the uniform price \( p \). The central issue is therefore to identify how the market penetration, represented by \( t_* \), is affected by the tariff. In practice this is rarely predictable with accuracy; however, here we adhere to the assumption that customers’ benefits from subscribing to the tariff are exactly described by the model. Thus, the net benefit obtained by a subscriber of type \( t \) is \( U(q, t) - pq - P \) if he purchases \( q \) units, or in terms of the consumer’s surplus, it is

\[
\int_{p}^{\infty} D(\pi, t) \, d\pi - P.
\]

The model predicts, therefore, that the marginal subscriber is that type \( t_* \) for which the consumer surplus is zero, and in particular the fee \( P \) that results in the market penetration \( t_* \) is

\[
P = \int_{p}^{\infty} D(\pi, t_*) \, d\pi.
\]

Substituting this characterization into the formula for the profit contribution then poses the firm’s maximization problem as the joint determination of the optimal uniform price.
Two-Part Tariffs

and the market penetration $t_\ast$. Let

$$D(p, t_\ast) = \int_{t_\ast}^{\infty} D(p, t) dF(t)$$

and

$$\eta(p, t_\ast) = -p \frac{D(p, t_\ast)}{D(p, t_\ast)}$$

be the aggregate demand and its corresponding price elasticity when the market penetration is $t_\ast$.

We display only the condition that determines the price, deferring the condition for $t_\ast$ to later. Expressed in terms of the price elasticity it is that the percentage profit margin should be

$$\frac{p - c}{p} = \frac{1}{\eta(p, t_\ast)} \left[ 1 - \frac{D(p, t_\ast)}{D(p, t_\ast)} \right].$$

Because $D(p, t_\ast)/F(t_\ast)$ is just the average quantity demanded, the quantity in square brackets is one minus the ratio of the minimum to the average quantity demanded.\(^3\)

In contrast to ordinary uniform pricing that sets the percentage profit margin equal to the reciprocal of the price elasticity, in the case of a two-part tariff the profit margin is reduced to take account of the effect on the market penetration. In particular, a lower price allows a deeper market penetration and thereby enables collection of the fixed fee from a larger number of subscribers. This tradeoff between the profit margin and the number of subscribers paying the fixed fee is, of course, the essence of the firm’s decision problem in designing a two-part tariff.

A similar result obtains for a regulated monopoly. The objective in this case is to maximize total surplus subject to the condition that the firm’s profit contribution is sufficient to meet a revenue requirement. The previous formula for the firm’s profit contribution identifies the producer’s surplus $PS$; the corresponding formula for the consumers’ surplus is

$$CS = \int_{t_\ast} \left[ \int_{p}^{\infty} D(\pi, t) d\pi - P \right] dF(t),$$

which sums up the types’ surplus between the demand curve and the uniform price schedule $p$, net of the fixed fee $P$. The objective for the Ramsey pricing problem is therefore to maximize $CS + [1 + \lambda]PS$, using $\lambda$ as the Lagrange multiplier for the firm’s revenue requirement and $\alpha = \lambda/[1 + \lambda]$ as the corresponding Ramsey number. For this

---

\(^3\) In more general models studied in §8.4 this ratio can exceed unity because the average demands of marginal types $t_\ast(\theta)$ can exceed average demand, in which case the marginal price is less than the marginal cost; cf. Oi (1971) and Schmalansee (1981a).
formulation the condition determining the optimal profit margin is exactly analogous:

\[
\frac{p - c}{p} = \frac{\alpha}{\eta(p, t_\star)} \left[ 1 - \frac{F(t_\star)}{D(p, t_\star)} \right].
\]

As in the case of nonlinear pricing, the effect of relaxing the revenue requirement is to allow further emphasis on efficiency considerations by reducing the profit margin associated with any particular market penetration. In turn, this expands the market and enables a reduction also in the fixed fee.

An important special case is an ordinary linear tariff with no fixed fee and a uniform price \( p \). In this case the marginal subscriber is the type \( t_\star \) for whom it is marginally worthwhile to purchase at the uniform price \( p \). Thus, \( D(p, t_\star) = 0 \) and therefore the optimality condition for the uniform price \( p \) is

\[
\frac{p - c}{p} = \frac{\alpha}{\eta(p, t_\star)},
\]

where \( \eta(p, t_\star) \) is the price elasticity of aggregate demand. This is the standard condition derived from the Ramsey formulation of uniform pricing. The various forms of multipart and nonlinear pricing merely disaggregate this condition in varying degrees.

The same results can also be derived from a formulation in terms of the demand profile \( N(p, q) \). Relying on the supposition that customers’ demand functions do not cross, we observe that the market penetration can also be characterized in terms of the minimal purchase size \( q_\star \) that a customer must anticipate before being willing to pay the fixed fee. Consequently, the aggregate demand \( \bar{D} \) is

\[
\bar{N}(p, q_\star) = \int_{q_\star}^\infty N(p, q) \, dq.
\]

Further, if \( r_\star = F(t_\star) \) represents the rank of the marginal subscriber then \( q_\star = D(p, r_\star) \), using the demand function derived from the demand profile as in Section 2. The fixed fee that produces this market penetration is calculated as before:

\[
P = \int_{p}^{\infty} D(\pi, r_\star) \, d\pi.
\]

Using these relationships in the previous construction enables the calculations to be carried out using the summary data in the demand profile.

We illustrate these methods by deriving the optimal two-part tariff for an example, addressing only the case of a profit-maximizing monopolist.
### Example 6.1: Optimal Two-Part Tariffs

<table>
<thead>
<tr>
<th>( c/A )</th>
<th>( p/A )</th>
<th>( 10 \times PB/A^2 )</th>
<th>( Pft/A^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>.00</td>
<td>.39</td>
<td>.407</td>
<td>.15</td>
</tr>
<tr>
<td>.10</td>
<td>.45</td>
<td>.330</td>
<td>.12</td>
</tr>
<tr>
<td>.20</td>
<td>.51</td>
<td>.261</td>
<td>.09</td>
</tr>
<tr>
<td>.30</td>
<td>.57</td>
<td>.200</td>
<td>.07</td>
</tr>
<tr>
<td>.40</td>
<td>.63</td>
<td>.147</td>
<td>.05</td>
</tr>
<tr>
<td>.50</td>
<td>.70</td>
<td>.102</td>
<td>.04</td>
</tr>
<tr>
<td>.60</td>
<td>.76</td>
<td>.065</td>
<td>.02</td>
</tr>
<tr>
<td>.70</td>
<td>.82</td>
<td>.037</td>
<td>.01</td>
</tr>
<tr>
<td>.80</td>
<td>.88</td>
<td>.016</td>
<td>.01</td>
</tr>
<tr>
<td>.90</td>
<td>.94</td>
<td>.004</td>
<td>.00</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

##### Example 6.1:

Suppose that the \( r \)-th ranked customer has the demand function
\[
D(p, r) = \frac{r}{2}[A - p].
\]
This demand function implies that
\[
P = \frac{r_s}{2B}[A - p]^2 \quad \text{so} \quad r_s(P) = 2B P/[A - p]^2
\]
is the market penetration depending on the fixed fee. Therefore the firm’s profit, expressed in terms of the fixed fee \( P \) and the uniform price \( p \) is
\[
\text{Profit} = \frac{1}{2}[p - c][A - p][1 - r_s(P)^2] + P[1 - r_s(P)].
\]
The two-part tariff that maximizes this profit is:
\[
p = c + \frac{1}{2}[1 - r_s][A - c], \quad \text{and} \quad P = \frac{r_s}{2B}[A - p]^2,
\]
where
\[
r_s = \frac{5}{4} - \sqrt{17/16} \approx 1 - 0.78078.
\]
Thus the optimal two part-tariff aims to induce about 78% of the potential customers to subscribe. Table 6.1 tabulates for each unit cost \( c \) the optimal unit price \( p \), the fixed charge \( P \), and the resulting profit per customer in the population. Note that a high-cost firm makes little use of the opportunity to impose a subscription fee. Characteristically, the fixed fee and the unit price vary inversely as the unit cost varies.

---

159
The Optimality Condition as an Average

To introduce the approach taken in the next section, it is useful to recast the optimality conditions for a two-part tariff. The optimality condition for the marginal price $p$ can be written in the alternative form

$$\int_{t_*}^{\infty} \left[ (p - c) \cdot D_p(p, t) + \alpha \frac{F(t)}{f(t)} \cdot D_t(p, t) \right] dF(t) = 0.$$ 

This merely states that the optimality condition derived in Section 1 for a nonlinear tariff should be satisfied on average for a two-part tariff — where the average is with respect to the distribution of types making purchases. The optimality condition for the marginal type $t_*$ can be cast similarly as

$$\int_p^{\infty} \left[ (\pi - c) \cdot D_\pi(\pi, t_*) + \alpha \frac{F(t_*)}{f(t_*)} \cdot D_t(\pi, t_*) \right] d\pi = 0,$$

which can again be interpreted as an average with respect to prices between $p$ and the largest possible price. This is basically the approach taken in §4, where the optimality condition is cast in terms of the aggregate or average demand profile over the interval of quantities for which each marginal price applies. It is also a justification for the ironing procedure used to flatten the price schedule when it would otherwise have an increasing segment.

The optimality condition for the marginal type $t_*$ has its own special uses. For instance, a flat-rate tariff consists of a fixed fee without any marginal charges for usage; thus, $p = 0$. To determine the optimal fixed fee $P$ one uses the condition above to determine the marginal type $t_*$ willing to subscribe, and then the fixed fee is that type’s consumer surplus: $P = \int_0^{\infty} D(0, t_*) \, dp$. This same approach applies also to a three-part tariff in which the initial segment is constrained to be a flat fee with a zero marginal price up to a specified maximum purchase size, beyond which a uniform price applies.

In the next section we show that this construction generalizes directly to the characterization of each segment of a multipart tariff.

6.4. Multipart Tariffs

Implementations of nonlinear pricing often offer a menu of options comprising several two-part tariffs among which each customer chooses. Based on the net benefits anticipated from the various options, each customer chooses one two-part tariff that becomes
§6.4. Multipart Tariffs

The basis for the charges billed by the seller. Such a menu mimics a single piecewise-linear $n$-part tariff comprising a fixed fee plus a block-declining price schedule with $n-1$ segments.

The formal equivalence between a multipart tariff and a menu of optional two-part tariffs is exact if the multipart tariff is concave. That is, if each customer elects the tariff that minimizes the charge billed for his actual usage then the net effect of a menu of two-part tariffs is the concave tariff that is the lower envelope of these two-part tariffs. Concavity is violated in several popular tariff designs. An especially common form of a 3-part tariff interprets the fixed fee as providing a specified free supply, beyond which a uniform price $p$ applies: this is equivalent to specifying the marginal prices $p_1 = 0$ and $p_2 = p$ for the first and second segments of the price schedule. Similarly, lifeline rates allow a low price for purchases within a small amount, but charge a higher price for further purchases. In practice, however, there is the further difference that customers are usually unable to predict exactly which optional two-part tariff will be best over the ensuing billing period. This deficiency is partially remedied if the firm bills ex post according to the least costly option.

In spite of these considerations we describe the construction of an optimal multipart tariff in terms of a menu of optional two-part tariffs. As it turns out, the characterization obtained is valid even if the associated multipart tariff is not concave. The optimality conditions are derived initially without imposing explicitly the constraint that the marginal prices must be decreasing. If the tariff must be concave to allow implementation as a menu of optional two-part tariffs, then this constraint can be invoked to obtain a modified concave tariff.

The construction follows the approach used for a two-part tariff. Assume that the menu comprises $n-1$ two-part tariffs indexed by $i = 1, \ldots, n-1$. The $i$-th tariff charges a fixed fee $P_i$ and a marginal price $p_i$. Assume these are ordered so that $P_i < P_{i+1}$ and $p_i > p_{i+1}$. Use $i = 0$ with $P_0 \equiv 0$ and $p_0 \equiv \infty$ to represent the option of not purchasing from any tariff. Because the customers’ demand functions are assumed to be ordered by their types, the set of types $t$ electing the $i$-th tariff (if any) is an interval $t_i < t < t_{i+1}$, where $t_n = \infty$. The profit contribution for the seller and the net benefit obtained by the customer from type $t$’s purchase from the $i$-th tariff are
their producer’s and consumer’s surpluses:\(^4\)

\[
\text{PS}_i(t) = [p_i - c] \cdot D(p_i, t) + P_i, \quad \text{and} \quad \text{CS}_i(t) = \int_{t_i}^{\infty} D(p, t) \, dp - P_i,
\]

and the corresponding aggregates are

\[
\text{PS} = \sum_{i=1}^{n-1} \int_{t_i}^{t_{i+1}} \text{PS}_i(t) \, dF(t), \quad \text{and} \quad \text{CS} = \sum_{i=1}^{n-1} \int_{t_i}^{t_{i+1}} \text{CS}_i(t) \, dF(t).
\]

As usual, the Ramsey formulation of the tariff design problem can be posed as the maximization of a weighted sum \( \text{CS} + [1 + \lambda] \text{PS} \) of the consumers’ and producer’s surpluses, where \( \lambda \) is a Lagrange multiplier on the constraint representing the firm’s revenue requirement and \( \alpha = \lambda/[1 + \lambda] \) is the corresponding Ramsey number. The variables in the design problem are the \( n - 1 \) pairs \( \{P_i, p_i\} \) of fixed fees and marginal prices. These parameters of the tariffs determine the corresponding segmentation of the market for which the boundaries are specified by the types \( t_i \).

The condition that type \( t_i \) is indifferent between the tariffs \( i \) and \( i - 1 \) can be expressed in terms of type \( t_i \)’s consumer’s surplus as the equality

\[
P_i - P_{i-1} = \int_{p_i}^{p_{i-1}} D(p, t_i) \, dp.
\]

That is, in considering a move from tariff \( i - 1 \) to tariff \( i \), the customer must perceive that the increment in the fixed fee will be compensated by the net value of the greater usage engendered by the lower marginal price.\(^5\) This equality implies that

\[
P_i = P_0 + \sum_{j \leq i} \int_{p_j}^{p_{j-1}} D(p, t_j) \, dp,
\]

so the formulas for the producer’s and consumers’ surpluses can be expressed entirely in terms of the marginal prices \( p_i \) and the types \( t_i \) at the boundaries between adjacent market segments.

This construction is illustrated in Figure 6.7, which shows the block-declining price schedule obtained from a menu of four two-part tariffs \( \{P_i, p_i\}, i = 1, \ldots, 4 \). Each

\(^4\) In some applications the marginal cost also depends on the tariff selected, say as \( c_j \), or on the average purchase size among those customers electing that tariff. We omit these possibilities here.

\(^5\) This equality must be phrased differently if the menu includes an initial option that offers a fixed quantity \( q_1 \) for a flat fee \( P_1 > P_2 \), or more generally a fixed fee plus a uniform price \( p_1 < p_2 \). In this case, \( P_2 - P_1 = q_1 \cdot [p_1 - p_2] \) and \( q_1 = D(p_2, t_2) \).
increment $P_i - P_{i-1}$ in the fixed fee equals the consumer's surplus obtained by the boundary type $t_i$ in moving from the marginal price $p_{i-1}$ to the lower marginal price $p_i$ for all units purchased. The fixed fee $P_4$ is therefore the sum of these increments, as represented by the entire shaded area. The figure also shows the segmentation of the market into disjoint volume bands associated with each of the two-part tariffs. The separation between two adjacent volume bands occurs because a customer's demand increases substantially in switching from one marginal price to the next lower one.

After eliminating the fixed fees from the formulation, as above, the objective function
for the Ramsey pricing problem is

$$\sum_{i=1}^{n-1} \left\{ \int_{t_i}^{t_{i+1}} \left( \int_{p_i}^{\infty} D(p, t) \, dp + [1 + \lambda][p_i - c] \cdot D(p_i, t) \right) \, dF(t) + \lambda F(t_i) \int_{p_i}^{p_{i-1}} D(p, t) \, dp \right\}. $$

Within the summation, the first term represents the consumer’s surplus for type \( t_i \); the second is the profit contribution from this type’s purchases; and the third indicates that each type exceeding \( t_i \) selects some tariff \( j \geq i \) and therefore pays the increment \( P_i - P_{i-1} \) to the fixed fee: this increment equals type \( t_i \) ’s gain from reducing the marginal price from \( p_{i-1} \) to \( p_i \), which is why he is at the boundary between these two adjacent market segments.

The variables in this version of the design problem are the \( n - 1 \) pairs \( \langle p_i, t_i \rangle \). By definition, these are constrained by the requirements that \( p_i \leq p_{i-1} \) and \( t_i \geq t_{i-1} \), where \( t_0 \) is the least type, and \( t_n \) is the highest type. In addition, \( D(p_0, t_1) = 0 \) expresses the fact that type \( t_1 \) is indifferent about purchasing. Provided these relations are satisfied, the necessary condition for an optimal choice of the marginal price \( p_i \) requires that

$$\int_{t_i}^{t_{i+1}} \left\{ [p_i - c] \cdot D_p(p_i, t) + \alpha \frac{F(t)}{f(t)} \cdot D_t(p_i, t) \right\} \, dF(t) = 0. $$

This is an exact analog of the optimality condition for a fully nonlinear tariff. The only difference is that the integral represents an average over the subpopulation of customers in the market segment selecting the \( i \)-th tariff.

