Wireless Network Pricing

Jianwei Huang and Lin Gao
Network Communications and Economics Lab
Department of Information Engineering
The Chinese University of Hong Kong

SYNTHESIS LECTURES ON COMMUNICATION NETWORKS #1

MORGAN & CLAYPOOL PUBLISHERS

ABSTRACT

Today's wireless communications and networking practices are tightly coupled with economics considerations, to the extent that it is almost impossible to make a sound technology choice without understanding its economic implications. In this book, we will focus on how pricing theory will help us to understand and build better wireless networks.

We start in Chapter 1 by discussing the detailed motivation behind this book. In particular, we show that economic mechanism is becoming an indispensable part of wireless network planning and operating, mainly due to the inherent conflict between the limited wireless resources (supply) and the fast growing wireless needs (demand). Then in Chapter 2, we introduce the basics of microeconomics, which will be especially useful for readers mainly having an engineering background. From Chapter 3 to Chapter 6, we introduce the key aspects of wireless network pricing one by one. The structure of each chapter is similar. We first introduce the necessary theoretical background, and then give two examples in wireless networking to illustrate the theory. The examples are meant for illustration purposes, and are biased as they are based on our own research results. More specifically, Chapter 3 focuses on social optimal pricing. Chapter 4 looks at the issue of monopoly, where a single service provider dominates the market and wants to maximize its profit, sometimes through price differentiation. Chapter 5 concerns the price competition among multiple service providers. Chapter 6 talks about the issue of network externalities. Finally, in Chapter 7, we come back to the larger topic of wireless network economics, and discuss the connections between pricing and several other economic mechanisms such as auction, contract, and bargaining.

We want to thank members of the Network Communications and Economics Lab (NCEL) in the department of Information Engineering at the Chinese University of Hong Kong. Many NCELers have contributed to the book through active discussions and providing useful suggestions. We also want to thank the co-authors of several prior papers that form the basis of many wireless applications introduced in this book: Mung Chiang, Lingjie Duan, Vojislav Gajic, Xiwei Huang, Aggelos K. Katsaggelos, Kwan Fong Erica Leung, Shuqin Helen Li, Shuo-Yen Robert Li, Zhu Li, Bixio Rimoldi, and Biying Shou. We also thank Tianyi Sky Hu, for helping drawing many of the figures in the theory part of this book. The first author would especially thank his PhD advisors, Randall Berry and Michael Honig, for bringing him into the fascinating world of network economics.

KEYWORDS

network economics, game theory, monopoly, oligopoly, price discriminations, market competition, network externalities, network optimization, distributed algorithms, wireless communications, wireless networks, resource allocation, power control, cognitive radio, dynamic spectrum management, access pricing

Contents

CHAPTER 1

Introduction

1.1 WHY THIS BOOK?

Today's wireless communications and networking practices are tightly coupled with economics considerations, to the extent that it is almost impossible to make a sound technology choice without understanding its economic implications. This simple fact motivates us to take a close and systematical look at how the economics interact with wireless technologies. In this chapter, we will outline the big picture of the wireless network economics, centered around the following two questions:

- Why should we care about economics in wireless networks?
- What are the unique challenges of wireless network economics?

We want to point out that wireless network economics is a large topic that is difficult to cover with a book of 200 pages, especially if we want to provide some concrete examples with some analytical details. Therefore, in the rest of the book, we choose to focus on one key aspect of wireless network economics - wireless network pricing - to give readers a partial but hopefully more focused and in-depth view of the challenging economics issues of the wireless networks. We will briefly outline the key topics of later chapters at the end of this chapter.

1.2 THE WIRELESS REALITY

Let us first imagine a world of “wireless utopia”, where the wireless spectrum is unlimited, the wireless technologies can provide a communication speed comparable to wireline networks, heterogeneous wireless technologies co-exist perfectly without mutual interferences, wireless users have reasonable demands that can always be satisfied without overburdening the network, and wireless service providers aim to maximize the social welfare instead of their own profits. In this perfect world, every user can enjoy seamless and high speed wireless connections whenever and wherever, and there is no reason to worry about the economics issues.

However, the reality of wireless networks is (almost) exactly the opposite. The wireless spectrum is very limited and overly crowded due to the static licensing policy, the communication speed of the latest wireless technologies is nowhere close to that of wireline networks when the distance between transmitter and receiver goes beyond tens of meters, heterogeneous wireless networks often exist with little or no coordinations, heavy mutual interferences between networks and devices are the norm rather than the exceptions, the exploding growth of wireless data traffic is far beyond the growth of wireless capacity, and the wireless service providers often care more about profits than the social welfare. Some of the above issues can be alleviated by the advance of wireless technologies; many others can only be addressed with a combination of technology advances, economic innovations, and policy reforms.

Next we will illustrate in a bit more details about several issues mentioned above, and outline how economics can help to improve the overall performance of the wireless networks and satisfaction levels for both users and service providers.

1.3 TENSION BETWEEN LIMITED WIRELESS SUPPLY AND GROWING DEMAND

One key reason for studying wireless network economics is to resolve the tension between limited wireless resources and the fast growing wireless demands.

Wireless resource is limited. Radio spectrum is limited, and only a fraction of them (mostly the lower frequency part) is useful for wireless communications over reasonable ranges. Because of the limited availability of wireless spectrum, it has been a tightly controlled resource worldwide since the early part of the 20th century. The traditional way of regulating the spectrum is the static licensing approach, which assigns each wireless application a particular piece of spectrum at each particular location. Currently, almost all spectrum licenses belong to government identities and commercial operators. This can be clearly seen in the frequency allocation map of any country or region.

However, new wireless technologies and services are emerging rapidly. This means that every new wireless commercial service, from satellite broadcasting to wireless local-area network, has to compete for licenses with numerous existing sources, creating a state of spectrum drought (1).