Note that if the market segments and the other tariffs are fixed, then satisfying this condition entails selection of both the marginal price \( p_i \) and the corresponding fixed fee \( P_i \) that leaves the two boundaries of the \( i \)-th market segment unchanged. In other cases the market segments vary to satisfy auxiliary conditions. For example, this condition can be used to determine the optimal uniform price \( p_1 \) by using the auxiliary condition \( D(p_1, t_1) = 0 \) to determine the marginal purchaser when the fixed fee is nil, and \( t_2 = \infty \).

The necessary condition for an optimal choice of the boundary type \( t_i \) is analogous:

$$\int_{p_i}^{p_{i-1}} \left\{ [p - c] \cdot D_p(p, t_i) + \alpha \frac{F(t_i)}{f(t_i)} \cdot D_t(p, t_i) \right\} \, dp = 0. $$

This condition is again equivalent to an average, in this case over the interval of marginal prices between those charged by adjacent tariffs.
As the number \( n - 1 \) of two-part tariffs in the menu increases, the optimality conditions for \( p_i \) and \( t_i \) converge to the same condition, which is precisely the optimality condition for a fully nonlinear tariff corresponding to \( n = \infty \). However, a nonlinear tariff is costly or impractical to implement in most applications, so nonlinear tariffs are approximated by a menu with several optional two-part tariffs, a single tariff with several linear segments, or a block-declining price schedule with several steps. Fortunately, several can mean few: a menu offering just four or five two-part tariffs usually suffices to realize most of the gains from a completely nonlinear tariff. We demonstrate this feature in the following examples; in §8.3 we describe why it is true universally.

\( \diamond \) **Example 6.2**: For this example, assume that type \( t \) ’s demand function is \( D(p, t) = t[1 - p] \), the types are uniformly distributed, the firm’s costs are nil, and \( \alpha = 1 \). In this case the optimality conditions are

\[
p_i = 1 - \frac{[t_i + t_{i+1}]}{2} \quad \text{and} \quad t_i = 1 - \frac{[p_i + p_{i-1}]}{2}.
\]

These are solved subject to the boundary conditions \( t_n = 1 \), which is the maximum type, and \( p_0 = 1 \), which is the customers’ valuation of an initial unit. The solution of these conditions indicates that the \( i \)-th boundary type and the \( i \)-th marginal price are

\[
t_i = \frac{i - .5}{n - .5} \quad \text{and} \quad p_i = 1 - \frac{i}{n - .5}.
\]

The \( i \)-th fixed fee is therefore

\[
P_i = \frac{1}{3} \cdot \frac{[i - .5][i + .5]}{[n - .5]^3}.
\]

The volume band for the \( i \)-th tariff is the interval of purchase sizes \( q_i \leq q \leq Q_i \), where

\[
(q_i, Q_i) = [i/(n - .5)^2] \cdot (i - .5, i + .5),
\]

which is centered on \( \tilde{q}_i = [i/(n - .5)^2] \). Note that the volume bandwidths vary from \( 1/[n - .5]^2 \) for \( i = 1 \) to \( [n - 1]/[n - .5]^2 \) for \( i = n - 1 \), even though the sizes of the market segments are equal. The firm’s profit or producer’s surplus is

\[
PS(n) = \frac{1}{6} \cdot \left[ 1 - \frac{1}{4(n - .5)^2} \right].
\]

Thus, the profit lost from using only a few options is inversely proportional to the square of \( n - .5 \). This indicates that a menu with only a few options is sufficient to obtain most of the potential profit.
Figure 8: The optimal nonlinear, 5-part, and two-part tariffs for Example 6.2.

The equivalent piecewise-linear tariff is shown in Figure 6.8 for the case $n = 5$. The fully nonlinear tariff, corresponding to $n = \infty$, and the optimal two-part tariff, corresponding to $n = 2$, are shown for comparison; each tariff is displayed only for the range between the minimum and maximum purchase sizes it produces. Evidently the 5-part tariff is virtually the same as the nonlinear tariff for practical purposes. The two-part tariff differs significantly, and in particular it restricts both small and large purchases substantially. The firm’s profits from the 5-part and two-part tariffs are 98.8% and 88.9% of the profits from the nonlinear tariff.

Example 6.3: We reinforce these conclusions with an example in which the role of the Ramsey number can be included explicitly. The demand function is $D(p, t) = t - p$ and again the types are uniformly distributed and the marginal cost is nil. The parameters
of the optimal $n$-part tariff in this case are

$$t_i = 1 - \frac{n - i}{d(n)}, \quad p_i = \alpha \frac{n - i - .5}{d(n)},$$

$$P_i = \alpha i / d(n) - \alpha [1 + \alpha] \frac{n - .5(i + 1)}{d(n)^2},$$

where $d(n) = [1 + \alpha] [n - .5] - .5$. The price schedule is shown in Figure 6.9 for $n = 5$ and the two cases $\alpha = 0.5$ and $\alpha = 1$. The price schedules for the nonlinear tariffs corresponding to $n = \infty$ are also shown. In each case the sloping portion represents the demand function for the type $t_i$ that is indifferent between purchasing from the two adjacent segments. The firm’s profit and the consumers’ surplus are:

$$\text{PS}(n) = \frac{\alpha}{3(1 + \alpha)^2} \left\{ 1 - \frac{1}{4d(n)^3} [(1 + \alpha^3)(n - .5) - .5] \right\},$$

$$\text{CS}(n) = \frac{1}{6(1 + \alpha)} \left\{ 1 - \frac{\alpha}{4d(n)^3} [d(n)(2 + \alpha) - \alpha] \right\}.$$

For each of these, the loss from using only $n - 1$ options in the menu is again of order $1/n^2$. If $n = 2$ then compared to a nonlinear tariff the shortfalls in producer’s surplus are 5.5% or 4% for $\alpha = 0.5$ or 1, and the shortfalls in consumers’ surplus are 10% or 4%. But if $n = 5$ then for all values of $\alpha$ both producer’s and consumers’ surplus are within 1% of the corresponding amounts from a nonlinear tariff.

\[\diamond\] **Example 6.4**: For this example we use the model from Example 4.3 with the following specification of the parameters: the substitution parameter is $\alpha = 0.4$, the correlation between the type parameters is $r = 0.4$, the price of the substitute commodity is $p^* = 0.5$, the Ramsey number is $\alpha = 1$, and the marginal cost is $c = 0$. Figure 6.10 shows the optimal nonlinear price schedule as well as the three $n$-part price schedules with $n = 4, 6,$ and $11$. Each slanted segment of one of the $n$-part schedules represents the demand function of the customer of type $t_i$ who is indifferent between the $(i - 1)$-th and $i$-th two-part tariff in the menu with $n - 1$ options. The degree to which these $n$-part schedules approximate the nonlinear schedule is not fully apparent visually; in fact, all three obtain profits that are within a fraction of a percent of the profit from the nonlinear schedule. This is revealed more clearly in Table 6.2. The top panel displays the data for the 6-part schedule, using $m_i \equiv F(t_i) - F(t_{i+1})$ to indicate the fraction of potential customers who elect the $i$-th two-part tariff, and $\bar{q}_i$ to indicate the average purchase size selected by these customers. The “Averages” row shows the average fraction of customers in each volume band, and then the average among all customers served of the fixed fees, the marginal prices, and the purchase size. The bottom panel
Figure 9: Example 6.3: The price schedules for the 5-part tariff using the Ramsey numbers $\alpha = 0.5$ and $1.0$.

shows for each of the three $n$-part tariffs the market penetration (fraction of potential customers served), the profits obtained from fixed fees and marginal charges, and the total quantity sold. The fixed fees represent about a quarter of the profit, regardless of how many options are offered, and the market penetration is approximately half of the potential market of 0.8086 consisting of those customers with positive demands at a zero price. For these three $n$-part tariffs the total profits from fixed fees and marginal charges are $0.3610$, $0.3615$, and $0.3618$ respectively, which differ by at most 0.2%. Thus, the potential profit is nearly realized by offering a menu with only three two-part tariffs, provided they are designed optimally.

In §8.3 we demonstrate with considerable generality that compared to a fully nonlinear tariff, the loss from using an $n$-part tariff is approximately proportional to $1/n^2$. 


**Figure 10:** Example 6.4: The nonlinear price schedule and the multipart price schedules for $n = 4, 6, 11$. The multipart schedules approximate the nonlinear schedule and yield nearly the same profits.

### A Demand-Profile Formulation

Multipart tariffs can also be constructed from the demand profile if it is interpreted as reflecting the demand behavior of a population of customers indexed by a single type parameter. For this we interpret the price schedule as in Figure 6.9 so that $p(q) = p_i$ in the interval $q_i \leq q \leq r_i$ between the limits $q_i = D(p_i, t_i)$ and $r_i = D(p_i, t_{i+1})$ of the $i$-th volume band, and $p(q)$ is the demand price of the boundary type $t_i$ in the interval $r_{i-1} < q < q_i$.

Represent the demand function $D(p, t_i)$ of type $t_i$ as the function $q(p; p_i, q_i)$ for which $N(p, q(p; p_i, q_i)) = N(p_i, q_i)$. Note that $q(p_i; p_i, q_i) = q_i$ and $q(p_{i-1}; p_i, q_i) = r_{i-1}$. 

169
Also, the increment in the fixed fee,
\[ P_i - P_{i-1} = \int_{p_i}^{p_{i-1}} q(p; p_i, q_i) \, dp, \]
matches type \( t_i \)'s gain in moving from the marginal price \( p_{i-1} \) to \( p_i \).

Consumers' surplus is formulated as in §5. The firm's profit contribution can be formulated as the producer's surplus
\[ PS = \sum_{i=1}^{n-1} \left\{ N(p_i, q_i) \cdot \int_{p_i}^{p_{i-1}} q(p; p_i, q_i) \, dp + [p_i - c] \cdot \int_{q_i}^{r_i} q \, dN(p_i, q) \right\}, \]
expressed in terms of the demand profile. For each two-part tariff \( i \) in the summation, the first term represents the increment \( P_i - P_{i-1} \) in the fixed fee collected from the number \( N(p_i, q_i) = \tilde{F}(t_i) \) of customers among those types opting for tariffs \( j \geq i \) and the second term represents the profit margin \( p_i - c \) collected from those customers purchasing amounts in the \( i \)-th volume band, for which the integral states the average purchase size.

The previously derived condition for the optimal choice of the \( i \)-th marginal price \( p_i \) translates directly into the condition
\[ \int_{q_i}^{r_i} \left\{ a N(p_i, q) + N_p(p_i, q) \cdot [p_i - c] \right\} dq = 0, \]
expressed in terms of the demand profile. Again, and as in §4.2, this condition specifies that the optimality condition for a nonlinear tariff is satisfied on average over the volume band associated with the \( i \)-th price. Similarly, the condition for the optimal choice of the boundary type \( t_i \) translates as

\[
\int_{p_i}^{p_{i+1}} \left\{ \alpha N(p, q_i) + N_p(p, q(p; p_i, q_i)) \cdot [p - c] \right\} \, dp = 0 ,
\]

expressed in terms of the demand profile.

### 6.5. Nonlinear Tariffs

As in the formulation for \( n \)-part tariffs, we interpret the nonlinear tariff design problem as the assignment of a price \( p(t) \) to each type \( t \). In these terms, the firm’s profit from a nonlinear tariff can be written as the producer’s surplus:

\[
PS = \int_0^\infty [p(t) - c] \cdot D(p(t), t) \, dF(t) - \int_0^\infty \tilde{F}(t) \cdot D(p(t), t) \, dp(t) ,
\]

This formula is obtained from the formula for the profit from an \( n \)-part tariff by taking the limit as \( n \to \infty \), using the convention that \( p(t) \) is the limit of the price \( p_i \) for which \( t \) is the limit of the market segment boundary \( t_i \). As for an \( n \)-part tariff, \( p(t) \) is the uniform price paid by type \( t \) for its purchase \( D(p(t), t) \) and every type \( t' \geq t \) pays the increment \( D(p(t), t) \, dp(t) \) in the fixed fee to obtain the decrement \( dp(t) \) in the marginal price.

This formula can be combined with the consumer’s surplus \( W(p, t) \equiv \int_p^\infty D(p, t) \, dp \) for type \( t \) from an optimal purchase in response to the uniform price \( p \). In total, the objective function for the Ramsey pricing problem is:

\[
\int_0^\infty \left\{ \left[ W(p(t), t) + [1 + \lambda][p(t) - c] \cdot D(p(t), t) \right] f(t) - \lambda \tilde{F}(t) \cdot D(p(t), t) \, dp(t) \right\} \, dt ,
\]

where \( \lambda \) is the Lagrange multiplier on the firm’s revenue constraint. Selection of the price assignment \( p(t) \) maximizing this objective requires that the Euler necessary condition is satisfied:

\[
[p(t) - c] \cdot D_r(p(t), t) f(t) + \alpha \tilde{F}(t) \cdot D_t(p(t), t) = 0 ,
\]

where \( \alpha = \lambda/[1 + \lambda] \). This is precisely the limit, as \( n \to \infty \), of the optimality conditions for an \( n \)-part tariff.

---

\( ^6 \) This formula can be derived from the definition \( \int_0^\infty [P(q(t)) - c \cdot q(t)] \, dF(t) \) using integration by parts and a change of variables. A more complicated version is obtained when the marginal cost is not constant.
§6. SINGLE-PARAMETER DISAGGREGATED MODELS

A Direct Derivation

To derive this result directly, we simplify by assuming that the marginal cost $c$ is constant and that the marginal price schedule is nonincreasing. The quantity variable is eliminated by representing the customer’s preferences in terms of the so-called dual or indirect or utility function

$$W(p, t) = \max_q \{ U(q, t) - pq \},$$

that indicates the customer’s attainable net benefit from the uniform price $p$. This is just the consumer’s surplus under the demand function as described previously. The optimal quantity selection is specified by the demand function $D(p, t)$, assuming as previously that the utility function is strictly concave in $q$, so that the demand function is decreasing with respect to the price. In this case the derivative of the indirect utility function with respect to the price is just $W_p(p, t) = -D(p, t)$. Consequently, after finding the price assignment $p(t)$ we can infer that type $t$ purchases the quantity $q(t) \equiv D(p(t), t)$. As in previous sections, if higher types have higher demands at every price then this quantity assignment must be nondecreasing to assure that the customers’ second-order necessary conditions for an optimum are satisfied; for instance, it is amply sufficient that the price assignment is nonincreasing, which we assume hereafter.

The tariff is now construed as a one-dimensional locus of pairs $\langle P, p \rangle$ indicating a fixed fee $P$ and a marginal price $p$. This locus can be interpreted directly as a tariff by using the quantity $q$ to parameterize the locus: if $\langle P(q), p(q) \rangle$ is the $q$-th pair along the locus then the charge for a purchase of size $q$ is

$$P(q) = P(q) + p(q)q.$$  

Alternatively, assuming that the resulting charge $P(q)$ is a concave function of the quantity $q$, the tariff can be interpreted as a menu of optional two-part tariffs, in which case

$$P(q) = \min_{\langle P, p \rangle} \{ P + pq \}$$

is the minimal charge payable for a purchase of size $q$, and this is obtained by selecting the two-part tariff $\langle P(q), p(q) \rangle$.

Using this formulation, the Ramsey pricing problem can be cast as the construction of an assignment of each type $t$ to a pair $\langle P(t), p(t) \rangle$ comprising a fixed fee $P(t) \equiv P(q(t))$ and a marginal price $p(t) \equiv p(q(t))$. The consumer’s surplus for type $t$ is therefore $W(p(t), t) - P(t)$ net of the fixed fee, and the firm’s profit contribution is
6.5. Nonlinear Tariffs

This assignment is tightly constrained, however, by the customer’s freedom to self-select his preferred pair among the entire menu of options offered. As in the derivation of multipart tariffs, this constraint can be expressed by the statement that a customer is unwilling to pay a higher fixed fee unless the lower marginal price it brings provides sufficient benefits. We write this as

\[ P'(t) + D(p(t), t)p'(t) = 0, \]

indicating that for type \( t \) the slightly lower marginal price obtained by imitating a slightly higher type is exactly compensated by the slightly higher fixed fee required. Alternatively, the constraint can be stated as the requirement that the fixed fee is the accumulation of the increments paid by lower types, namely

\[ P(t) = \int_0^t D(p(s), s)[-p'(s)] \, ds, \]

using \( P(0) = 0 \) to express the fact that a customer also has the option to forgo purchasing altogether. Using this relationship to integrate by parts, the firm’s total profit contribution from fixed fees can be written as

\[ \int_0^\infty P(t) \, dF(t) = \int_0^\infty \bar{F}(t)D(p(t), t)[-p'(t)] \, dt. \]

On the right side, the integrand states that all types exceeding \( t \) elect to pay the increment \( P'(t) \) in the fixed fee in order to obtain marginal prices less than \( p(t) \).