A key challenge for government regulators is how to allocate these ever decreasing and precious spectrum resources wisely to achieve the maximum benefits for the society. Among many possible solutions, the spectrum auction has been advocated and successfully implemented in many countries. This will help to allocate the spectrum to service providers who value the resources most, as these providers are typically the ones who have the best technologies and thus the capability to provide the maximum benefits to the customers.

A more revolutionary approach is to enable unlicensed wireless users to opportunistically share the spectrum with licensed users through dynamic spectrum management. This is motivated by the fact that many licensed spectrum bands are not efficiently utilized (2). For example, FCC has recently decided to open up the TV spectrum for unlicensed use, as long as the licensed users communications are protected. Microsoft has already built a testbed over its Redmond campus to demonstrate the practicality of such sharing (3).

Note that there are two economic issues under this dynamic spectrum management regime. First, the regulators need to provide enough economic incentives for the license holders to open up spectrum for sharing, otherwise complicated legal issues might arise. The law suite between FCC and National Association of Broadcasters in 2009 is a good example of this (4). Second, it remains an open question in terms of what kind of services and commercial business models can succeed in this newly open spectrum bands, considering the potentially unregulated interferences among multiple unlicensed service providers (5).

The other perspective of the limited wireless spectrum is the tension between the low and often unreliable data rates provided by today's wireless technologies and the fast growing needs of wireless users. One may argue that the Wi-Fi technology (e.g., IEEE 802.11 family) can already provide a speed of hundreds of Mbps, which is good enough even for high definition video streaming. However, the Wi-Fi technology has a very limited coverage (e.g., from 20 to 230 meters for indoor communications), and thus cannot provide a ubiquitous wireless access experience. The cellular network still remains as the only wireless technology that has the potential to provide seamless access and mobility solutions. Today's 4G cellular networks can provide a theoretical peak download speed of 100 Mbps, although the actual speed can be less than 10% of the theoretical one. The speed will be even less when tens of users are sharing resources of a same base station, which is often the case in practice. On the other hand, thanks to the introduction of sophisticated smartphones and tablets (especially iPhone and iPad), users have growing needs to enjoy high quality and highly interactive contents on-the-go. Consider, for example, the very popular video streaming application of Netflix, which has been available on the iPad platform since 2010. To stream a high-quality video, Netflix recommends a data rate of at least 5Mbps. An always smooth playback requires the speed to be much higher. Applications like these make the current cellular network very stressful. It is widely known that AT&T networks in big US cities such as New York City and San Francisco often experience very heavy congestions and low effective user data rates, ever since AT&T introduced iPhone on their networks from 2007. During the Christmas season of 2009, AT&T even tentatively stopped selling iPhone in New York City, and many suspected that it was due to AT&T's fear of not being able to support fast growing new iPhone users. Finally, Cisco has predicted that the global mobile data traffic will grow from 0.6 exabytes (10^{18}) per month in 2011 to 10.8 exabytes per month in 2016 (6), which corresponds to close to 80% of growth per year.

Due to the limited spectrum and the constraints of today's wireless cellular technologies, it is impossible to over-provision the wireless network as what we did for fiber-based wireline networks. In other words, technology alone is not enough to resolve the tension between the supply and demand in the wireless market, no matter in the short run or in the long run. It is thus very important to use economics to guide the operation of the market.

1.4 COUPLING BETWEEN ECONOMICS AND WIRELESS TECHNOLOGIES

The economics of wireless networks can be quite different from economics of other industries, mainly due to the unique characteristics of the wireless technology and applications.

From the wireless technology side, there are many choices today in the market, and each has its unique strength and weaknesses. For example, Wi-Fi technology can provide high data rates within a short distance, and the cellular technology provides a much better coverage with a much lower data rate. The economic models for these two technologies are thus very different. In practice, the commercial Wi-Fi usage are often charged based on connection time without data limit, while the cellular usage are often charged based on wireless data volume with a more relaxed time constraint.

In terms of wireless applications, each application has a unique Quality of Service requirement, resource implication on the networks, and sensitivity to price. For example, a video streaming application requires a wireless connection that supports a high data rate and stringent delay requirements. It is possible to charge a high price for such an inelastic application, although providing data rate higher than needed will not be useful. A file transfer application can adapt to different transmission speeds, but requires a very low bit error rate to ensure correct decoding. Such elastic application will be very sensitive to price, and can be arranged to be delivered when the network is not congested and the delivery cost per bit is low.

The key challenge of wireless network economics is to properly match the wireless technologies with the wireless applications via the most proper economical mechanisms. We also want to emphasize that the choices of wireless technologies and applications are not static. Which technology and application will dominant the market at what time will also heavily depend on the economic implications. For example, although the 4G cellular technology has been available to many operators globally, only a small subset of them are considering upgrading to 4G networks in the near future. The factors to be considered include how the upgrade costs evolve over time, how fast the users will accept the 4G technology, what types of applications will emerge and fully take advantage of the new technology, how the market competition will affect the pricing strategies, and how the network effect will affect the value of the new service (7). Thus it is critical to understand the impact of economics on the evolution of wireless network technologies and applications.

1.5 EFFECT OF MARKET DEREGULATION

The deregulation of telecommunication markets in many countries has made the study of wireless network economics more impotent than ever. In the past, very often there was only one major wireless service provider enjoying the monopoly status in a particular local (or national) market. Examples including AT&T in the US, China Mobile in China, and

Telcel in Mexico. However, the recent telecommunication deregulation leads to several major players in a single market. Examples include AT&T, Verizon, T-Mobile, and Sprint in the US, as well as China Mobile, China Unicom, and China Telecom in China. As a wireless service provider is ultimately a profit-maximizing entity, it needs to optimize the technology choices and pricing mechanisms under intense market competition.