Combining these results, the objective function for the Ramsey pricing problem is, as before,

\[ \int_0^\infty \left\{ W(p(t), t) + [1 + \lambda] \cdot [p(t) - c] \cdot D(p(t), t) \right\} f(t) - \lambda \bar{F}(t) \cdot D(p(t), t) \cdot p'(t) \, dt. \]

This objective is to be maximized by choosing the assignment \( p(t) \) from types to marginal prices. Where the optimal price assignment is nonnegative and nonincreasing, it is characterized by the Euler condition from the calculus of variations:

\[ p(t) - c = \alpha \begin{bmatrix} \bar{F}(t) \\ f(t) \end{bmatrix} \begin{bmatrix} D_t(p(t), t) \\ -D_p(p(t), t) \end{bmatrix}, \]

where \( \alpha = \lambda / [1 + \lambda] \). This version of the optimality condition is equivalent to the ones derived previously, differing only in notation. In terms of elasticities, it states as in §4 that

\[ \frac{p(q) - c(q)}{p(q)} = \frac{\alpha}{\eta(p(q), q)}, \]

173
where the imputed price elasticity of the demand profile is
\[ \eta(p, q) = \frac{\eta_p}{\eta_t} \frac{\phi_t}{\eta_t}, \]
expressed in terms of the absolute values of the price and type elasticities \( \eta_p \) and \( \eta_t \) of the demand function \( D(p, t) \), and the elasticity \( \phi_t \) of the type distribution \( F \). This derives from the identity \( N(p, q) = F(t) \), which holds when \( q = D(p, t) \), to characterize the demand \( N(p, q) \) for the \( q \)-th increment at the price \( p \). An alternative version is stated in terms of the profit margin \( \pi = p - c \) and the complement \( s = F(t) \) of type \( t \)'s rank: the ratio of the elasticities of demand with respect to \( \pi \) and \( s \) should be the Ramsey number \( \alpha \).

This version has the practical advantage that it uses directly the type-dependent demand function. In general, the problem posed in finding the optimal price schedule is one of solving two simultaneous nonlinear equations for the marginal price \( p \), and the corresponding least type \( t \) that pays this price, for each purchase size \( q \).

### 6.6. Some Examples

We now examine several examples that illustrate the methods described above.

- **Example 6.5**: We begin with examples illustrating that a uniform price is optimal in special cases.\(^7\) Consider the case that a customer of type \( t \) has the demand function \( D(p, t) = kt - \log(p) \) and the distribution of types in the population has an exponential distribution function \( F(t) = 1 - e^{-t/m} \) so that the mean of the type index is \( m \). In this case the optimal tariff charges a uniform price
\[ p = c/[1 - \alpha m], \]
when \( \alpha < 1/m \). This result is essentially a consequence of the fact that the induced price elasticity of the demand profile is the constant \( \eta = 1/m \). Similarly, if \( D(p, t) = kt^a p^{-b} \) and \( F(t) = t^{-1/\gamma} \) then \( \eta = b/a\gamma \) and \( p = c/[1 - \alpha/\eta] \) when \( \alpha < \eta \).

Other examples of optimal uniform prices occur whenever the price schedule would be increasing over its entire domain were it not flattened by the ironing procedure described previously.

---

\(^7\) Sufficient conditions for a uniform price to be optimal are given by Salant (1989), and Wilson (1988) studies a special case adapted to labor markets and to airlines’ advance-purchase airfares.
Example 6.6: An example with a linear price schedule is obtained from the demand function \( D(p, t) = A - Bp/t \) in the case that the type distribution is exponential, as in the previous example. In this case, the optimal price schedule is \( p(q) = c + \alpha m[A - q]/B \).

Alternatively, if the demand function has either of the two forms \( D(p, t) = A - Bp/t \) or \( D(p, t) = tA - Bp \), and the type distribution is uniform so that \( F(t) = t \), then again the price schedule is linear of the form \( p(q) = [1 - \beta]c + \beta [A - q]/B \), where \( \beta = \lambda/[1 + 2\lambda] \).

However, the second of these demand functions combined with an exponential type distribution yields a uniform price \( p = c + \alpha m A/B \).

Example 6.7: For this example, assume that a customer of type \( t \) has the linear demand function

\[
D(p, t) = \frac{t}{B} [A - p],
\]

corresponding to the benefit function

\[
U(q, t) = Aq - \frac{B}{2t} q^2.
\]

Assume, moreover, that the customers’ types are uniformly distributed in the population. Inserting this specification into the optimality condition indicates that the tariff should induce customers of type \( t(q) \) to purchase \( q \) units, where \( t(q) \) is the larger root of the quadratic equation

\[
t^2 [A - c] - [1 - \alpha]Bqt - \alpha Bq = 0.
\]

The price schedule for marginal units that accomplishes this is then

\[
p(q) = v(q, t(q)) = A - Bq/t(q).
\]

For instance, in the special case that no weight is given to the profit contribution, namely \( \lambda = 0 \) and \( \alpha = 0 \), this produces the uniform price \( p(q) = c \) equal to marginal cost, as one expects. In the other extreme case, all the weight is given to the profit contribution so \( \alpha = 1 \). The price schedule is then

\[
p(q) = A - \sqrt{Bq[A - c]},
\]

corresponding to the tariff

\[
P(q) = q[A - \frac{2}{3} \sqrt{Bq[A - c]}].
\]

That is, the firm offers a nominal uniform price \( p = A \) but then offers a rebate of \( \frac{2}{3} \sqrt{Bq[A - c]} \) on each unit the customer buys if his purchase size is \( q \). With this pricing
policy, the average profit contribution per customer is \( (A - c)^2 / 6B \), which is greater by a third than the firm can obtain with the optimal uniform price \( p = \frac{1}{2} [A + c] \) that ignores the heterogeneity among customers. The optimal price schedule, by the way, could also have been derived directly as in §4 by maximizing the profit contribution

\[
N(p(q), q) \cdot [p(q) - c]
\]

from the market for the \( q \)-th increment, using the demand profile

\[
N(p, q) = 1 - Bq / [A - p],
\]

which measures the fraction of the population purchasing the \( q \)-th increment at the price \( p \).

A tariff expressed in square roots is presumably impractical. However, Table 6.3 shows a menu of four two-part tariffs obtained as tangents to the optimal monopoly tariff. This menu obtains nearly the same profit contribution for the firm even though it is not the optimal menu.

\[\text{Example 6.7:} \text{ Tangential Approximation of the Optimal Tariff}\]

\[
\begin{array}{cccc}
\text{Tangent Point (} \tilde{q} \text{)} & 0 & .2 & .4 & .7 \\
\text{Fixed Charge ($)} & 0 & .0283 & .0800 & .1852 \\
\text{Price ($/unit)} & 1.000 & .576 & .400 & .206 \\
\text{Interval (units)} & 0 \leq q \leq .067 & .067 \leq q \leq .294 & .294 \leq q \leq .543 & .543 \leq q \leq .794 \\
\end{array}
\]

\[\text{Example 6.8:} \text{ For this example assume that customers are alike except that a customer of type } t \text{ incurs a transport cost of } $t \text{ per unit purchased. In particular, the demand function of a customer of type } t \text{ is linear of the form}\]

\[
D(p, t) = a - b[p + t],
\]

and the fraction of customers having transport costs less than \( t \) is \( F(t) = t^k \). For instance, if the firm has a plant at a central location in a region over which customers
are uniformly distributed, then typically \( k \approx 2 \). Since the demand function decreases as \( t \) increases (rather than increases as assumed previously) we use \(-F\) rather than \( F\) in the optimality condition. Using this specification in the optimality condition indicates that the price schedule should be designed to induce the customer of type

\[
t(q) = \frac{k}{k + \alpha} \left[ \frac{a - q}{b} - c \right]
\]

to purchase \( q \) units. The price schedule for marginal units is therefore

\[
p(q) = \frac{k}{k + \alpha} c + \frac{\alpha}{k + \alpha} \left[ \frac{a - q}{b} \right],
\]

and the tariff is

\[
P(q) = q \left\{ \frac{k}{k + \alpha} c + \frac{\alpha}{k + \alpha} \left[ \frac{a - \frac{1}{2}q}{b} \right] \right\}.
\]

As in the previous example, we see that if \( \alpha = 0 \) then a uniform price equal to marginal cost is used, and otherwise a higher nominal price is offered with quantity discounts.

The incidence of charges in excess of actual cost is

\[
P(q(t)) - cq(t) = \frac{\alpha}{2b} \left[ \frac{(a - bc)^2}{k + \alpha} - (k + \alpha)(b/k)^2 t^2 \right],
\]

where \( q(t) \) is type \( t \)'s purchase size. Thus customers' contributions to the firm's revenue requirements are derived mostly from those types with low transport costs, and they decrease quadratically as the transport cost increases. This feature is fairly general; namely, customers with advantages in using the product tend to bear the greater burden of meeting the firm's revenue requirements. It partly accounts for the acceptance of nonlinear pricing by regulatory agencies.

However, the fact that nonlinear pricing prescribes a particular pattern of incidence can result in its modification or rejection if this pattern is unacceptable. For instance, in the case of transport costs, a popular alternative is for the firm to absorb some of the costs of transport in order to increase the market penetration to distant customers. Considering only a profit-maximizing monopolist (that is, \( \alpha = 1 \) and \( k = 2 \), suppose that the firm offers a uniform price \( p(t) \) that depends on the delivery cost \( t \), which in this case is incurred entirely by the firm. For the example above, the optimal price is

\[
p(t) = \frac{a}{2b} + \frac{1}{2} [c + t],
\]
indicating that the firm charges the usual monopoly price and absorbs half of the transport cost. In this case the customer of type $t$ contributes

$$[p(t) - c - t]q(t) = \frac{1}{b} \left[(a/2b) - \frac{1}{2}(c + t)\right]^2$$

to the firm’s profit, which is again quadratic but the incidence is quite different than previously, as is the distribution of net benefits among customers. Figure 6.11 shows the price schedule $p(t)$ for the case $a/b = 1$ and $c = 0.1$. A significant feature is that the market penetration is much deeper with optimal freight absorption: with the optimal schedule customers as far away as $t = 0.9$ make positive purchases, where as shown in the figure, FOB pricing stops at $t = 0.6$ and a uniform delivered price also stops at $t = 0.6$ because the firm is unwilling to serve customers farther away.

Nonlinear pricing applies equally well to contexts in which customers’ preferences

\[178\]
are derived from other considerations. An important instance is a manufacturer’s choice of wholesale prices to retailers who in turn select their own retail prices for resale. Franchising is similar: the franchiser shares in some of the profits of its higher-volume franchises by pricing the franchised product nonlinearly. We illustrate below with an example.

\begin{example}
Consider a manufacturer that sells its product to many retail stores. Suppose the stores are sufficiently separated spatially that their retail markets can be considered independent. In particular, each store chooses its retail price based on the manufacturer’s wholesale price and the store’s local demand conditions. Recognizing this, the manufacturer can obtain advantages from a nonlinear price schedule that offers quantity discounts to high-volume retailers. To represent the simplest example, suppose that stores are classified into various types indexed by a single parameter \( t \), and that the demand function of a store of type \( t \) is \( D(p, t) = t - p \), measured in, say, units per month. That is, local demand conditions are assumed to raise or lower the demand function uniformly. The optimal retail price at such a store is

\[ p_t = \frac{1}{2} \left[ t + c_s + p(q_t) \right], \]

where \( c_s \) is the marginal cost at the store and \( p(q_t) \) is the manufacturer’s marginal price for the last unit \( q_t \) bought by the store each month. The monthly quantity sold by the store is then predicted to be

\[ q_t = \frac{1}{2} \left[ t - c_s - p(q_t) \right]. \]

Let \( c_m \) be the manufacturer’s marginal cost, and consider now the manufacturer’s pricing problem. If the types of the stores are uniformly distributed, then the manufacturer’s problem can be posed as follows. Let \( t(q) \) indicate the type of store that buys \( q \) units. Then a fraction \( 1 - t(q) \) of the stores will buy a \( q \)-th unit, and on these sales of \( q \)-th units the manufacturer obtains a profit margin \( p(q) - c_m \). Assuming that the stores follow the optimal retail pricing strategy derived above, this profit margin can be alternatively formulated as

\[ t(q) - c_s - c_m - 2q, \]

and similarly the gross profit on this market segment is

\[ \left[ 1 - t(q) \right] \left[ t(q) - c_s - c_m - 2q \right]. \]

Choosing the corresponding type \( t(q) \) optimally to maximize the gross profit, the manufacturer prefers that \( t(q) = \frac{1}{2} \left[ 1 + c_s + c_m + 2q \right] \). The marginal price that achieves this result is

\[ p(q) = t(q) - c_s - 2q = \frac{1}{2} \left[ 1 - c_s + c_m \right] - q. \]

The corresponding total charge for \( q \) units is therefore

\[ P(q) = \frac{1}{2} q \left[ 1 - c_s + c_m \right] - \frac{1}{2} q^2. \]
Consequently, the predicted quantity sold to a store of type \( t \) is

\[
q_t = t - \frac{1}{2} \left[ 1 + c_s + c_m \right],
\]

which is positive only if the type \( t \) is sufficiently large; that is, the manufacturer prefers to forego sales to the stores with small markets. Using this pricing policy, the manufacturer’s gross profit in total (normalized by the number of retailers) is \( \frac{1}{24} \left[ 1 - c_s - c_m \right]^3 \).

If both marginal costs are zero then this is \( \approx .0417 \); in contrast, using a single uniform price the manufacturer would charge \( p = \left[ 8 - \sqrt{7} \right] /9 \approx .301 \) and obtain the gross profit \( \approx .0351 \) — thus, the nonlinear pricing policy yields 19% greater profits. In practice it would suffice for the manufacturer to offer only three or four price breaks to achieve most of this potential gain in profits.

6.7. Fixed Costs and Fixed Fees

The optimal tariff includes a fixed fee whenever the firm incurs a fixed cost of serving an individual customer. The construction of this fixed fee is exactly analogous to the construction of the optimal price schedule. We show, however, that if the firm’s fixed cost is nil then so too is the fixed charge when a fully nonlinear price schedule is used. Multipart tariffs typically include a positive fixed fee for the initial segment but it is small if several options are offered: it is positive only because of the limitation on the number of options.

As in previous sections, customers’ types are described by a one-dimensional index \( t \). The number or fraction of customers having types exceeding \( t \) is \( \hat{F}(t) \). Type \( t \) has the benefit function \( U(q, t) \) for a purchase of size \( q \) and the marginal benefit function \( v(q, t) = U_q(q, t) \) for a \( q \)-th unit, both of which increase as the type index increases, and \( v \) declines as \( q \) increases. Alternatively, the benefit is the area under the demand curve \( D(p, t) \). The firm’s total and marginal cost functions are \( C(q) \) and \( c(q) \) to supply an individual customer. A fixed cost component is interpreted as \( C(0) \), with the understanding that this is incurred only if the customer makes a positive purchase.

The tariff \( P(q) \) and its marginal price schedule \( p(q) \) are specified similarly, again with the proviso that the fixed fee \( P(0) \) is paid only if the customer makes a positive purchase. In effect, demand for access is equated with demand to purchase. The tariff therefore consists of two ingredients:

- A minimum price \( P_* = P(q_*) \) is charged for a minimum purchase \( q_* \). This purchase or some larger one is bought by all types exceeding a minimum type \( t_* \),
making a purchase.

- A marginal price schedule \( p(q) \) imposes charges for increments \( q > q_s \). These increments are bought by types \( t > t_s \).

Thus, the minimum quantity \( q_s \) is sold as a block for the minimum charge \( P_s \). Alternatively, the firm can extend the price schedule to increments \( q < q_s \) and charge a fixed fee \( P_o \), so that the minimum charge is

\[
P_s = P_o + \int_0^{q_s} p(x) \, dx.
\]

and the total tariff is

\[
P(q) = P_s + \int_0^{q} p(x) \, dx.
\]

We use a somewhat different formulation of the Ramsey pricing problem in this section in order to address the determination of the fixed fee most simply. The firm’s profit contribution or producer’s surplus can be formulated in terms of the assignment \( q(t) \) specifying each type’s purchase:

\[
PS = \int_0^{\infty} [P(q(t)) - C(q(t))] \, dF(t).
\]

Integration by parts provides an equivalent formula:

\[
PS = \hat{F}(t_s) \cdot [P_s - C(q_s)] + \int_{t_s}^{\infty} \hat{F}(t(q)) \cdot [p(q(t)) - c(q(t))] \, dq(t),
\]

where again \( q_s = q(t_s) \) is the least purchase made, \( t_s \) is the type making this purchase, and \( P_s = P(q_s) \) is the tariff charged for the purchase \( q_s \). In this formula, the first term indicates that the measure of those types \( t \geq t_s \) making positive purchases is \( \hat{F}(t_s) \).

From each of these types the firm’s profit contribution from the minimum purchase \( q_s \) is \( P_s - C(q_s) \), namely the tariff net of the firm’s total cost of supplying this purchase. In effect, \( F(t_s) \) is the demand for access at the price \( P_s \) when it includes an allowance \( q_s \).