Industry deregulation also brings more choices to the wireless consumers. For example, a user may freely compare and choose services from different service providers based on the service quality and cost. A user may even use different service providers for different types of services, such as using both cellular service and wifi service through the same cell phone. A service provider may no longer have complete control and knows the complete information of each of its subscribers. All these bring interesting and sometimes new economic questions that are not present in other industries.

1.6 WE ARE TALKING ABOUT WIRELESS

One may argue that researchers have studied Internet economics for more than a decade, and the lessons and results learnt there can be carried over to the wireless industry. However, the wireless network economics is significantly different from Internet economics in several ways.

First, characterization of network resources in wireless networks is more difficult than in wireline networks. Although wireless spectrum can be measured in hertz, the network resource corresponding to each hertz of spectrum is not easy to characterize. First, the wireless data rate is often highly stochastic over time due to shadowing, fading, and mobility. Second, the wireless resource is spatially heterogeneous, and the same spectrum may be concurrently used by multiple users who are physically far apart without affecting each other. Third, the wireless data rates are affected by mutual interferences. Although there are many analytical models characterizing the interference relationships, they can be either clean yet stylized (e.g., the protocol interference model) or precise yet complicated (e.g., the signal-to-interference-plus-noise ratio (SINR) model). There does not yet exist a model that is precise and analytically trackable for all practical wireless networks.

Second, the characterization of end users can be more complicated in wireless networks. A wireless user may have many different attributes, such as utility function (determined by the application type), total energy constraint and energy efficiency (determined by battery technology and charging levels), and channel conditions (determined by node locations and mobility). Also, the users performances are often tightly coupled due to mutual interferences.

Third, the interactions between wireless users heavily depend on the specific choice of wireless technology. In random medium access protocols such as slotted Aloha, users are coupled through their channel access probabilities. In Code Division Multiple Access (CDMA) network, users are coupled through mutual interferences. When we consider a spectrum overlay in cognitive radio networks, unlicensed users can not transmit simultaneously with the licensed users in the same channel at the same location. In a spectrum underlay cognitive radio network, unlicensed users are allowed to transmit simultaneously with the licensed users, as long as the total unlicensed interference generated at a particulate location is below an interference threshold. Different interactions and couplings between users lead to different types of markets and economic mechanisms.

Fourth, the coupling between technology, policy, and economics is different in wireless networks. We can use cognitive radio as an example to illustrate this point. Cognitive radio technology enables more flexible radio transmitters and receivers, and makes it feasible for wireless devices to sense and opportunistically utilize the spectrum holes. However, how and when cognitive radio technology should be used heavily depends on the type of spectrum band, which determines the types of licensed users and how they value the pros and cons of the new technology. Regulators in some countries are also more conservative than others in approving the new technology and changing the existing licensing practice. In fact, many wireless technologies can only work under the proper policy framework together with the right economic mechanisms that incentivize all parties involved.

1.7 OVERVIEW OF THE BOOK

This book will discuss how to use pricing in wireless networks through several chapters.

In Chapter 2, we introduce the basics of microeconomics. In particular, we derive the consumer's demand functions and the firms' supply functions, and explain the concept of market clearing where demand equals supply. Roughly speaking, pricing is a mechanism to regular market demand and supply under different system design objectives.

From Chapter 3 to Chapter 6, we introduce the key aspects of wireless network pricing one by one. The structure of each chapter is similar. We first introduce the necessary theoretical background, and then give two examples in wireless networking to illustrate the theory. The examples are meant for illustration purposes, and are biased as they are based on our own research results. Interested readers can refer to references for detailed technical discussions (such as proofs).

Chapter 3 focuses on social optimal pricing. This corresponds to the case where a service provider's interests are properly aligned with the regulator's interests through some economic mechanisms. We first introduce the theory of convex optimization, with focus on dual-based distributed optimization. Then we introduce the first example of resource allocation for uplink and downlink wireless video streaming. The prices correspond to the dual variables, and they are used to coordinate the resource allocation to satisfy the heterogeneous Quality of Service (QoS) requirements of multiple users in a single cell. Finally, we discuss the second example of downlink resource allocation among multiple base stations. Each base station has its own limited resource, and each mobile user in the network has different channel conditions to different base stations. The prices announced by the base stations will help to coordinate the users' base station selections and resource demands to achieve the social optimality.

Chapter 4 looks at the issue of monopoly, where a single service provider dominates the market and wants to maximize its profit, sometimes through announcing different prices to different users. We first introduce the theory of monopoly pricing and three types of price discriminations. Then we introduce the first example of revenue management in cognitive underlay networks, where the spectrum owner can control the demand elasticity of the users by adjusting total available bandwidth and tolerable interference level. In the second example, we study how to design an incentive-compatible pricing menu under incomplete information, in order to achieve the same maximum revenue under price differentiation with complete information. We also discuss how to optimize the price differentiation parameters when the service provider is constrained in terms of the number of prices it can select.

Chapter 5 concerns the price competition among multiple service providers. Such competition can be analyzed using game theory, which is introduced first in the chapter. Then we look at multiple classical market competition models, including Cournot competition based on output quantities, Bertrand competition based on pricing, and Hoteling model that captures the location information in the competition. In terms of applications, we first revisit the multiple base station model discussed in Chapter 3. Instead of looking at social optimal pricing as in Chapter 3, we will study how multiple base stations compete in the market by pricing their resources to attract customers and maximize their own revenues. A surprising result is that such a competition can lead to the social optimal resource allocation under proper technical conditions, even when the number of base station is small. In the second example, we examine how two secondary wireless service providers compete by leasing resources from spectrum owners and provide services to the same group of customers.