We make substitutions in this formula to take account of customers’ purchase behaviors. The first is that \( v(q(t), t) = p(q(t)) \) for those types \( t \geq t_s \), indicating that \( q(t) \) is on \( t \)’s demand curve at its assigned marginal price. The second indicates that the least type \( t_s \) to purchase is the one for whom the net benefit \( U(q_s, t_s) - P_s \) is nil; so
we substitute \( U(q_s, t_s) = P_s \) in the formula. These substitutions can also be made in the formula for the consumers’ surplus, so that integration by parts yields:

\[
CS = \int_{t_s}^{\infty} [U(q(t), t) - P(q(t))] dF(t) \\
= \int_{t_s}^{\infty} \frac{F(t)}{f(t)} U_q(q(t), t) \, dt.
\]

The Ramsey formulation of the pricing problem is again to maximize consumers’ surplus subject to a revenue constraint for the firm, say \( CS + [1 + \lambda]PS \) where \( \lambda \) is the multiplier on the firm’s revenue constraint and \( \alpha = \lambda/[1 + \lambda] \) is the corresponding Ramsey number. An analysis of the optimal schedule of marginal prices yields exactly the same condition as derived previously in Sections 1 and 5. It remains, therefore, to determine the minimal purchase \( q_s \), the minimal charge \( P_s \), or the least type \( t_s \) served. Using the formulas above, the necessary condition for an optimal choice of any one of these is:

\[
U(q_s, t_s) - C(q_s) - \alpha \frac{F(t_s)}{f(t_s)} U_t(q_s, t_s) = 0.
\]

This optimality condition parallels exactly the optimality condition for the purchase \( q(t) \) assigned each type \( t > t_s \). Using \( q_s = q(t_s) \) it provides an equation that determines \( t_s \), and thereby also \( P_s = U(q(t_s), t_s) \).

An alternative construction uses the auxiliary demand profile \( M(P, q) = \# \{ t \mid U(q, t) \geq P \} \). The necessary condition above is essentially equivalent to selecting the minimal charge \( P_s \) to maximize the contribution

\[
\int_{P_s}^{\infty} M(P, q_s) \, dP + [1 + \lambda] \cdot M(P_s, q_s) \cdot [P_s - C(q_s)],
\]

to consumers’ and producer’s surplus from customers purchasing the minimal quantity \( q_s \), for which the necessary condition is

\[
\alpha M(P_s, q_s) + M_P(P_s, q_s) \cdot [P_s - C(q_s)] = 0.
\]

This condition is applied by interpreting its solution \( P_s(q_s) \) as a function of the minimal purchase \( q_s \) and then determining \( q_s \) and \( t_s \) by the requirement that

\[
M(P_s(q_s), q_s) = N(p(q_s), q_s) = \frac{F(t_s)}{}.
\]

This construction suffices under the assumptions used in this chapter, but more general formulations require methods such as those in §4.4 and §8.5.
The interpretations are also analogous. For example, the percentage profit margin on the minimal purchase \( q_\ast \) can be interpreted as inversely proportional to the price elasticity of the demand for this initial block. Moreover, as in the analysis of multipart tariffs, this condition can be interpreted simply as an average of the corresponding condition for a nonlinear price schedule, in this case averaged over the units \( q \leq q_\ast \) purchased by the single type \( t_\ast \) purchasing these units:

\[
\int_0^{q_\ast} \left\{ v(q, t_\ast) - c(q) - \alpha \frac{F(t_\ast)}{f(t_\ast)} v_t(q, t_\ast) \right\} dq = C(0).
\]

An important corollary of this condition is that the fixed fee is nil if the fixed cost is nil. To see this, observe that the benefit from a purchase size of zero is nil independently of the customer’s type; that is, \( U(0, t) = 0 \) and therefore \( U_t(0, t) = 0 \). Consequently, if \( C(0) = 0 \) then the condition is satisfied by the minimal purchase \( q_\ast = 0 \). In this case, \( t_\ast \) is merely the least type willing to make a purchase from the price schedule, namely \( q(t_\ast) = 0 \). This result evidently depends on the assumption that access is not valued apart from the opportunity to purchase; it is presumably false in the case of telephone service used only to receive calls.

When the fixed cost of service is positive, it is useful to apply the optimality condition to obtain the formulas

\[
P_\ast = U(q_\ast, t_\ast) = C(q_\ast) + \alpha \frac{F(t_\ast)}{f(t_\ast)} U_t(q_\ast, t_\ast)
\]

\[
= C(0) + \int_0^{q_\ast} \left[ c(q) + \alpha \frac{F(t_\ast)}{f(t_\ast)} v_t(q, t_\ast) \right] dq
\]

\[
= C(0) + p_\ast q_\ast + \int_0^{q_\ast} q[c'(q) + \alpha \frac{F(t_\ast)}{f(t_\ast)} v_{tt}(q, t_\ast)] dq,
\]

where \( p_\ast = p(q_\ast) \) and the third line is obtained from the second using integration by parts. The third line indicates that if marginal cost is constant and \( v_{tt} = 0 \) then the minimal charge is \( P_\ast = C(0) + p_\ast q_\ast \), which is merely a two-part tariff with a fixed fee equal to the fixed cost and a marginal price equal to one charged on the price schedule for the \( q_\ast \)-th unit. Only if marginal cost or the slopes of customers’ demand functions vary significantly will the minimum charge deviate much from this approximation. The following example illustrates this feature.

\[\diamond \textbf{Example 6.10} : \] For this example, suppose that \( \alpha = 1 \) and

\[
U(q, t) = ta q - \frac{1}{2} bg^2, \quad C(q) = C_0 + cq, \quad F(t) = t,
\]
where \( C_0 \) is the fixed cost of serving a customer. Then the optimal price schedule is \( p(q) = \frac{\sqrt{a + c - bq}}{2} \) and type \( t \) purchases \( q(t) = \frac{a + c - 2t}{b} \). Applying the optimality condition for the minimal purchase yields

\[
q_* = \sqrt{2C_0 / b}, \quad \text{and} \quad P_* = C_0 + p_* q_*.
\]

Thus, the minimum charge \( P_* \) consists of a fixed fee \( P_0 = C_0 \) equal to the fixed cost, plus a uniform price \( p_* \equiv p(q_*) \) for those units in the minimal purchase \( q_* \). The firm’s profit margin on the minimal purchase can therefore be interpreted as independent of the fixed cost per se. Rather, it derives entirely from the uniform price charged for the minimal purchase:

\[
P(q_*) - C(q_*) = [p_* - c] \cdot q_*.
\]

The optimal price schedule can also be interpreted as the minimum of the price schedules for a two-part tariff that charges the fixed cost plus the uniform price \( p_* \) per unit, and the nonlinear tariff that charges the schedule \( p(q) \) for incremental units; that is, marginal prices above \( p_* \) are excluded.

Figure 6.12 shows the optimal price schedule constructed in this way. It also shows that the fixed fee in the amount of the fixed cost is exactly equal to the consumer’s surplus for the least type \( t_* \), calculated using the uniform price \( p_* \) for the first \( q_* \) units. From another viewpoint, the price schedule can be interpreted as the charging the maximum of \( v(q, t_*) \) and \( p(q) \) for each unit \( q \geq 0 \).

### 6.8. Summary

In this chapter we replace the demand profile as a summary of customers’ behavior with a disaggregated formulation in which customers’ benefit or demand functions are specified explicitly. The model relies on a one-dimensional parameter to identify customers’ types or market segments. It is therefore limited by the maintained assumption that demand functions of different types of customers do not intersect.

We consider models in which the types are discrete and also models with a continuum of types. The latter are used to characterize the construction of multipart and nonlinear tariffs. In each case the same basic optimality condition recurs, though the form differs depending on the model used. This optimality condition is essentially the same as the one obtained from the demand-profile formulation. The optimality conditions for multipart tariffs are simply averages of the corresponding condition for a nonlinear tariff.
This same interpretation applies to the determination of a fixed fee, although such a fee is positive only if the firm incurs a fixed cost in providing access to service.

**Figure 12:** The optimal price schedule with a fixed fee $P_0 = C(0)$.

---

**SUPPLEMENT**
Chapter 7

OTHER APPLICATIONS OF NONLINEAR PRICING

Product pricing is one of several applications of the basic principles of nonlinear pricing. This chapter reviews briefly a few other applications, some of which reveal additional motivations for using nonlinear schedules. These applications have the common feature that the population is diverse, reflecting private information about preferences or technology. The tariff, price schedule, or other incentive scheme promotes efficiency partly by eliciting this information indirectly as revealed by the behavior it induces. The principles of Ramsey pricing provide a unifying framework to derive characterizations of efficient incentive schemes.

Section 1 summarizes additional contexts that use nonlinear pricing of products and quality-differentiated services. Section 2 describes three of many contexts in which nonlinear schedules are used to cope with incentive effects in insurance markets and other situations affected by private information. Section 3 reviews a portion of the literature on contracting, especially in labor markets where incentive effects are prominent. Section 4 describes applications to the design of taxation and regulatory policies. We omit technical details except for brief summaries in subsections indicated with an asterisk (*). Although terse, these formulations indicate how the principles of Ramsey pricing are applied to other contexts with incentive effects stemming from private information.

Good texts on the topics in this chapter are the two books and the survey article by Phlips (1983, 1988a, 1988b), and the four related surveys by Baron, Braeutigam, Varian, and Stiglitz in the Handbook of Industrial Organization (1989).

7.1. Product Pricing Applications

Previous chapters address mainly the case that a firm offers an explicit tariff to a large population of customers. In contrast, the tariff is often implicit in the applications in this section.
7.1. Product Pricing Applications

Product Lines

A firm offering a product line usually organizes it along a one-dimensional scale representing the magnitude of a prominent attribute. In the simplest case the products are identical except for differing package sizes. The schedule of package prices is essentially a tariff; and, the price increments charged for size increments constitute the marginal price schedule. The same is true if the products differ along quality dimensions. If customers buy multiple units of various qualities, then joint pricing of quality and quantity is advantageous. The analysis in §10.2 of Ramsey pricing of priority service is an example.

In important cases, qualities are measured as output rates, and from a customer’s viewpoint they are like quantities. For instance, a faster computer or copier is a substitute for several slower machines, and is therefore like a compact package. Nevertheless, product lines of this sort have other motivations. Design considerations may enable the firm to enhance the operating rate, durability, or maintenance costs of a machine with a less than proportionate increase in fabrication costs. For a customer, therefore, choices among machines involve tradeoffs between investment costs and operating costs. If a customer has a variable demand for usage, namely a peakload problem, then it may be preferable to purchase an expensive (large or fast) machine to serve baseloads, as well as a cheaper machine to serve peak loads. Discounts for multiple purchases from a product line are also used by some firms when customers incur costs of switching from one vendor to another; switching costs are potentially an important source of monopoly power.

Delivery Conditions

Some quality dimensions occur pervasively: size, output rate, reliability, durability, speed, precision, ease of maintenance, operators’ skill levels, et cetera. Indeed, several of these have small literatures applying principles of nonlinear pricing. Conditions of delivery are especially common and we mention two of the most important: place and time of delivery.

Spatial Pricing

Firms use many different schemes to recoup delivery costs and to differentiate products via place or cost of delivery. As mentioned in §6.6, the two extreme forms are FOB pricing, in which the customer takes delivery at the firm’s location or pays for shipment costs; and delivered pricing, in which the firm absorbs all delivery costs. An analysis
based on nonlinear pricing generally indicates that an optimal scheme is intermediate between these two extremes. Suppose for instance that the populations of customers at different distances from the firm’s plant are the same whereas the cost of serving customers increases with distance. Then generally it is optimal for the firm to absorb some delivery costs. Partial absorption of transport costs is implemented in many ways, including warehousing, basing points, and direct credits to customers.\footnote{Although it is widely used elsewhere, basing-point pricing in oligopolistic markets violates antitrust laws in the United States.} When firms enjoy local monopolies, but compete near the boundaries of their service areas, this can produce strong effects; in particular, profit margins can be driven to zero at the boundaries.\footnote{See Anderson and Thisse (1988), Gabszewicz and Thisse (1990), and Spulber (1984) for more elaborate analyses of spatially differentiated pricing in this vein.} In an oligopoly usually neither FOB pricing nor full absorption of delivery costs is a stable pattern (some firm has an incentive to switch to the other mode) but partial absorption is stable.

*Temporal Pricing*

Nondurable products produced continually rarely offer significant opportunities for differentiation by time of delivery. Time is important for some durable products because unserved demand accumulates. An example is household goods, such as sheets and towels, which wear out steadily but can be replaced opportunistically by taking advantage of periodic sales. The deterioration of existing stocks and formation of new households produce a steady influx of potential demand, part of which is unserved until the backlog grows so large that the firm finds it advantageous to tap this market with a lower sale price. Stores selling such items typically use a recurring pattern in which periods of high prices are followed by sales at low prices. Customers with urgent needs buy immediately at whatever price prevails, whereas customers who can defer or store inexpensively prefer to wait until a sale. The price schedule over a cycle of price variation represents a form of nonlinear pricing in which the quality attribute is the time of delivery and customers differ in terms of their patience to wait for a sale or their valuations of the items.\footnote{See Conlisk, Gerstner, and Sobel (1984) and Sobel (1991). In an oligopoly the timing of sales must be random, since each firm has an incentive to preempt the others. There is also a literature on firms’ use of multiple prices at a single time, a practice called price dispersion, that we do not address here; cf. Stiglitz (1989) and Varián (1989). Such models require nonconvexities in their formulations and often imply randomization of prices. However, Charles Wilson (1988) shows that, for a monopolist firm using uniform}
Even when demand is not recurring, the seller of a durable good generally prefers to differentiate prices according to the time of delivery. When there is no limitation on the supply and storage costs are positive or customers are impatient, this practice is an instance of inefficient quality distortion in the form of delayed delivery, of the sort described in §13, Figure 13.2. Typical examples are new books, films, and computer programs for which some customers are willing to pay more for earlier delivery. The prices of these products typically decline over time. For instance, hardcover editions of books are offered at higher prices initially than later soft cover editions; films in theaters are more expensive initially than later on videocassettes, and still later on television.

The rate at which prices decline depends greatly on whether the firm can commit in advance to its price schedule. If the firm cannot commit then, after serving a high-valuation or impatient segment of the market, the firm perceives advantages from cutting its price beyond what it previously would have wanted. Indeed, anticipating this, the high-valuation customers are reluctant to pay as high a price initially because the advantage of waiting for the next, lower price is greater. An extreme form of this phenomenon is the Coase property: if the firm’s marginal cost is constant and the firm can change its price rapidly, then the optimal price schedule is little more than marginal cost and most sales are made immediately.4 A mild version of the Coase property is seen in the following example, where the initial price of $0.232 is less than half the monopoly price the firm would use in every period if it could commit to its price schedule. Further, if the firm changes its price weekly rather than monthly, then the initial price is $0.142 and all customers are served after nine weeks, rather than eight months.

Example 7.1: Suppose customers differ in their valuations of the product, but they are equally impatient. In particular, a customer whose valuation is $v$ obtains the net benefit $[v - p_k]\delta^k$ if he purchases at the price $p_k$ in month $k$, and zero if he never purchases. The firm’s objective in choosing the sequence $p_k$ of prices is to maximize the present value of its revenues, using the same discount factor $\delta$. Table 7.1 shows the firm’s optimal monthly price schedule for the case that there are ten types of customers, prices, one or two prices are optimal if customers arrive randomly and buy at the lowest price for which supply remains, the aggregate demand curve is downward sloping, and marginal costs are increasing or supply is limited. Two prices are optimal when the optimal or limited supply occurs where the firm’s marginal revenue curve is increasing.

---

each comprising 10% of the population, whose valuations are $0.10, $0.20, \ldots, $1.00; the firm’s marginal cost is zero, and the discount factor is $\delta = 0.95$. Anticipating this sequence of declining prices, each customer waits to purchase until the price is less than his valuation, and then waits further until he prefers to purchase immediately rather than wait for the next, lower price.\footnote{The fraction of customers remaining (those who have not yet purchased) in some periods is not a multiple of 0.10 because the previous price made customers at the margin indifferent whether to purchase that month or wait until the next. The fraction predicted to accept is determined by customers’ responses to deviant prices that might be offered by the firm. See Gul, Sonnenschein, and Wilson (1986) for details.}

\begin{table}[ht]
\centering
\begin{tabular}{llll}
\hline
Period $k$ & Fraction of Customers Remaining $\frac{1}{10}$ & Price $p_k$ & Fraction of those Remaining who Purchase $\frac{1}{2}$ \\
\hline
1 & 1.000 & $\$.232$ & .199 \\
2 & .801 & .197 & .182 \\
3 & .655 & .170 & .196 \\
4 & .527 & .148 & .221 \\
5 & .410 & .129 & .263 \\
6 & .303 & .115 & .339 \\
7 & .200 & .105 & .500 \\
8 & .100 & .100 & 1.000 \\
\hline
\end{tabular}
\caption{Price Schedule for Example 7.1}
\end{table}

Firms adopt various tactics to avoid the implications of the Coase property. The problem is severe in the case of durable capital equipment. The Coase property takes the form that the manufacturer is in effect selling items now with which its later output will compete via customers’ opportunities to purchase in secondary resale or rental markets. This is another version of the requirement for fully optimal nonlinear pricing that resale markets must be excluded. To alleviate this problem the firm can reduce the durability of the product, which is usually inefficient, or control the resale and rental markets by only leasing the product.\footnote{Bulow (1986) gives an example of optimal inefficient durability. Since IBM and
commitments to limited capacity or a production technology that has increasing marginal costs.\footnote{These strategies are analyzed by Stokey (1981) and Kahn (1986), respectively.}

Temporal differentiation via nonlinear pricing applies also to the design of advance-purchase discounts. Most airlines offer such discounts: §2.6 describes the schedule offered by Delta Airlines. The diversity among customers that is important in designing such discounts concerns the expected cost of committing early to travel plans. In addition, firms schedule crews and equipment more efficiently when demand uncertainties are reduced in advance, so the firm’s marginal cost is also affected. Gale (1992) shows that advance-purchase discounts can be an optimal way to induce those customers who can schedule their trips flexibly to avoid congested peak periods when capacity is scarce.

*An Illustrative Formulation*

We sketch a simplified formulation of the design of an airline’s advance purchase fares adapted to the item-assignment model in §9.4. Customers’ demands are interpreted as items requiring service in the form of trips requiring flights. Ticket purchases farther in advance of departure are interpreted as lower quality because there is a greater chance that intervening events will force cancellation.

Tickets are neither refundable nor transferable. To avoid complications, we make the extreme assumption that each customer buys tickets repeatedly until he is able to complete a single trip. That is, demand for a trip persists even if events force cancellation of a reservation. Moreover, a customer has ample time to reschedule the trip: there are no time constraints on the next reservation. We assume the firm has a monopoly on the route and we ignore capacity constraints.\footnote{An oligopoly can be addressed using the methods in §12. The demand persistence assumption is used here to obtain the simplification that a customer’s benefit from a completed trip has no effect on his choice of the optimal time to purchase a ticket.}

Each customer is described by a pair \((v, t)\) in which \(v\) is the value of a completed trip and \(t\) is a parameter affecting the probability \(r(s, t)\) that he will be able to make the flight if he buys a ticket the duration \(s\) beforehand. That is, \(1 - r(s, t)\) is the probability that some event in the intervening \(s\) days will force postponement of his travel plans. Usually one specifies that \(t\) is the mean arrival rate of events that would force cancellation of a reservation; that is, \(t\) tends to be higher for commercial travelers and therefore \(r(s, t)\) is lower, and by definition \(r(s, t)\) is lower if \(s\) is higher, so one

---

\footnote{Xerox were enjoined from leasing but refusing to sell, this practice has been construed as a violation of the United States’ antitrust laws.}
assumes that \( r \) is a decreasing function of both parameters. If the price of a ticket purchased \( s \) days in advance is \( p(s) \) then his expected net gain can be stated recursively as

\[
V = \max_{s \geq 0} \{ r(s, t)v + [1 - r(s, t)]V - p(s) \}.
\]

This says that if the present reservation is cancelled then the process repeats until the trip is completed. Equivalently, for the optimal choice of \( s \), \( V = v - p(s)/r(s, t) \), indicating that the expected sum of the prices paid for all the tickets purchased until the trip is completed is \( p(s)/r(s, t) \). Only those customers for whom \( V \) is positive will purchase any ticket, but assuming a ticket is bought then the condition characterizing the optimal duration is

\[
r_s(s, t)[v - V] - p'(s) = 0,
\]

or

\[
r_s(s, t)/r(s, t) = p'(s)/p(s),
\]

independently of \( v \).

To complete the formulation, suppose that the density of those customers with parameter \( t \) and \( v \geq P \) is \( D(P, t) \). This is the demand function for refundable tickets at the price \( P \) among those customers with parameter \( t \). This demand is the arrival rate of customers with newly acquired motives for an eventual trip. To apply these specifications to the present case, observe that if it is type \( t(s) \) who purchases \( s \) days in advance, then the arrival rate of new customers for such tickets is \( D(p(s)/r(s, t(s)), t(s)) \cdot |t'(s)| \) and to each such customer the firm sells on average \( 1/r(s, t(s)) \) tickets at the price \( p(s) \) before his trip is completed. (The absolute value is used because presumably \( t(s) \) is a declining function.) The firm’s expected profit contribution, also interpreted as an average rate per unit time, from a fare schedule \( p \) is therefore

\[
\int_0^\infty [P(s) - c] \cdot D(P(s), t(s))|t'(s)| \, ds,
\]

where \( P(s) = p(s)/r(s, t(s)) \) is the average total revenue obtained from each customer purchasing at time \( s \) beforehand and \( c \) is the marginal cost of the single trip provided. Thus the two functions \( P(s) \) and \( t(s) \) are chosen to maximize this expected profit contribution subject to the customer’s optimality constraint. This constraint can be restated in terms of \( P \) as:

\[
P'(s) = P(s)q(s, t(s))t'(s), \quad \text{where} \quad q(s, t) = -r_t(s, t)/r(s, t).
\]

Having found \( P(s) \) and \( t(s) \), the fare schedule is obtained via \( p(s) = P(s)r(s, t(s)) \).
The optimal schedule can be characterized using the Euler condition from the calculus of variations in a fashion similar to the analysis of the item-assignment formulation in §9.4. For the practical applications, however, one considers only block-declining fare schedules that are piecewise constant, say differing fares for purchases within a few days, a week, two weeks, and a month of departure. Extensions to Ramsey pricing must include the costs imposed on customers; in particular, customers may incur substantial costs adhering to the original itinerary to avoid the high price of interrupting a roundtrip reservation midway in the trip.

**Price Discrimination and Inefficient Quality Degradation**

Some of the topics in this section are versions of price discrimination. The inefficiencies and distributional consequences they engender cast a shadow on the benefits of non-linear pricing in other contexts where it enhances efficiency. As we emphasize in §5 and §10, careful application of nonlinear pricing can enhance efficiency without disadvantaging any customer. But versions of nonlinear pricing can also be used to exploit monopoly power inefficiently and with significant welfare consequences. Airlines’ advance purchase discounts are possibly an example. These discounts are implemented by offering lower prices for tickets purchased in advance, but usually the quality is degraded by making the tickets nontransferable and partially or entirely nonrefundable; this is deemed necessary to make them unattractive to customers who highly value flexibility. The lower prices promote efficiency by enabling more customers to travel, but against this gain must be set the losses from quality degradation. Nontransferability produces empty seats when customers cancel trips, and nonrefundability causes costly consequences for travelers enroute who cannot change their itineraries by paying the actual cost imposed on the airline. The fact that firms with monopoly power often use inefficient quality degradation to implement nonlinear pricing of a product line of unbundled quality-differentiated services was noted by Dupuit (1844) in the first treatise addressing the subject.

The problem must always be studied with care. Even in the case of inefficient delays caused by temporal discrimination, there can be offsetting gains from customers who buy late at low prices, whereas at a single uniform price above marginal cost they would not be able to buy at all. In other contexts, differentiation of services and products by unbundling along one or more dimensions of quality generally improves efficiency, provided the differentiated products are priced appropriately. The case of priority service
studied in §10 provides an example of quality differentiation in which a spectrum of lower and higher qualities is offered at prices that increase all customers’ net benefits. A similar analysis of advance purchase discounts would determine the pricing policy and restrictions on itineraries that ensure that the net effect is to recover the firm’s costs while extending service to customers with low valuations of service and low costs of restrictions on flexibility.

7.2. Incentives in Markets with Private Information

We turn now to the role of nonlinear pricing in coping with incentive effects in markets affected by private information. Previous chapters deal with this topic only in the limited sense that the firm offers a single tariff to a diverse population of customers. The firm knows the distribution of customers’ types in the population and the tariff is offered on equal terms to all; it is immaterial therefore whether the firm knows the types of individual customers. Of course this assumes that the firm’s costs do not depend on the individual served. We consider first the alternative case that the firm is concerned about each individual’s type, which materially affects the seller’s costs and about which it is uncertain (although each individual knows his own type).

Insurance Markets

The incentive problems resulting from private information are acute in insurance markets, so we use them to illustrate. Medical and property insurance are affected by private information in the form of a customer’s superior information about the chances of incurring illnesses or damages for which claims would be payable by the firm. The firm is in effect buying a lottery about which the seller (the insured) has better knowledge. This situation is called adverse selection by actuaries since the firm anticipates that customers with greater chances of claims have greater incentives to purchase policies.

A typical device to mitigate adverse selection is a form of nonlinear pricing. The firm offers policies for which the premium is disproportionately higher if the coverage is greater or the deductible is lower. The premium is more than proportionately greater for greater coverage because the firm (even a mutual company or beneficial society)

---

9 This produces a premium schedule that is an increasing and usually convex function of the coverage. A useful interpretation of the convexity of the schedule derives from the observation above that the firm can be interpreted as the buyer of a lottery rather than as the seller of a policy.
§7.2. Incentives in Markets with Private Information

anticipates that those customers purchasing greater coverage impose larger or more numerous claims. This has the inefficient consequence that low-risk customers are deterred from buying full coverage. Sometimes this is ameliorated by group policies that are affected less by adverse selection, or by preferred-risk policies that require the applicant to provide evidence of good health or a record of few claims. On the other hand, the firm collects sufficient premiums for each coverage to pay claims arising from those policies.¹⁰

An Illustrative Formulation *

To illustrate details of a formulation of this problem, suppose that the firm is a profit maximizing monopolist. Assume that customers’ types are described by their expected losses, so a customer of type \( t \) has expected losses equal to \( t \), and the distribution function of these types in the population is \( F(t) \). The firm then wants to maximize its expected profit

\[
\int_T \left[ P(q(t)) - tq(t) \right] dF(t),
\]

where \( q(t) \) is the coverage selected by type \( t \), namely, the fraction of \( t \)’s losses paid by the firm. The premium \( P(q) \) charged for the coverage \( q \) is a tariff chosen in anticipation that type \( t \) will select the coverage that maximizes the expected utility

\[
\int_Z U(qz - z - P(q)) dG(z; t),
\]

where \( U \) is the customer’s utility function describing his aversion to risk. Also, \( G \) is the distribution function of the customer’s loss \( z \) conditional on his type being \( t \); in particular, the mean of \( z \) is \( \int_Z z \ dG(z; t) = t \) according to the definition of \( t \). This formulation differs from §6.2 in that the customer is risk averse, so his choice process is more complicated, but the crucial feature is that the firm’s marginal cost of coverage depends on (in fact, is) the customer’s type \( t \). It is this feature that characterizes adverse selection.

Using nonlinear pricing to assure that premiums match claims for each coverage assumes that cross-subsidization among risk classes is excluded. An alternative is a uniform price per unit of coverage, but this brings its own inefficiency: high-risk customers prefer to buy coverage exceeding their losses (if allowed) and in any case low-risk customers prefer to underinsure.

¹⁰ See Rothschild and Stiglitz (1976) and Stiglitz (1977).
Example 7.2: To illustrate, suppose that customers’ losses are distributed independently and that type $t$'s loss has a Normal distribution with mean $t$ and variance $\sigma^2$. Further, suppose that all customers have the same utility function $U$, which is exponential with the risk aversion parameter $r$. In this case, type $t$’s certainty equivalent for the net risk $qz - z - P(q)$ after purchasing coverage $q$ is

$$u(q, t) = qt - t - P(q) - \frac{1}{2} r \sigma^2 [1 - q]^2,$$

measured in dollars. Using this data, the condition that characterizes optimal Ramsey pricing, as in §5.1, is

$$\alpha \hat{F}(t) - f(t) r \sigma^2 [1 - q(t)] = 0,$$

where $\alpha$ is the Ramsey number and $q(t)$ is the coverage selected by type $t$. The corresponding marginal price schedule that implements this allocation is

$$p(q) = t(q) + r \sigma^2 [1 - q],$$

where $t(q)$ is the type purchasing coverage $q$. For example, if the types are uniformly distributed then $\hat{F}(t) = 1 - t$; therefore,

$$q(t) = 1 - \frac{\alpha}{r \sigma^2} [1 - t],$$

if $t \geq t* \equiv \max \{0, 1 - r \sigma^2 / \alpha\}$, and zero otherwise. The optimal price schedule is

$$p(q) = 1 - [(1 / \alpha) - 1] r \sigma^2 [1 - q],$$

provided $r \sigma^2 \leq \alpha / [1 - \alpha]$ so that $p(q) \geq 0$. Notice that with full monopoly power the price of coverage is uniform: if $\alpha = 1$ then $p = 1$ uniformly. Excluding a fixed fee, the firm’s net revenue is

$$\text{Net Revenue} = \frac{1}{3} [1 + \alpha][1 - t_*]^3 / [r \sigma^2 / \alpha] - \frac{1}{2} [1 - t_*]^2.$$

This example shows some of the novel features that arise when the tariff design is affected by adverse selection. The price schedule is increasing, indicating that the tariff is convex. Also, the net revenue is negative if the Ramsey number is too small, indicating that some monopoly power is necessary to counter the effects of adverse selection. 

\[\diamond\]
Auctions and Trading Procedures

Auctions are another example of markets affected by private information. Indeed, auctions are allocative mechanisms designed mainly to elicit buyers’ estimates of the values of the items offered for sale. Familiar auctions in which each item is sold to the highest bidder are efficient in special cases; for example, when the buyers’ valuations are statistically independent. However, if their valuations are dependent or their estimates are correlated, as in the usual case that they are all estimating a common value, then the seller can increase the expected sale price by using nonlinear pricing and allocation rules. The literature on designing auctions that are optimal for the seller is essentially a translation of the standard theory of a multiproduct monopolist using nonlinear pricing. We describe briefly below its generalization to more general contexts in which the procedural rules for trading are designed to promote efficient outcomes.

An Illustrative Formulation *

The theory of optimal trading procedures is based on models like the following. Consider a group of traders comprising several buyers and sellers. Each buyer $i$ is willing to pay at most $v_i$ for a single item, and his gain is $v_i - p$ if he obtains an item at the price $p$. Each seller $j$ has a single item to sell and his gain is $p - c_j$ if he sells it at the price $p$. The sellers’ items are identical and each has only one item to sell; also, the buyers have ample money and want only one item. Say that a subset $K$ of the traders is feasible, in the sense that they can exchange items among themselves, if it has equal numbers of buyers and sellers (possibly zero), and let $\mathcal{K}$ be the collection of feasible subsets. Each individual trader $k$ has a privately known type parameter $t_k$ that is his valuation of an item: if $k = i$ for a buyer then $t_k = v_i$ and if $k = j$ for a seller then $t_k = -c_j$, and similarly their money transfers $y_k$ when they trade are either the price $y_k = p$ for a buyer or its negative $y_k = -p$ for a seller. A general trading procedure, like an auction, allows each trader $k$ to submit a bid $b_k(t_k)$ depending on his type and then based on these bids a feasible subset $\tilde{K}$ is selected and the traders in $\tilde{K}$ exchange items; also, trader $k$ gets the money transfer $y^K_k(b)$ depending on the list $b = (b_k)$ of bids submitted. Let $x_K(b)$ be the probability that $\tilde{K}$ is the selected feasible set when the list $b$ is submitted. Feasibility requires that $\sum_{K \in \mathcal{K}} x_K(b) = 1$ for each list $b$ and $\sum_k y^K_k(b) = 0$ for each

---


12 This is demonstrated by Bulow and Roberts (1989).
§7. OTHER APPLICATIONS OF NONLINEAR PRICING

$K \in \mathcal{K}$ and each list $b$. For the following derivation we assume that the traders' types are independently distributed: let $F_k(t_k)$ be the distribution function of $k$'s type $t_k$ as perceived by other traders, and let $f_k(t_k)$ be the corresponding density function. Assume that the hazard rate $f_k(t_k) / \hat{F}(t_k)$ is increasing, where $\hat{F}(t_k) = 1 - F_k(t_k)$.

We use the method in §8.2 to characterize the trading procedure that maximizes the traders' expected total surplus, calculated ex ante before they learn their types. The bid $b_k(t_k)$ submitted by trader $k$ must maximize his expected profit:

$$U_k(t_k) = \max_{b_k(t_k)} \mathcal{E}_k \left\{ \sum_{k \in K \in \mathcal{K}} x_K(b(t)) t_k - \sum_{k \in K} x_K(b(t)) \tilde{y}_K(b(t)) \right\},$$

where in this case the expectation $\mathcal{E}_k$ is calculated over the possible types of all traders other than $k$. The envelope property therefore requires that

$$U_k'(t_k) = \mathcal{E}_k \left\{ \sum_{k \in K \in \mathcal{K}} x_K(b(t)) \right\},$$

which is just $k$'s perceived probability that he gets to trade. The individual rationality or participation constraint from §6.1 requires that $U_k(t_k) \geq 0$ so that $k$ has a motive to participate in the procedure. Because the exchanges balance between the buyers and the sellers, feasibility implies that

$$\mathcal{E} \left\{ \sum_k U_k(t_k) \right\} = \mathcal{E} \left\{ \sum_{K \in \mathcal{K}} x_K(b(t)) \sum_{k \in K} t_k \right\}.$$

This appears to be a very weak constraint, requiring only that exchanges balance in expectation, but actually it is stronger in combination with the other constraints: the individual rationality constraint will ensure that the least type of each trader gets zero profit and then the envelope property determines the expected profit of each type. In sum, therefore, the design problem is to select a specification $\langle x, y \rangle$ of the procedure that maximizes the expected total surplus $\mathcal{E} \left\{ \sum_k U_k(t_k) \right\}$ subject to the envelope property, the individual rationality constraints, and the balance constraint.