Chapter 6 talks about the issue of network externalities. There are two types of network externalities, depending on whether network entities positively or negatively affect each other. We introduce the basic theory of externality first, and then present two examples. In the first one, users generate negative externalities to each other due to interferences. The key idea to resolve this is to internalize the externality through Pigovian tax, which is called interference price in our example. We propose an Asynchronous Distributed Pricing algorithm, which globally and rapidly converges the global optimal solution (if it is unique). In the second example, two wireless networks interconnect with other to increase the customer base. We examine how wireless networks determine the access pricing to maximize either the social welfare or their individual profits.

Finally, in Chapter 7, we come back to the larger topic of wireless network economics, and discuss the connections between pricing and several other economic mechanisms such as auction, contract, and bargaining. Hopefully we will be able to discuss these topics in more details in a future book.

CHAPTER 2

Economics Basics

In this chapter, we will follow the convention of economics and use the terms of “firm” and “consumer”. A firm may represent an wireless service provider or a wireless spectrum owner, and a consumer can present a wireless user or a lower tier wireless service provider. In later chapters, we will give more concrete examples of firms and consumers in different wireless networks. The theory introduced in this chapter closely follow popular microeconomics textbooks such as (8; 9; 10).

2.1 SUPPLY AND DEMAND

Supply and demand in a market are both functions of market prices. When prices increase, usually the market supply increases as firms have more incentives to produce, and market demand decreases as consumers have less incentives to purchase. We first study how the demand and supply change with prices, and then characterize what prices lead to a market equilibrium where supply equals demand.

2.1.1 MARKET DEMAND FUNCTION

Let us consider a consumer who subscribes to a wireless cellular data plan. We may characterize the consumer demand as a function of price by the following table:

Table 2.1: Relationship between the monthly wireless data demand and the price per Gigabytes

Price Per Gigabytes	Wireless Data Demanded Per Month
$1	50 Gigabytes
$2	22 Gigabytes
$10	4 Gigabytes
$20	1.5 Gigabytes

Other consumers may have different demands for wireless data. If we add up all consumers’ demands together, we will obtain the relationship between the aggregate demand and the price, which we call the market demand function.

Definition 2.1 The **market demand function** $D(\cdot)$ characterizes the relationship between the total demand quantity Q_d and the product price P as follows

$$Q_d = D(P). \tag{2.1}$$

Figure 2.1 gives an example of the market demand function. Here we adopt the convention of placing price at the vertical y-axis and quantity (can be demand here or supply later on) at the horizontal x-axis. When the price decreases from P_1 to P_2, the demand increases from Q_1 to Q_2. There are two reasons for this inverse change of demand. First, the existing consumers who have positive demands at price P_1 will increase their demands when the price drops. Second, some consumers did not purchase at price P_1 may decide to purchase at the lower price P_2.

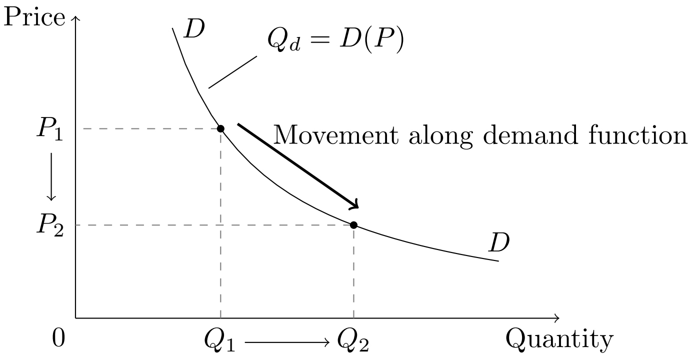

Figure 2.1: The market demand function $Q_d = D(P)$, and the shift along the function due to a price increase.

Besides shifting *along* the demand function due the price change, the demand function itself might also shift due to several reasons: (1) the change of consumers' income, (2) the price change of other products, and (3) the change of consumers' taste. Figure 2.2 illustrate such an example. Let us take wireless data service as an example. When consumers' income increases, the aggregate cellular data demand will increase (and thus the demand function will shift to the right), as consumers are more willing to use high price services (such as high-definition video streaming). When a price of a substitutable product (such as the price of commercial Wi-Fi access points) decreases, the aggregate cellular data demand decreases, as consumers are more willing to use the substitutable product. Finally, when consumers' tastes change due to education, the demand function may also shift.

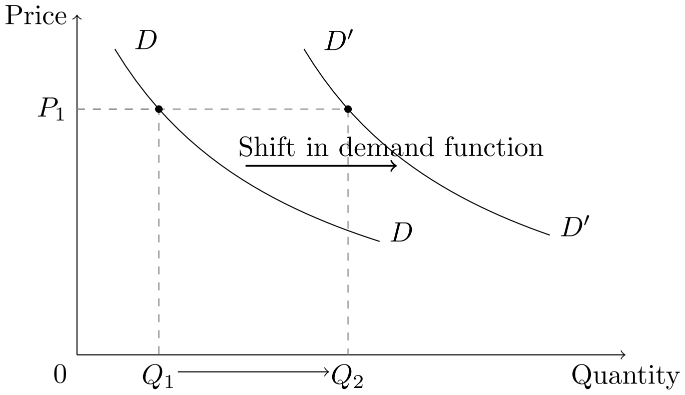

Figure 2.2: The shift of market demand function from $Q_d = D(P)$ to $Q'_d = D'(P)$.

2.1.2 MARKET SUPPLY FUNCTION

Similar as Definition 2.1, we can define the market supply function as follows.

Definition 2.2 The **market supply function** $S(\cdot)$ characterizes the relationship between the total supply quantity Q_s and the product price P as follows

$$Q_s = S(P). \tag{2.2}$$

Imagine the case where each firm does not have a capacity limit, and then the total market supply will increase with the price, as shown in Figure 2.3.