As in §8.2, this problem is addressed by maximizing an augmented objective that includes a Lagrange multiplier $\mu_k(t_k)$ for the envelope property and another multiplier $1 + \lambda$ for the balance constraint. From this formulation one derives several necessary

---

13 The trader's second-order necessary condition can be shown to require that $U''(t_k) \geq 0$; that is, higher types must have higher probabilities of trading, which is ultimately a consequence of the increasing hazard rate assumption.
7.2. Incentives in Markets with Private Information

conditions, which are also sufficient due to the increasing hazard rate assumption. First, optimization of type $t_k$’s profit $U_k(t_k)$ requires that $\mu_k(t_k) = \lambda F_k(t_k)$ for those types large enough to obtain positive expected profit. Then, optimization of the probabilities $(x_K(b(t)))_{K \in \mathcal{K}}$ implies that it is sufficient to assign probability $x_K(b(t)) = 1$ for some feasible set $K \in \mathcal{K}$ for which the value of

$$\sum_{k \in K} \left[ t_k - \alpha \frac{F(t_k)}{f(t_k)} \right]$$

is maximal, where $\alpha = \lambda/[1+\lambda]$ — including the value zero for the empty set of traders. This is the key result, since it identifies those types that get to trade. In effect, it says that the gains from trade are maximized with respect to “virtual” valuations that take account of buyers’ incentives to submit bids below their true valuations and seller’s incentives to submit offers above their true costs. The transfers and thereby the prices at which transactions are consummated can subsequently be inferred from these conditions. In fact, a large variety of transfer schemes will work, but in some implementations it suffices to select a single price that clears the market for the submitted bids and offers from the buyers and sellers, as we illustrate with an example.\(^\dagger\)

\(\diamond\) **Example 7.3**: Suppose there is a single buyer and a single seller whose valuations are uniformly distributed between zero and one. Thus, $F_k(t_k) = t_k$ if $k = 1$ and $t_1 = v$ is the buyer’s valuation, and $F_k(t_k) = 1 + t_k$ if $k = 2$ and $t_2 = -c$ is the seller’s cost. The preceding analysis implies that they should trade if and only if

$$[v - \alpha (1 - v)] - [c + \alpha c] \geq 0,$$

or $v - c \geq \beta \equiv \alpha /[1 + \alpha]$. Applying this characterization to the balance condition indicates that feasibility requires $\beta = 1/4$, corresponding to $\lambda = 1/2$ and $\alpha = 1/3$. An implementation that achieves this result stipulates that they trade if the buyer’s bid $b(v)$ exceeds the seller’s offer $a(c)$, in which case the price splits the difference: $p(a, b) = [1/2][a + b]$. The seller’s optimal offer and the buyer’s optimal bid are

$$a(c) = \max \{ c, 1/4 + [2/3]c \} \quad \text{and} \quad b(v) = \min \{ v, 1/12 + [2/3]v \}.$$

Caution is advised, however: this procedure allows other pairs of mutually optimal strategies that do not attain the efficient outcome. \(\diamond\)

\(^\dagger\) This example is due to Chatterjee and Samuelson (1983) and Myerson and Satterthwaite (1983).
This style of analysis can be applied to a great variety of optimal design problems, including ones involving many traders with elastic supplies and demands. Ramsey pricing as addressed in this book is merely the special case in which the purpose of the procedure is to identify an allocation of the firm’s supply among customers that is efficient subject to the firm’s revenue requirement. And, we focus on implementations that accomplish this allocation by having the firm offer a nonlinear tariff from which customers choose their preferred purchases based on private information about their types.

Risk Sharing

Partnerships and other joint ventures encounter different versions of adverse selection. Casualty insurance is sold to many customers with independent risks of claims, and therefore total claims are largely predictable. In contrast, partners share a common risk regarding the financial success of the enterprise. Efficient risk sharing therefore depends sensitively on the nature of the uncertainty and on each member’s tolerance for risk. Private information can intrude in many ways: each member’s risk tolerance is private information, and each may have privileged information about the prospects of the enterprise. This information can be important both for risk sharing and for investment decisions. A few special utility functions and probability distributions have the ideal property that linear sharing rules are efficient for risk sharing and also provide sufficient incentives for revelation of private information (Wilson (1984)). In general, however, nonlinear sharing rules are required for efficient risk sharing, and they must be further modified (with some loss in efficiency) to encourage revelation of private information. The formulation used for the preceding analysis of optimal trading procedures can encompass the design of sharing rules in a partnership, as we illustrate briefly.

An Illustrative Formulation *

We envision several members indexed by $i$ who can choose among several risky projects indexed by $j$. If chosen, the $j$-th project yields a net income $y_j(\theta)$ depending on the realization of a random variable $\theta = (t, \tau)$ that has a known distribution function $F(\theta)$. The portion $\tau$ is never observed but $t = (t_i)$ is a list of parameters known privately by the members; that is, member $i$ knows $t_i$ initially, and we interpret $t_i$ as the type of member $i$. A decision rule is a pair $\langle x, s \rangle$ in which $x = (x_j)$ lists the probabilities that each project is chosen and $s = (s_{ij})$ lists the members’ shares of the income obtained. Naturally, $\sum_j x_j = 1$ and $\sum_i s_{ij} = y_j$. The key feature is that each member $i$ can
submit a report \( t_i \) regarding his private information; consequently, both \( x_j(t) \) and \( s_{ij}(y_j, \hat{t}) \) can depend on the reports submitted, and of course the shares must depend on the income obtained. Incentive compatibility is interpreted in this context as the requirement that each member has an incentive to report truthfully if he expects others to do so. Thus, the decision rule must be designed both to induce truthful reporting and to promote efficient selection of the project. Assume that member \( i \) has a utility function \( u_{ij} \) that can depend on the project \( j \) selected as well as his share and the outcome. Then his expected utility is

\[
U_i(t_i) = \max_{t_i} \mathcal{E}\{ \sum_j x_j(t)u_{ij}(s_{ij}(y_j, \hat{t}), \theta) \mid t_i \},
\]

where the conditional expectation is calculated given \( t_i \) and the presumption that others report truthfully: \( \hat{t}_k = t_k \) for members \( k \neq i \). The envelope property requires in this case that

\[
U_i'(t_i) = \mathcal{E}\{ \sum_j x_j(t)[u_{ij}(\theta) + v_{ij}] \mid t_i \},
\]

where if \( f(t \mid \tau) \) and \( f_i(t_i) \) are the conditional and marginal density functions then

\[
\phi_i(\theta) = \frac{\partial f(t \mid \tau) / \partial t_i}{f(t \mid \tau)} - \frac{\partial f_i(t_i) / \partial t_i}{f_i(t_i)},
\]

\[
v_{ij}(s; \theta) = \partial u_{ij}(s; t, \tau) / \partial t_i.
\]

The participation constraint is expressed by the requirement that \( U_i(t_i) \geq U^\circ_i(t_i) \), where \( U^\circ_i(t_i) \) represents a utility level that member \( i \) can obtain by severing his membership.\(^{15} \)

Finally, in view of the envelope property, one more constraint is needed to determine the absolute level of member \( i \)'s expected utility: the identity

\[
\mathcal{E}\{ U_i(t_i) - \sum_j x_j(t)u_{ij}(s_{ij}(y_j, \hat{t}), \theta) \} = 0
\]

suffices as this feasibility constraint. An efficient decision rule is therefore one that maximizes some weighted sum \( \mathcal{E}\{ \sum_i \lambda_i U_i(t_i) \} \) of the members' expected utilities subject to the restrictions represented by the envelope property, the participation constraint, and the feasibility constraint. The weights \( \lambda_i \) are intended to summarize a welfare criterion: they could also represent the outcome of a bargaining process among the members or in

\(^{15} \) Some formulations also impose additional participation constraints to assure each coalition of members what they could obtain separately.
CHAPTER VII
OTHER APPLICATIONS OF NONLINEAR PRICING

a market context the values of the resources they contribute, in which case each weight might depend on the member’s type.

To apply the method of §8.2, a Lagrange multiplier $\mu_i(t_i)$ is associated with the envelope property, and another multiplier $\lambda_i$ is associated with the feasibility constraint. For simplicity we consider here only the case of one-dimensional types and presume sufficient monotonicity conditions so that $U_i(t_i)$ will be an increasing function of $i$’s type. In this case, the necessary condition for an optimal choice of $U_i(t_i)$ indicates that

$$\mu_i(t_i) = [\lambda_i - \lambda_i] F_i(t_i)$$

if $U_i(t_i) > U^*(t_i)$. Consequently, we define the virtual utility

$$\hat{v}_{ij}(s, \theta) = u_{ij}(s, \theta) - \alpha_i \frac{F_i(t_i)}{f_i(t_i)} [u_{ij}(s, \theta)\phi_i(\theta) + v_{ij}(s, \theta)],$$

where $\alpha_i = [\lambda_i - \lambda_i]/\lambda_i$. Then the optimal shares assigned to each project $j$ are selected to attain the aggregate benefit measure

$$\hat{a}_{ij}(y_j, t) = \max_{\sum_i s_i = y_j} \mathbb{E}\{ \sum_i \lambda_i \hat{v}_{ij}(s_{ij}, \theta) \mid y_j, t \}.$$ 

Moreover, an optimal project is one that attains the maximal value of $\mathbb{E}\{ \hat{a}_{ij}(y_j(\theta), t) \mid t \}$. These calculations depend on the multipliers $\lambda_i$, but actually each can be determined from the feasibility condition, one form of which is

$$\mathbb{E}\{ \sum_j x_j [\lambda_i u_{ij} - \lambda_i \hat{a}_{ij}] \} = 0,$$

where arguments of functions are omitted for simplicity. If the members’ utilities are linear in their shares then maximization of expected total surplus entails $\lambda_i = 1$ and $\lambda_i = \lambda$ independently of $i$; further, $\lambda$ is chosen so that

$$\mathbb{E}\{ \sum_j x_j \sum_i [u_{ij} - \frac{F_i}{f_i} [u_{ij}\phi_i + v_{ij}]] } = 0$$

for the optimal decision rule.

A significant implication of this formulation is that private information useful for project selection affects efficient risk sharing. In effect, the members share risk according to their virtual utility functions rather than their actual ones. This provides incentives for truthful revelation of productive information, but it also hinders the more efficient risk sharing that would be possible if informational asymmetries did not intervene.

This formulation illustrates the variety of problems to which the methods of §8 can be applied. These methods were originally devised to study Ramsey pricing and optimal
7.3. Incentives in Labor Contracting

Many kinds of contracting invoke nonlinear pricing but it is especially significant in labor contracting. Its role stems from informational differences between the parties and the importance of risk sharing.\(^{16}\) If labor inputs and outputs are verifiable by both parties then a linear compensation scheme such as a piece rate or a salary suffices; or failing this, if the employee is not risk averse, then the employer can as well sell the activity to the employee, since a worker bearing all the risk has ample incentive to expend effort and care in the task. More commonly, however, it is costly or impossible for the employer to observe some inputs or outputs, and the employee is unable or unwilling to bear the entire risk. In such cases it is efficient for the employer and the employee to share risk; moreover, the worker’s share is an incentive to pursue the task vigorously. The worker’s share, based on observable performance measures, is a form of nonlinear pricing of output.

We consider only the case that output is observed directly, and ignore team production in which an individual’s contribution cannot be distinguished. Compensation that is a nonlinear function of measured output arises in this case when labor inputs by the worker are observed imperfectly by the employer. This input might be the worker’s effort or care, or other factors that are costly for the worker to provide: we measure all outputs and inputs in money terms. It is important to realize that this case depends also on the presence of either private information or an exogenous source of uncertainty. Otherwise, from the output the input might be inferred. In the case of sharecropping, for instance, the size of the crop is observed but the effort required to produce that crop cannot be inferred: the worker may have information about fertility, pests and rainfall, but even given this information, output depends stochastically on other factors unobserved by either party. Although labor contracting usually has some aspects of bilateral

\(^{16}\) In the technical literature, contracting with these features is sometimes called a principle-agent problem, after the legal terminology referring to contracts affected by statutes and common law precedents regulating agency relationships: the worker’s performance of a task as an agent for the principle allows discretionary choices by the worker.
monopoly, we assume here that the employer has all the bargaining power: the employer offers a compensation schedule and the worker either accepts or takes another job at the prevailing market wage.

Most labor contracting has important dynamic elements, since employment relationships often continue over time. Long term and repeated relationships differ greatly from short term ones, but these changes do not alter the role of nonlinear pricing, so we omit them here.\textsuperscript{17}

**Nonlinear Compensation for Output**

The simplest case omits consideration of risk sharing. Suppose the worker knows a parameter $t$, interpreted as his type, but the employer is uncertain: a probability distribution $F(t)$ describes this uncertainty. The output $q$ depends on both $t$ and the worker’s input $x$. Typically the type represents factors that reduce the effort required from the worker to produce a given output: in the sharecropping context these factors could include soil fertility, rainfall, and an absence of pests. This formulation can be stated alternatively by saying that the worker’s net benefit is $P(q) - C(q, t)$ if he is paid $P(q)$ for the output $q$ and he incurs the cost $x = C(q, t)$ to produce this output when the type parameter is $t$. Thus, the employer’s objective is to maximize the expected value of output net of compensation, $q - P(q)$, anticipating that the worker will choose to produce the output $q(t)$ that maximizes his net benefit when he knows the type is $t$. Moreover, the employer must offer compensation sufficient to induce the worker to accept employment rather than to seek the alternative wage:

$$P(q(t)) - C(q(t), t) \geq w,$$

for each type $t$, assuming the alternative wage $w$ is measured net of the worker’s inputs.

This formulation is evidently the same as the formulation of the nonlinear pricing formulation in §6 except payments are reversed: the firm (employer) pays the customer (the worker). This is often true of pricing factor inputs: payments are reversed but otherwise the nonlinear pricing formulation remains intact. The reversed direction of payment does have the effect that typically the compensation function $P(q)$ is increasing and convex, rather than concave, as seen previously in the case of an insurance market.

\textsuperscript{17} A survey of contracting, including dynamic features, is by Hart and Holmström (1987).
Example 7.4: Suppose the worker's cost is \( C(q, t) = [1-t]q \), so that \( 1-t \) represents directly the worker's cost per unit of output. If the type distribution has the special form \( F(t) = t^a \) then the employer's optimal compensation function provides a fixed wage plus a constant price \( p \) per unit of output; for instance, \( p = \$0.50 \) if \( a = 1 \). But for most other distribution functions the schedule is nonlinear, as for the examples in §6.6.¹⁸

To illustrate an alternative interpretation, consider a negotiation between a homeowner and a construction contractor.¹⁹ If the owner is uncertain about the contractor's costs, and prefers different designs depending on the prices charged, then a protracted negotiation might be avoided by a procedure in which the owner offers a schedule of prices and designs (say, size of the house) from which the contractor chooses. That is, the contractor selects from the menu of options based on private information about his costs. In general, the advantages of nonlinear pricing derive from offering a menu of options from which customers or contractors select based on private information. In the case of bilateral negotiations, a further advantage is that it can avoid costly delays in reaching an agreement.

Risk Sharing

The employer's compensation design problem is more complicated when the worker is risk averse and there is exogenous uncertainty. Part of the complexity stems from the need to share risk efficiently; that is, the employer provides some insurance for the worker, but of course the worker cannot be insured entirely against output risks without eliminating his incentive to expend effort. We omit this aspect below and simply assume that the employer aims to minimize the expected cost of inducing the worker to expend a specified effort \( x^0 \).

An Illustrative Formulation *

It simplifies notation to assume that the worker chooses directly the probability distribution of output. Thus, more effort or care makes greater or better outputs more likely, and others, less likely. In particular, suppose there are \( n \) possible output levels \( i = 1, \ldots, n \) and the net effect of the worker's effort is to choose the probabilities \( x = (x_i)_{i=1,\ldots,n} \) that

---

¹⁸ The uniform price that splits profits equally with the worker is also the optimal maximin strategy for the employer over a wide class of possible cost functions for the worker. That is, rather than assessing the distribution function \( F \) the employer maximizes his minimum net profit, where the minimum is with respect to all the cost functions the worker might have. See Hurwicz and Shapiro (1978).

¹⁹ I am indebted to Gyu Wang for this illustration.
each of these occurs, where of course \( \sum_i x_i = 1 \). Let \( U(P(q) - C(x)) \) be the worker’s utility function defined on his net monetary gain after subtracting his cost \( C(x) \) depending on the distribution \( x \) that he chooses. The utility function \( U \) is increasing, and strictly concave if the worker is risk averse. The employer’s problem is to choose the utility level \( u_i \) that the worker obtains if the \( i \)-th output level \( q_i \) occurs. The expected cost of doing this when the worker is induced to choose the specified distribution \( x^0 \) is

\[
\sum_{i=1}^n x_i^0 P(q_i) = \sum_{i=1}^n x_i^0[V(u_i) + C(x^0)],
\]

where \( V \) is the inverse of the utility function \( U \). Namely \( U(V(u)) = u \) for each possible value of \( u \), so \( V(u) \) is the net monetary amount that yields the utility \( u \). Thus, the actual compensation if the output \( q_i \) occurs is \( P(q_i) = V(u_i) + C(x^0) \). Note that \( V \) is increasing and convex if \( U \) is increasing and concave.