Similarly, the market supply function itself may shift when the price of a raw material (used for production) changes or the production technology changes. For example, consider a wireless service provider selling wireless services (e.g., data rates) to customers. The supply of wireless resource may change if the price for wireless spectrum (raw material that provides data rates) changes or the physical layer technologies changes (such as upgrading from the 3G CDMA-based cellular network to a more efficient 4G OFDMA-based network). We will leave it as an exercise for the readers to draw a figure of shifting market supply function similar to Figure 2.2.

2.1.3 MARKET EQUILIBRIUM

Now let us look at the interactions between supply and demand, which is a stable predication of the market.

Definition 2.3 At a **market equilibrium**, the aggregate demand equals the aggregate supply.

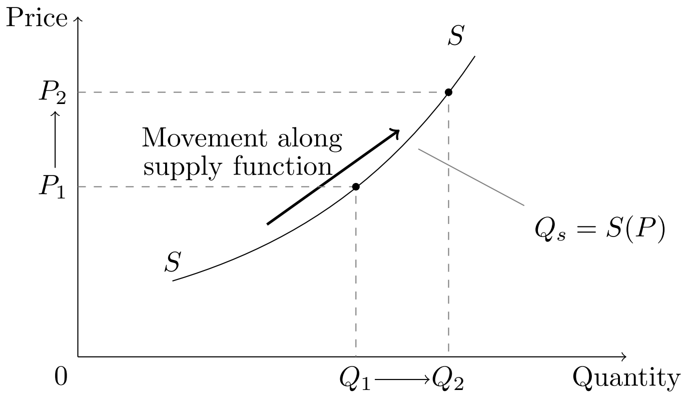

Figure 2.3: The market supply function $Q_s = S(P)$ and the shift along the function due to a price increase.

Apparently there will be a price associated with a market equilibrium. If the demand and supply functions are monotonic in terms of the price, then there is a unique interaction point which corresponds to the unique market equilibrium. The corresponding price is denoted by P_e and the (same) aggregate demand and aggregate supply is denoted as Q_e, i.e.,

$$Q_e = D(P_e) = S(P_e). \tag{2.3}$$

Figure 2.4 illustrates the market equilibrium. We want to emphasize that equilibrium is a prediction of how the actual market will look like, as the market is stable at the equilibrium and is unlikely to change once it has reached there. When the market price is lower than the equilibrium price P_e, for example, the aggregate demand is higher than the aggregate supply. In this case, consumers are willing to pay more to secure the limited supply, and the firms have incentives to produce more to earn more profits. As a result, the market price decreases until the equilibrium is reached. Of course, the real process of reaching the market equilibrium is more complicated than this, and we will provide some examples in later chapters.

When either market demand function or market supply function shifts due to factors other than the price, the equilibrium will also change accordingly.

2.2 CONSUMER BEHAVIOR

Now let us zoom into the behavior of a particular consumer, and understand how the market demand function $Q_d = D(P)$ is derived.

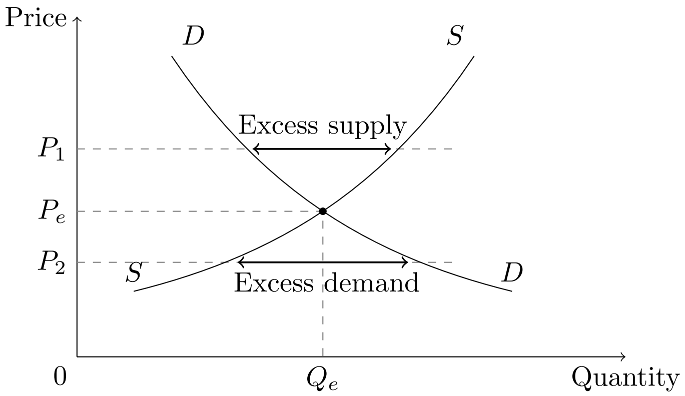

Figure 2.4: The market equilibrium price P_e and quantity (demand and supply) Q_e.

2.2.1 INDIFFERENCE CURVES

In order to understand a single consumer's demand, we first need to understand how a consumer evaluates the benefit of consuming certain products. For example, how would a consumer evaluate the satisfaction level of watching a 60-min action movie and playing 30 mins of video games on his iPad? To explain this, we first define the concept of market basket.

Definition 2.4 A **market basket** specifies the quantity of different products.

If we consider "watching movies" and "playing games" as two types of products, then watching a 60-movie movie and playing 30 mins of game can be represented by the market basket $(60, 30)$. We can use a utility function U to characterize the consumer's satisfaction level of consuming a certain market basket (x, y), i.e.,

$$U = U(x, y). \tag{2.4}$$

In Figure 2.5, we represent the basket $(60, 30)$ as point 1. We also add several baskets, where basket 2 is $(45, 40)$, basket 3 is $(30, 60)$, basket 4 is $(25, 25)$, and basket 5 is $(75, 65)$. Assuming that the consumer's utility is increasing in both x and y, then point 5 leads the maximum utility (among five baskets) and point 4 leads to the minimum utility. If we further know that that the consumer is indifferent among baskets 1, 2, and 3, then we say that these three baskets are on the same indifference curve.

Definition 2.5 An **indifference curve** represents a set of market baskets where the consumer's utility is the same.