Inducing the worker to choose the specified distribution \( x^0 \) requires that these utility levels ensure that \( x^0 \) is the worker’s optimal choice. This constraint requires that for each feasible choice \( x \) of the output distribution,

\[
\sum_{i=1}^n x_i^0 u_i \geq \sum_{i=1}^n x_i U(V(u_i) + C(x^0) - C(x)),
\]

so that the worker’s expected utility from choosing \( x^0 \) is no less than it is from any alternative choice. In addition, the feasibility constraint

\[
\sum_{i=1}^n x_i^0 u_i \geq U(w),
\]

is imposed to ensure that the worker is willing to accept the contract. These constraints are linear in the variables \( (u_i) \) if \( U \) is an exponential utility function. It is useful to note that if the worker also has some private information, then such constraints must be imposed for each possible state of the worker’s information; that is, for each possible type \( t \) the worker must prefer to take the action \( x^0(t) \) intended by the employer.

Taken together, these components of the employer’s design problem define a nonlinear constrained maximization problem. From the perspective of nonlinear pricing, the key feature is that the net result is a compensation scheme that usually depends nonlinearly on output.\(^{20}\)

\(^{20}\) One dynamic version of this formulation has the property that the worker’s compensation depends only on aggregates constructed from the numbers of times that the various output levels occur; cf. Holmström and Milgrom (1987).
Example 7.5: An alternative formulation supposes that the agent chooses an action that determines the probability distribution of output. Let \( g(q \mid a) \) be the density function of the output \( q \) depending on the action \( a \) chosen by the agent. Suppose further that the agent’s utility function has the additive form \( U(P(q)) - C(a) \) in which the agent’s cost \( C(a) \) is measured in terms of utility. Assume that the range of possible outputs is independent of \( a \). In this case, the condition that determines the optimal remuneration has the form

\[
1/U'(P(q)) = \lambda - \mu \frac{g_a(q \mid a)}{g(q \mid a)},
\]

where the multiplier \( \lambda \) is chosen large enough to ensure the agent’s participation, and the multiplier \( \mu \) is chosen to ensure that the agent prefers the action selected by the principle. For instance, if the agent chooses the mean of the output distribution, which is either an exponential distribution or a Normal distribution, then \( g_a/g \) is a decreasing linear function of the output \( q \). Consequently, if the utility function \( U \) is an exponential function then the remuneration \( P(q) \) is proportional to the logarithm of a linear function of output. Or, if \( U(P) = \sqrt{P} \) then \( P(q) \) is a quadratic function of output.

### 7.4. Taxation and Regulatory Policies

In this section we mention two applications in the public sector. The theory of Ramsey pricing was originally developed to provide a systematic framework to study commodity and income taxation. We mention here how the material in previous chapters can be interpreted in terms of income taxation; §16 provides a brief history of the subject. We then describe recent applications to the design of regulatory policies.

#### Taxation

In §6 we formulated Ramsey pricing as the design of a tariff that maximizes consumers’ surplus subject to the constraint that the firm’s net revenue is sufficient to cover its total costs. This formulation can be applied simplistically to the design of an income tax schedule by a reinterpretation. Interpreting the firm as the government, its objective is to maximize citizens’ surplus subject to the requirement that it raise sufficient tax revenues to cover its expenditures for public goods. The tariff \( P(q) \) is the tax assessed on the income \( q \) and in total the tax receipts must meet the government’s revenue requirement. Citizens or households are diverse and the government is assumed to know the distribution of their types in the population. One of type \( t \) obtains the net
benefit $U(q, t) - P(q)$ if its gross income is $q$ and its net income is $q - P(q)$. The gross benefit $U(q, t)$ must be interpreted in this formulation as net of all nonmonetary expenditures of effort required to generate the income $q$ given one’s type $t$.

Practical applications address complications omitted in this formulation. One is that progressive taxation, in which the marginal tax rate is an increasing function of income, may be motivated by ethical considerations that in terms of the Ramsey formulation must be included by specifying individuals’ welfare weights that are dependent on their incomes or their types, especially for types disadvantaged in producing income; or, it may reflect political realities, since votes are allocated equally. In addition, in some cases the tax schedule treats different classes of citizens differently, such as households with differing numbers of dependents; and incomes (and expenditures) from different sources are treated differently, such as the distinction between wage income and capital gains. A second is that the formulation must be considerably enriched to include risk aversion and other income effects as in §7; a variety of dynamic effects over time, such as individuals’ investments in education and skills; and incentives to expend effort in work, especially when income derives from a concatenation of effort, luck, and inheritance — both financial and genetic. These complications are inadequately addressed by the simplistic interpretation above and explain why we omit a systematic presentation of the general theory of taxation even though it encompasses nonlinear pricing by firms as a special case.

Design of Regulatory Policies

In the United States many public services are provided by privately owned utilities whose operations are regulated by state commissions. A recent literature interprets the agency relationship between the regulator and the firm as a kind of contracting in which, via its policies regarding investments, products, and prices, the regulator provides incentives for the firm to provide services efficiently in the public interest. Ramsey pricing is an older approach to this topic, but the recent literature examines also the role of private information known by the firm regarding technology and costs. Some of the analyses along this line parallel the ones above: the regulator offers the firm a nonlinear schedule of allowed profits depending on observable magnitudes such as investments, outputs, and prices, and then the firm selects an option from this menu based on its more detailed knowledge of technology and costs.21

21 This is an overly simplified view of this subject, especially because regulation is
An Illustrative Formulation *

A simple static formulation is the following. The regulatory commission is assumed to maximize a welfare measure that assigns weights 1 and 1 - \( \alpha \) to expected consumers’ and producer’s surplus, respectively.\(^{22}\) Producer’s surplus is uncertain because the firm’s cost cannot be observed by the commission and it depends on a type parameter \( t \) known privately by the firm. The commission is therefore limited to specifying the revenue \( P(q) \) that the firm is allowed, depending on a measure \( q \) of the output that it provides. In response to this incentive, the firm chooses an output level to maximize its profit:

\[
R(t) = \max_q \{ P(q) - C(q, t) \},
\]

where \( C(q, t) \) is its cost function depending on output \( q \) and its type parameter \( t \). The constraint \( R(t) \geq 0 \) expresses the requirement that the firm must break even. In addition, as in §8, the envelope property indicates that the firm optimizes its response to the incentive offered: \( R'(t) = -C_t(q(t), t) \), where \( q(t) \) is the optimal output level.

Suppose that total surplus is the difference \( W(q) - C(q, t) \) between a measure \( W(q) \) of the social value of output and the firm’s actual cost. Also, let \( F(t) \) be the distribution function of the firm’s type parameter. Then the commission’s objective is to choose the function \( P(q) \) to maximize the expectation

\[
\int_0^\infty \{ W(q(t)) - C(q(t), t) - \alpha R(t) \} \, dF(t)
\]

of the welfare measure, subject to the constraints mentioned above. Assume that the cost function \( C(q, t) \) and marginal cost function \( c(q, t) \equiv C_q(q, t) \) are increasing functions of the output, and decreasing functions of the type parameter; in particular, \( R(t) \) is increasing. Then the methods of §8 applied to this problem indicate that the optimal assignment of types to output levels is characterized by the condition that

\[
W'(q) = c(q, t) + \alpha \frac{F(t)}{f(t)} c_t(q, t)
\]


\( ^{22} \) One assumes \( \alpha > 0 \) to capture the political reality that regulatory commissions tend to favor consumers more than owners of regulated firms, perhaps because owners are not voters in the firm’s service territory. An alternative view is that \( \alpha \) reflects distortionary effects of taxation to provide transfers to the firm. Laffont and Tirole (1986) study an alternative formulation in which the firm’s cost is observed by the regulator.
at \( q = q(t) \), provided this assignment is nondecreasing. Having identified the assignment \( q(t) \) from this condition, the marginal revenue \( p(q) \equiv P'(q) \) is inferred from the property that \( p(q(t)) = c(q(t), t) \). In turn, the revenue function is \( P(q) = P_0 + \int_0^q p(x) \, dx \), where the fixed fee \( P_0 \) is chosen to satisfy minimally the requirement that the least type breaks even: \( R(0) = 0 \). In an application, one might assume that an inverse demand function \( D^{-1}(q) \) specifies the market price at which the demand for output is \( q \). Before cost is subtracted, total surplus is therefore the area under this demand function and the marginal gross benefit is \( W'(q) = D^{-1}(q) \) if there are no income effects among consumers.

\[ W'(q(t)) = 1/t + \alpha[b - t][1/t^2]. \]

Also, because \( p(q(t)) = C_q = 1/t \) the marginal revenue allowed the firm satisfies the analogous condition,

\[ W'(q) = [1 + \alpha]p(q) - \alpha bp(q)^2, \]

and therefore the optimal marginal revenue function is

\[ p(q) = W'(q) \div \{ \beta + \sqrt{\beta^2 - \alpha b W'(q)} \}, \]

where \( \beta = [1 + \alpha]/2 \), provided \( W'(q) \leq \beta^2/\alpha b \), and otherwise \( p(q) = W'(q)/\beta \). The key feature of this solution is that if \( \alpha > 0 \) then the marginal revenue allowed the firm exceeds the demand price \( D^{-1}(q) \) for the same output: the difference is made up in expectation by the fixed fee \( P_0 \), which may be negative and therefore constitute a subsidy. The net result is that output is less than the perfectly efficient quantity, but in expectation output is more than the firm would provide if it were allowed to maximize its profit as an unregulated monopolist. Similarly, if \( C(q, t) = C_0 + qt \) and \( D^{-1}(q) = 1 - q \) then \( p(q) = 1 - q/[1 - \alpha] \), which again inflates marginal revenue and encourages the firm to expand output beyond the quantity an unregulated monopolist would provide.

\[ \diamond \]

7.5. Summary

The forms of nonlinear pricing studied in previous chapters are motivated by rate design in the communications, power, and transport industries. Firms in these industries
specify tariffs to define service conditions for a large and stable, but diverse, population of customers with continuing demands. As in Ramsey pricing and its variants, increments in purchase size are differentiated to meet revenue requirements for cost recovery efficiently. The efficiency gains result primarily from taking account of the higher price elasticities of demands for increments to large purchase sizes. Via tactics such as Pareto-improving pricing, the firm can assure that no customer is disadvantaged compared to the uniform price that would raise the same revenue. The net effect is only to increase the supply provided by the firm.

Nonlinear pricing can also be used to exploit monopoly power. When it is applied inefficiently, without benefits for customers, it falls within the penumbra of price discrimination. An extreme example of price discrimination is temporally differentiated pricing in the case the firm disposes of an initial stock via a declining sequence of prices: the resulting delivery delays for customers are purely inefficient. In this and most other examples of the deleterious effects of price discrimination, the inefficiencies stem from quality degradation. That is, delivery time is a quality attribute and uniform pricing would result in the highest quality, namely immediate delivery. Regulatory agencies therefore examine proposals to ensure that inefficient quality degradation does not offset the gains from product differentiation, such as increased quantities supplied and a spectrum of qualities better adapted to customers’ preferences.

The product pricing applications reviewed in Section 1 are only a sampling of the pervasive role of nonlinear pricing. They suffice nevertheless to indicate that the subject is richer than the coverage in this book. Among the broader topics are pricing and product design to meet customers’ peakload requirements, the more general topic of bundling, and various forms of differentiation of delivery conditions. The latter in particular introduce complicating factors, such as the Coase property of temporal pricing.

The applications in Section 2 to markets with private information introduce a different motivation for nonlinear pricing. Adverse selection is important when one party to a transaction is affected by factors known only by the other. Nonlinear pricing can eliminate this problem by assuring that prices are sufficient to cover the costs imposed by those types selecting each option on the menu offered. Although this can impose inefficiencies, uniform prices are also inefficient in such situations.

Risk sharing among partners is a pure case of beneficial nonlinear pricing. Nonlinear sharing rules are generally essential to allocate risk efficiently, and also to promote revelation of private information useful for investment and production decisions.
The applications in Section 3 to labor contracting indicate, however, that the problem is complicated when risk sharing is further affected by the incentives it creates for expenditure of effort by workers. Absent risk sharing, optimal compensation schemes are direct applications of nonlinear pricing (with reversed payments). Related applications in Section 4 extend even to the design of regulatory policies in the context that the firm has private information about its technology or costs. But with risk sharing, designing an optimal compensation scheme requires solving a constrained optimization problem that is more complex than the ones studied in previous chapters. That the compensation function is designed in anticipation of the worker’s selection of effort (in response to the incentives provided) is still the guiding principle, but the conditions required to induce selection of the actions preferred by the employer must usually be treated explicitly as auxiliary constraints. Previous chapters allow omission of such constraints because the customers’ only actions are to select purchase sizes.
Chapter 8

BIBLIOGRAPHY

Because few references are provided in the text, this chapter provides a bibliography of the literature on nonlinear pricing and a capsule summary of its theoretical development. No attempt is made to be exhaustive; rather, the main ideas are emphasized and references are given to the key technical articles on which the various chapters are based. This literature provides supplementary material on the topic of each chapter.

Stephen Brown and David Sibley (1986) provide a comprehensive exposition of nonlinear pricing that complements much of the material here. Bridger Mitchell and Ingo Vogelsang (1991) summarize the elements of the theory and review applications in the telecommunications industry. For expositions and bibliographies of a larger literature on related topics, see the texts by Louis Phlips (1983, 1988a, 1988b) and Jean Tirole (1988, Chapter 3) and the review articles by Ronald Braeutigam (1989) and Hal Varian (1989).

8.1. A Short History

Although nonlinear pricing has long been used in practice, its theoretical development is recent. The principal achievement was also the first: the construction of a theory of optimal nonlinear taxation by James Mirrlees (1971). It is the basis for all the material in this book. This work is refined and exposited by Mirrlees (1976, 1986, 1990) and alternative formulations are developed by Kevin Roberts (1979) and Roger Guesnerie and Jesus Seade (1982). Tuomala (1990) reviews these contributions.

provides a survey and bibliography of recent developments in the theory of Ramsey pricing using uniform prices.

Mirrlees’ work is cast in a general model that allows many complicating features, such as income effects and choice of production technology, and it focuses on the construction of welfare-maximizing tax policies. Its application to firms’ pricing policies was not fully appreciated initially. Some contributions were made steadily (A. Gabor (1955) and M.M. Murphy (1977)), especially to the simpler theory of two-part tariffs; including Martin Feldstein (1972), Stephen Littlechild (1975), Walter Oi (1971), Yew Kwan Ng and M. Weissner (1974), and Richard Schmalensee (1981a). Three influential articles by George Akerlof (1970), Michael Spence (1973), and Michael Rothschild and Joseph Stiglitz (1976) emphasized the role of self-selecion and implicitly the necessity of differentiated prices to achieve efficiency in markets with a heterogeneous population of customers. Development of the theory of nonlinear pricing as an integral part of the theory of the firm blossomed suddenly in the late 1970s. The initial flurry included articles by Gerald Faulhaber and John Panzar (1977), Michael Mussa and Sherwin Rosen (1978), Panzar (1977), Roberts (1979), Joel Sobel (1979), Spence (1976a, 1980), and Robert Willig (1978). An application to utility pricing is reported by Roger Koenker and Sibley (1979) and there have been several applications to related topics in general microeconomic theory; for instance, Stiglitz (1977). The articles by Willig (1978) and Janusz Ordover and Panzar (1980, 1982) emphasize the implications of nonlinear pricing for efficient regulatory policies.


With few exceptions, much of the work on nonlinear pricing focuses on the special case of a monopolist vendor of a single product and/or a parameterized model of customers’ benefits with a single type parameter. This book does the same in Part II to introduce the reader to the subject, but as we emphasize in §8, for a single product much of the analysis is unaffected by the restriction to a single type parameter. We also attempt to make the theory accessible and applicable by concentrating in Part I on the use of the demand profile to represent market data. This eliminates reliance on type parameters and it enables numerical calculation of multiproduct tariffs as in §14.

The recent work on nonlinear pricing developed in close relation to the more general theory of optimal auctions and other trading processes that account for incentive effects — such as customers’ self-selection of their preferred purchases in the present context. The seminal articles are by Roger Myerson (1981), Myerson and Mark Satterthwaite (1983), and John Moore (1984). An application to regulatory policies developed by David Baron and Myerson (1982) initiated a literature that is surveyed by Baron (1989) and Spulber (1989b). The monograph by Laffont and Tirole (1992) develops a comprehensive theory of regulatory incentives as part of a general exposition of the principles involved in designing incentive provisions of procurement and agency contracts. An especially careful application to the design of incentives for an agent employed by a principal is by Guesnerie and Laffont (1984); as they illustrate, their construction also applies to a wider class of problems with a one-dimensional type parameter. Portions of this literature applicable to auctions and bargaining are surveyed by Wilson (1992) and John Kennan and Wilson (1992).

The Literature on Price Discrimination

Price discrimination has had negative connotations at least since the indictment by Dupuit (1844) and subsequent analyses by Arthur C. Pigou (1920, 1932) and Joan Robinson (1933). A sampling of subsequent literature includes James Buchanan (1953), Ralph Cassady

The association of nonlinear pricing with price discrimination stems from an inadequate and confusing categorization of pricing policies by Pigou. His analysis largely ignored customers’ self-selection of preferred choices from a menu of options available equally to all customers: prices were implicitly assumed to be uniform and different prices were offered to different customers or market segments. The term first-degree price discrimination was coined by Pigou to describe pricing practices that exhaust all or most of consumers’ surplus by charging each customer its reservation price, which was presumed known to the firm. Third-degree price discrimination referred to similar practices applied to separated market segments. Second-degree price discrimination referred to intermediate cases in which perfect price discrimination in a single market was limited by the seller’s inability to distinguish among customers or an inability to prevent arbitrage by customers. Subsequent authors, such as Tirole (1988) and Varian (1989), include nonlinear pricing within the penumbra of second-degree price discrimination by interpreting it as an imperfect form limited by each customer’s ability to select any one among the menu of the options offered. We have not used this terminology because it is hard to reconcile the pejorative interpretation of discrimination with the efficiency properties of nonlinear pricing derived from Ramsey pricing formulations.

The main strand of the recent literature on second-degree price discrimination uses nonlinear pricing of qualities as a paradigm, as in Mussa and Rosen (1978) and Maskin and Riley (1984a). A main theme of this literature is that a monopolist’s optimal prices lead each customer type to select a lesser quality and at a higher price than would result from efficient pricing at marginal cost — as shown in Figure 13.2 — which is termed quality distortion by some authors. The literature on intertemporal pricing of durable goods dissents from this view by showing that quality distortion in the form of delayed purchases is small when the seller has limited powers of commitment. Called the Coase Conjecture (Coase, 1972), this proposition has been studied by Jeremy Bulow (1982, 1986),

---

8.1. A Short History

Faruk Gul (1987), Gul, Hugo Sonnenschein, and Wilson (1976), Gul and Sonnenschein (1988), Nancy Stokey (1979, 1981), and Peter Swan (1972). A different dissent is registered by Srinagesh and Ralph Bradburd (1989), who posit a plausible model in which quality is enhanced because customers with lower total benefit nevertheless have higher marginal valuations of quality increments.

In Pigou’s lexicon, the term third-degree price discrimination described different (presumably uniform) prices offered to separate market segments. A typical example of third-degree price discrimination is a discount based on an observable attribute such as age (via discounts for senior citizens), location, or class, as in the usual distinction between residential, commercial, and industrial customers of regulated utilities. Subsequent interpretations of this category have included prices conditioned on other observable aspects, such as education, that are endogenously chosen signals of unobserved attributes, such as measures of ability in a labor market context; the seminal work by Spence (1973, 1974) initiated a large literature that in the product market context is described by Tirole (1988, §2.6.1.2). One strand of the literature on third-degree price discrimination emphasizes market tests for unfavorable welfare consequences as compared to uniform pricing. Robinson (1933) established that charging different uniform prices in separated markets with independent linear demand functions reduces total surplus. Schmalensee (1981b) shows more generally that an increase in aggregate output is a necessary condition for an increase in total surplus; Varian (1985) extends the analysis to cases with dependent demands and obtains general bounds on the change in welfare that allow, for instance, that differentiated prices are beneficial if they enable a market to be served that would not be otherwise.  

To counter negative connotations of second-degree price discrimination, we follow Brown and Sibley (1986) in emphasizing that nonlinear pricing is universally beneficial compared to uniform pricing provided options, such as a Pareto-improving tariff, are offered to ensure that no customers are disadvantaged. This is consistent with an analysis by J. Stephen Henderson and Robert Burns (1989) that exempts nonlinear tariffs derived via Ramsey pricing from the undue price discrimination label used in statutory provisions of laws establishing standards for public utility regulation. However, little in this book addresses the potential problem of quality degradation described in §1 and §15, except the analysis of priority service in §10. Our analysis of this topic illustrates that differentiation

\[ \text{2 For related work, see DeGraba (1990), Edwards (1950), Finn (1974), Greenhut and Ohta (1976), Katz (1987), Nahata et al. (1990), and Schwarz (1990).} \]
of service qualities, in this case service reliabilities, can be accompanied by efficient
pricing policies ensuring that all customers’ net benefits are improved, without altering
the firm’s net revenue.

Summary

The methodology used to analyze nonlinear pricing stems from Mirrlees’ adaptation
of Ramsey’s formulation of the problem of choosing prices to maximize an aggregate
measure of total welfare subject to a revenue requirement. The optimal tariff (or tax
schedule in Mirrlees’ context) is generally nonlinear if the population of customers is
diverse. Nonlinear pricing can be interpreted within Pigou’s category of second-degree
price discrimination by noting that customers with different price elasticities of demand
pay different average prices. Nevertheless, it is an efficient means of raising the required
revenue: the nonuniformity of marginal prices is explained solely by the fact that the
price elasticities of aggregate demand (as measured by the demand profile) are different
for different increments. Thus, its main features derive from differentiated pricing of
increments. As with linear pricing, each customer chooses a preferred selection from a
menu of options, but the schedule of prices for successive increments is not restricted to
be constant.

The technical literature developed since the late 1970s has brought nonlinear pricing
within the standard economic theory of the firm, both in the analysis of quantity dis-
counts and in the analysis of a product line comprising a spectrum of qualities. It has
been applied to monopoly, oligopoly, and regulated contexts as well. This work relies
substantially on models with convenient regularity properties, a single product, and a
single type parameter to avoid technical problems. However, it can be extended to the
more general contexts of multidimensional and multiproduct pricing addressed in Parts
III and IV. The basic principle derived from this literature interprets nonlinear pricing
as differentiated pricing of increments. The basic theory of bundling provides a unified
methodology in which to conduct the analysis; in addition, the example in §13.5 suggests
a wider role for bundling in multiproduct contexts. In the special context of quantity
discounts, Mirrlees’ formulation in terms of Ramsey pricing provides a framework that
encompasses most of the applications to regulated industries, as well as many of the ap-
plications in §15 to markets affected by informational disparities. Extensions to multipart
tariffs and other restricted forms follow modified versions of the same principles.

Nonlinear pricing is used by firms in many industries, although the forms of imple-
mentation vary substantially. Regulatory agencies increasingly condone offering a menu of optional tariffs provided old options are preserved to ensure that no customers are disadvantaged. This stems in part from acceptance of the basic conclusion that an efficient way to meet a regulated firm’s revenue requirement is to exploit its monopoly power according to the principles of Ramsey pricing, possibly modified to meet distributional constraints. Even if nonlinear pricing has the onerous connotations of second-degree price discrimination, it is an efficient means of meeting the firm’s revenue requirement. This conclusion about allocative efficiency is subject however to the important provisos that the firm’s operations are productively efficient, and that the efficient spectrum of qualities of products or services is provided; in particular, the latter is always suspect because of the evidence that quality degradation can be used to facilitate price discrimination.

8.2. Chapter References

In this section we direct the reader’s attention to portions of the literature pertaining to particular chapters.

Mathematical Techniques: The exposition in Parts I-III uses elementary methods of algebra and calculus except for occasional references to the Euler condition and the transversality conditions from the calculus of variations. These necessary conditions as well as various sufficiency conditions are presented in standard texts; I rely on Elsgolc (1961). Part IV depends on advanced methods of multivariate calculus but the exposition suppresses technical aspects in the formulations and derivations of optimality conditions. The formulations in §12 and §14 rely on the special form of multivariate integration by parts known variously as Green’s Theorem, the Divergence Theorem, or in a generalized form, the fundamental theorem of multivariate calculus, sometimes called Stokes’ Theorem; a popular reference is Schey (1973).

§1: Introduction. Some of the feasibility requirements for nonlinear pricing are mentioned by Scherer (1980, p. 315) who provides a long list of varieties of price discrimination, some of which can be interpreted as forms of nonlinear pricing.

§2: Illustrations. In addition to the illustrations in this chapter, several are presented in detail by Brown and Sibley (1986). Mitchell and Vogelsang (1991) provide a detailed description of applications to telecommunications. An application to pricing telephone services for customers with impaired hearing is described by Oren, Smith, and Wilson (1982a).
§8. BIBLIOGRAPHY

§3: Models and Data Sources. The material in this chapter is a composite of the articles cited above. Nearly all authors have made explicit assumptions that fit roughly within those cited in Section 4. Additional sufficiency conditions are presented in §8. The most general sufficiency assumptions are apparently those in Mirrlees (1976, 1986). In principle, the necessary and sufficient condition for the required monotonicity properties is based on the property of quasi-supermodularity developed by Milgrom and Shannon (1991).

§4: Tariff Design. The use of the demand profile to predict customers’ responses to nonlinear tariffs is drawn from Oren, Smith, and Wilson (1982b). It has not been adopted by other authors except that Brown and Sibley (1986) and Goldman, Leland, and Sibley (1984) use a parallel formulation for one style of proof; see also Tirole (1986, §3.5.1.3). Mitchell and Vogelsang (1991) provide an exposition and critique of this formulation based on a previous version of this manuscript. Basic references on bundling are Adams and Yellan (1976), McAfee, McMillan, and Whinston (1989), and Schmalensee (1984).

§5: Ramsey Pricing. The interpretation of Ramsey pricing as an extension of monopoly pricing is standard in the literature, dating at least from Ramsey (1927) and more recently Mirrlees (1971, 1976). For a recent analysis more in the spirit of Boiteux’s (1956) formulation, and with applications to regulatory policies, see Laffont and Tirole (1992). The initial work on Pareto-improving tariffs is by Faulhaber and Panzar (1977), Ordover and Panzar (1980), and Willig (1978). For an alternative analysis in a quality context based on regulation of minimum qualities or the imposition of a price cap, see Besanko, Donnenfeld, and White (1987). The material in this chapter is motivated by Brown and Sibley (1986, p. 83 ff.). The illustration in Section 3 on telephone tariffs is based on D. Heyman, J. Lazorchak, D. Sibley, and W. Taylor (1987). The application to a communication system for hearing-impaired customers is described in Oren, Smith, and Wilson (1982a). Train and Toyama (1989) describe an application to time-of-use tariffs for electricity used for pumping irrigation water by agricultural customers.

§6: Single-Parameter Disaggregated Models. The topics in this chapter are standard in the articles in the large journal literature on nonlinear pricing. Expositions of models with discrete types in which only the adjacent incentive compatibility constraints are binding include Cooper (1984) and Maskin and Riley (1984a); more general models are addressed by Guesnerie and Seade (1982) and Matthews and Moore (1987) and
8.2. Chapter References


7: Income Effects. Mirrlees (1976) and Guesnerie and Laffont (1984) include income effects in their formulations. The formulation in terms of the indirect utility function, introduced in §8, is used by Roberts (1979) to develop an alternative method of constructing tariffs when there are income effects of the type in the third example. His method has the advantage that it facilitates analysis of Ramsey pricing with variable welfare weights.

8: Technical Amendments. This exposition relies on Goldman, Leland, and Sibley (1984), Guesnerie and Laffont (1984), and Mirrlees (1976). The formulation of nonlinear tariffs in terms of the indirect utility function is based on Roberts (1979). The extension to multidimensional types is developed by Srinagesh (1985, 1991a). Another version of the analysis of multidimensional types is by McAfee and McMillan (1988), who invoke an assumption somewhat weaker than type-linearity of the marginal benefit functions to reduce the problem to one with a single type parameter. Their exposition provides explicit sufficiency conditions. The analysis of multipart tariffs is analogous to results in Chao and Wilson (1987) and Wilson (1989a).

9: Multidimensional Pricing. This chapter is based on Oren, Smith, and Wilson (1985), who provide sufficiency conditions. Section 4 on the item-assignment formulation is based on Wilson, Oren, and Smith (1980), who include a derivation of optimal fixed fees.

10: Priority Pricing. In Section 1 on priority service, the analysis of supply uncertainty is based on Chao and Wilson (1987) and Wilson (1989a); and of demand uncertainty, on Harris and Raviv (1981). Spulber (1990ab) provides an alternative analysis, including the special case that customers’ demands differ only by a multiplicative factor, for which equi-proportional curtailments are efficient. Section 2 on Ramsey pricing follows Wilson (1989b). The initial work on priority service is by Marchand (1974) and Tschirhart and Jen (1979); subsequent literature includes Chao, Oren, Smith, and Wilson (1986, 1989), Viswanathan and Tze (1989), and Woo (1990). Section 3 on priority queuing systems is based on Mendelson and Whang (1990). A different
formulation applied to priority scheduling of airport landing and takeoff slots is developed by Pitbladdo (1990).

**§11: Capacity Pricing.** This chapter is based on Oren, Smith, and Wilson (1985) and Panzar and Sibley (1978). See Srinagesh (1990b) for further analysis of the Panzar and Sibley model.

**§12: Multiple Products and Competitive Tariffs.** Sections 1 and 2 are special cases of the construction in §14: see the cautionary comments below. Srinagesh (1991b) studies the topic of Sections 1 and 2 in the case that one product is priced linearly and another is priced nonlinearly. He finds, for instance, that when demands for the products are independent a multiproduct monopolist nevertheless sets a lower price for the first product than would a duopolist selling only that product: the motive is to increase customers’ benefits that can then be extracted as profits via higher nonlinear prices for the other product. Section 3 regarding Cournot models is based on Oren, Smith, and Wilson (1982b), who also provide sufficiency conditions. The brief discussion in Section 4 is based on Meyer and Klemperer (1989) who provide detailed analysis of nonlinear pricing policies adapted ex post to realized demand conditions.

**§13: Multiproduct Pricing.** Sections 2 and 5 are based on Mirrlees (1976, 1986) and the analysis with one-dimensional type parameters in Section 4 is based on Mirman and Sibley (1980). Armstrong (1992) studies the example in Section 5, showing that a monopolist firm’s optimal tariff always excludes some customers from purchasing, and that a multiproduct firm gains from using a nonseparable tariff. See Champsaur and Rochet (1989) for a study of Bertrand competition between duopolists offering a spectrum of qualities, of which each customer purchases a single unit, in the case that customers’ type parameters are one-dimensional.

**§14: Multiproduct Tariffs.** This chapter is motivated by ideas presented in Wilson (1991). However, the formulas in that article are wrong (!) because own-price and substitution effects are not separated properly. Errors might remain in this revised version, which was not included in the manuscript reviewed by referees.

REFERENCES


REFERENCES


REFERENCES


REFERENCES


De Fontenay, Alain, Mary Shugard, and David Sibley (1990), *Telecommunications Demand Modelling*. Amsterdam: North-Holland.


REFERENCES


REFERENCES


REFERENCES


REFERENCES


REFERENCES


233
REFERENCES


REFERENCES


Energy Research Group, University of California, Berkeley, mimeo.


CONTENTS

Chapter/Section                                                                 Page
1. INTRODUCTION                                                                 4
   1.1. Motivations for Nonlinear Pricing                                         6
   1.2. Practical Uses for Nonlinear Pricing                                     9
   1.3. Feasibility of Nonlinear Pricing                                         12
   1.4. Illustration: The Electric Power Industry                               18
   1.5. Overview of the Chapters                                                 20
   1.6. Summary                                                                  27
2. ILLUSTRATIONS OF NONLINEAR PRICING                                           29
   2.1. Time and Newsweek’s Advertising Rates                                   30
   2.2. EDF’s Tariffs for Electric Power                                         32
   2.3. PG&E’s Curtailable Service                                               39
   2.4. MCI’s and AT&T’s Telephone Tariffs                                       43
   2.5. Federal Express’ Mail Rates                                              45
   2.6. Delta Airlines’ Frequent Flier Rebates                                   46
   2.7. Summary                                                                  50
3. MODELS AND DATA SOURCES                                                       52
   3.1. Descriptions of Demand Behavior                                          55
   3.2. Estimation of the Demand Profile                                         65
   3.3. Welfare Considerations                                                    70
   3.4. Cautions and Caveats                                                     75
   3.5. Standard Assumptions *                                                    77
   3.6. Summary                                                                  82
4. TARIFF DESIGN                                                                 83
   4.1. The Price Schedule of a Monopolist                                       83
   4.2. Extensions and Qualifications                                            92
   4.3. The Bundling Interpretation                                              100
   4.4. Fixed Costs and Fixed Fees                                               102
   4.5. Multipart Tariffs                                                        104

* Author’s address: Stanford Business School, Stanford, CA 94305-5015.
  Phone 650-723-8620; Fax 725-7979. Email: rwilson@stanford.edu.
This manuscript is derived from a book published by Oxford University Press in 1993.
REFERENCES

4.6. Summary 110

5. RAMSEY PRICING 112
  5.1. The Price Schedule of a Regulated Firm 113
  5.2. Pareto-Improving Tariffs 123
  5.3. Long-Distance Telephone Tariffs 136
  5.4. Summary 139

6. SINGLE-PARAMETER DISAGGREGATED MODELS 142
  6.1. A Model with Discrete Types 143
  6.2. Models with One-Dimensional Types 150
  6.3. Two-Part Tariffs 155
  6.4. Multipart Tariffs 160
  6.5. Nonlinear Tariffs 171
  6.6. Some Examples 174
  6.7. Fixed Costs and Fixed Fees 180
  6.8. Summary 184

7. OTHER APPLICATIONS OF NONLINEAR PRICING 186
  7.1. Product Pricing Applications 186
  7.2. Incentives in Markets with Private Information 194
  7.3. Incentives in Labor Contracting 203
  7.4. Taxation and Regulatory Policies 207
  7.5. Summary 210

8. BIBLIOGRAPHY 213
  8.1. A Short History 213
  8.2. Chapter References 219

Final Status for this chapter:
Chapter 8
Section 2
Table 0
Figure 0: , nf23, nf28, nf39, nf36b, nf37, nf38, nf4, nf11, nf21, nf9, nf33, nf77, nf102, nf3, nf87, nf102b, nf50, n
Example 0
Footnote 2
Page 240 (plus any graphs spilled over).