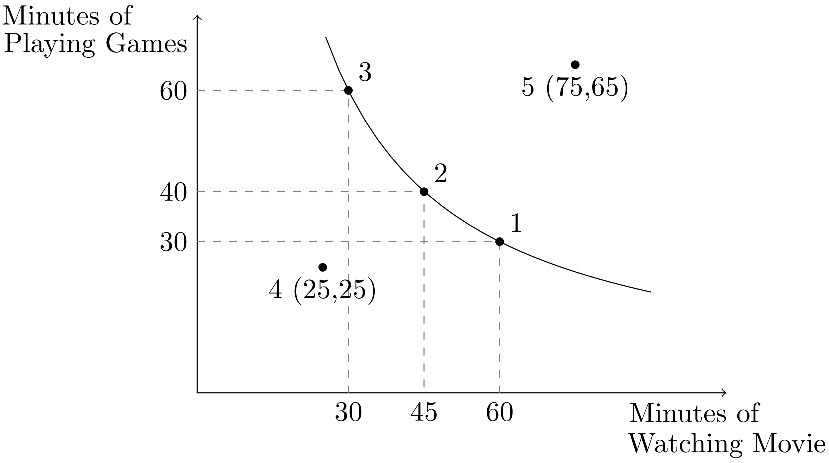

Figure 2.5: Market baskets and indifferent curve.

The indifference curve characterizes how a consumer trades off two different products. We can further draw an indifference map, which consists of all indifference curves of a consumer. That means in Figure 2.5, basket 5 will be on an indifference curve that has a higher utility than baskets 1, 2, and 3, and basket 4 will be on an indifference curve that has a lower utility than baskets 1, 2, and 3.

2.2.2 BUDGET CONSTRAINTS

If a consumer has enough income, he will definitely prefer to choose bask 5 in Figure 2.5 over other four baskets. However, the budget constraint will limit a consumer's choice.

Definition 2.6 The **budget constraint** characterizes which market baskets are affordable to the consumer.

In our example, we can consider the limited energy of the iPad battery as the budget. Assuming watching one minute of movie will cost 1 unit of energy, and playing one minute of game will cost 2 units of energy. Then the constraint of 100 units of energy leads to the budget constraint shown in Figure 2.6, which can be mathematically represented as $y = \frac{1}{2}(100 - x)$. The consumer can afford any market basket on or below the budget constraint. Alternatively, one can think of the price of watching movie as $P_x = 1/\text{min}$ and the price of playing game as $P_y = 2/\text{min}$. Thus the budget constraint can be represented by $x = \frac{I}{P_x} - \frac{P_y}{P_x} y$, where I is the fixed budget.

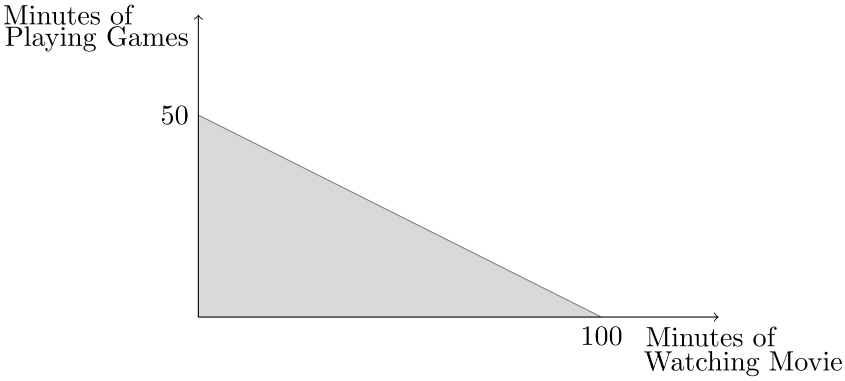

Figure 2.6: Budget constraint of 100 units of battery energy.

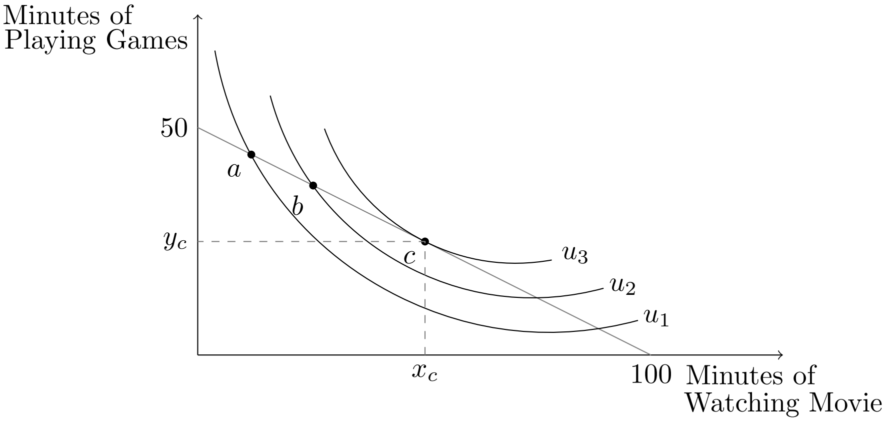

Figure 2.7: Consumer's optimal market basket choice is basket c.

2.2.3 CONSUMER CONSUMPTION PROBLEM

Once we consider both the consumer's indifference curve and the budget constraint, we will start to understand how a consumer decides which market basket to purchase. Essentially, the consumer wants to maximize its utility subject to the budget constraint. Geometrically, the consumer will find the highest indifference curve that "touches" the budget constraint.

Let us consider the illustration in Figure 2.7. It is clear that basket a or b does not maximize the utility, as basket c is on a higher utility indifference curve which "touches" the budget constraint (and thus is feasible). To be more precise, the derivative of the indifference curve with utility $U3$ at basket c equals to the slope of the budget constraint at basket c, i.e., the budget constraint is the tangent line to the indifference curve at basket c,

$$\left.\frac{\Delta y}{\Delta x}\right|_{U(x,y)=U_3,(x,y)=(x_c,y_c)} = -\frac{P_x}{P_y}. \tag{2.5}$$

The lefthand side of equation (2.5) is also called **marginal rate of substitution (MRS)**, which represents how much the consumer is willing to tradeoff one product with the other product. Here we constrain $U(x, y) = U_3$, which means that the MRS is measured along the indifference curve with a constant utility U_3. Unless in very special cases, the MRS along an indifference curve is not a constant, and that is why we need to specify basket c at $(x, y) = (x_c, y_c)$.

2.2.4 CONSUMER DEMAND FUNCTION

Now we are ready to derive a consumer's demand function, which characterizes how its demand of a product changes with the price of that product. The market demand function is simply the summation of all consumers' demand functions in the same market.

Assume that there are three games on iPad. The first one is a strategy game (e.g., Chess) that requires deep thinking and thus infrequent inputs and animations; the second one is a light game (e.g., Angry Birds) that contains some frequent simple animations; the third one is an action game (e.g., car racing) that features high-definition action-packed animations. The energy prices of these three games are 1/min, 2/min, and 4/min, respectively. With a total budget of 100 units of energy, the budget constraint will rotate around the point of $(100, 0)$ (under the fixed energy price of 1/min for movie watching), depending on which game to play. The optimal market basket that maximizes the consumer's utility will also change accordingly, denoted as baskets A, B, and C as shown in Figure 2.8.

The three points of A, B, C lead to three points on the consumer's demand curve (in terms of the demand of time for playing game). Connecting these points (or alternatively choosing different energy price for playing games and examining the utility maximizing baskets) will lead to the demand function as shown in Figure 2.9.

2.2.5 PRICE ELASTICITY

We notice that a consumer's demand is often downward slopping, i.e., a lower price leads to a higher demand. However, how fast the demand changes with the price depends on the nature of the demand. Consider the cellular wireless data usage as an example. A college student might be very price sensitive, and will dramatically decrease the monthly data usage if the price per Gigabytes cellular data increases. However, a business user might be much less sensitive and even did not notice the change of price until several months later. Such sensitivity of demand in term of price can be characterized by the price elasticity.

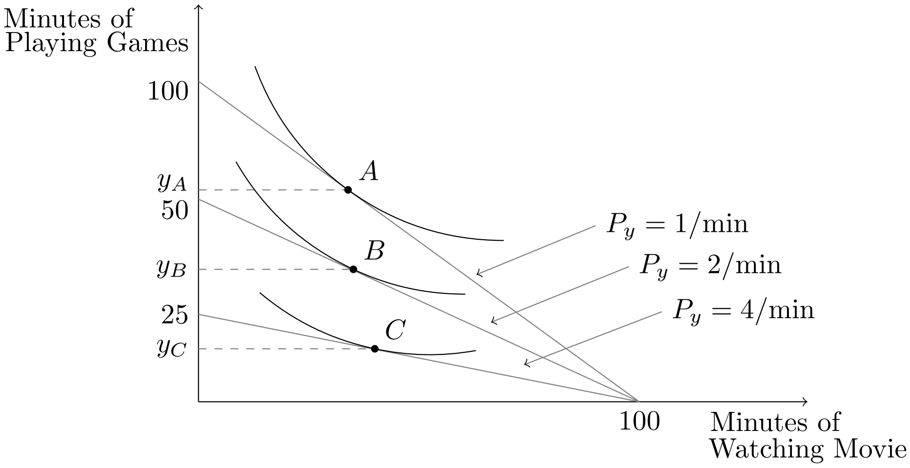

Figure 2.8: Consumer's different optimal market basket choices under different energy prices for playing games.

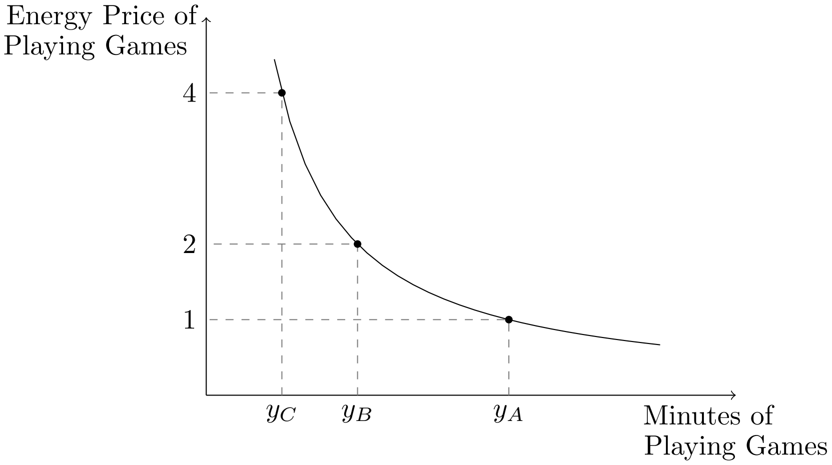

Figure 2.9: Consumer's demand function for playing iPad games as a function of the energy cost.

Definition 2.7 The **price elasticity of demand** measures the ratio between the percentage change of demand and the percentage change of price, i.e.,

$$E_d = \frac{\%\text{ change in demand}}{\%\text{ change in price}} = \frac{\Delta Q_d/Q_d}{\Delta P/P}. \tag{2.6}$$

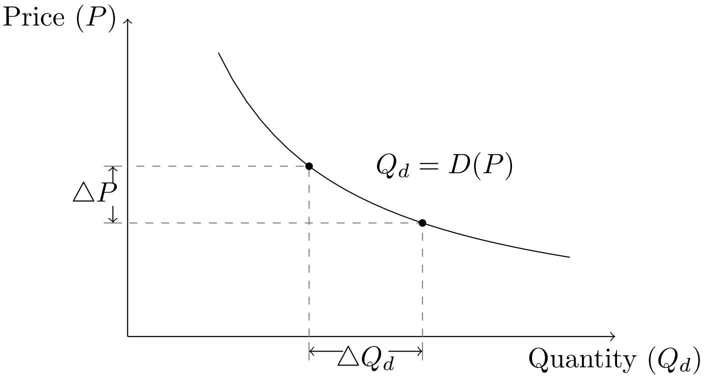

Figure 2.10: The change of demand ΔQ_d due to the change of price ΔP.

An illustrative example is shown in Figure 2.10. Here we use the market demand function $Q_d = D(P)$ to illustrate the concept of price elasticity, although the same concept can also be applied to consumer demand function. The value of E_d is often negative due to the downward slopping of the demand curve.

When the demand function Q_d is differentiable, we can compute the "point-price elasticity" by taking derivative of the demand function at a particular price P:

$$E_d = \frac{P}{Q_d}\frac{\partial Q_d}{\partial P}. \tag{2.7}$$

Depending on the value of E_d, the demand can be classified into three types:

- *Elastic demand:* the demand changes significantly with the price and $E_d < -1$.
- *Inelastic demand:* the demand is not sensitive to price and $-1 < E_d < 0$.
- *Unitary elastic demand:* $E_d = -1$.

Notice that different parts of the same demand function can have different price elasticities. If a firm can adjust the price P to maximize its revenue PQ_d, then it will decrease the price when the market demand is elastic, increase the price when the market demand is inelastic, and do not change the price when the market demand is unitary elastic.

2.3 FIRM BEHAVIOR

In this section, we will take a deeper look at the firm, and discuss how the market supply function $Q_s = S(P)$ is derived as a result of the firm's cost minimization behavior.

2.3.1 TOTAL AND MARGINAL PRODUCTION COST

In a market, a firm will product products based on certain technologies and sell the products in the market. How much to produce depends both on the production costs and the selling price in the market. We will start by understanding the types and impacts of production costs.

First, we can classify the cost by *explicit costs* and *opportunity costs*. Explicit cost of a wireless service provider may involve the cost of purchasing and installing the network equipments as well as the salary of the engineers. Opportunity costs represents the income that the firm loses due to utilizing the resources for a particular purpose. For example, if a spectrum owner (such as AT&T) decides to offer cellular services over its licensed spectrum, it explicitly forgoes the income that it can earn by leasing the spectrum to a third party (such as Google). The production cost thus includes both the explicit and opportunity costs.

The production cost will be different depending on whether we consider *short-term* or *long-term*. In general, we have less production choices in the short run than in the long-run. For example, in the long run a wireless service provider may be able to choose which technology to use (CDMA, TDMA, or OFDMA) and how much spectrum to obtain (through auction or leasing). In the short run, however, both the technology and total spectrum (and thus the network capacity) are fixed, and the service provider can only change the resource allocation among different cells, users, frequency bands, and time slots. In this section, we will focus the discussions on the short-term production cost. The discussions can be similarly generalized to the long-term production cost.

The total cost includes two parts: the *fixed* cost and the *variable* cost. The fixed cost F is the amount that a firm needs to pay independent of the quantity produced. The variable cost $V(q)$ depends on the production quantity q.

Definition 2.8 The **total production cost** includes both the fixed cost and variable cost, i.e.,

$$C(q) = F + V(q). \tag{2.8}$$

We are also interested in how the total cost changes when the firm changes with production quantity.

Definition 2.9 The **marginal cost** measures how the total cost changes with the production quantity, i.e.,

$$MC(q) = \frac{\text{\% change in total production cost}}{\text{\% change in production quantity}} = \frac{\Delta C(q)}{\Delta q} = \frac{\Delta V(q)}{\Delta q}. \tag{2.9}$$

Notice that the fixed cost F does not affect the computation of marginal cost. When the variable cost function $V(q)$ is differentiable, we have

$$MC(q) = \frac{\partial C(q)}{\partial q} = \frac{\partial V(q)}{\partial q}. \tag{2.10}$$

2.3.2 COMPETITIVE FIRM'S SUPPLY FUNCTION

Next we derive the supply function of a *competitive* firm.

Definition 2.10 A **competitive firm** is price-taking and acts as if the market price is independent of the quantity produced and sold by the firm.

The competitive firm accurately reflects the reality when the firm faces many competitors in the same market. In this case, each firm's production decision is unlike to significantly change the total quantity available in the market, and thus will not significantly affect the market price. The total revenue of a competitive firm will be $P \cdot q$, where P is the market price and q is the production quantity. This is assuming that the produced quantity can always be sold at the fixed market price P. The firm wants to choose the production amount q to maximize its profit.

Definition 2.11 A competitive firm's **profit** is the difference between revenue and total cost, i.e.,

$$\pi(q) = P \cdot q - V(q) - F. \tag{2.11}$$

If the firm produces $q = 0$, then the total profit is $-F$. Here we assume that the fixed cost F is also the *sunk cost*, i.e., a cost that the firm cannot avoid. This means that a firm will only produce when the revenue is no less than the variable cost, i.e., $Pq \geq V(q)$. At the optimal choice of q^* that maximizes the profit, we have

$$P = \frac{\partial V(q)}{\partial q} = MC(q), \tag{2.12}$$

which means that the price equals to the magical cost.

As we change the market price P, the competitive firm's optimal production quantity q changes according to (2.12). The firm's supply function is thus the firms' marginal cost function as long as revenue is no smaller than the variable cost.

2.4 CHAPTER SUMMARY

In this chapter, we introduce the basics of economic theory, including the relationship of supply and demand, the consumer behavior model, and the firm behavior model. In particular, we show how the market supply and demand are derived based on the behaviors of individual consumers and competitive firms. This chapter serves as the basis for discussions of pricing models in later chapters.

CHAPTER 3

Social Optimal Pricing

This chapter will focus on the issue of social optimal pricing, where a service provider chooses prices to maximize the social welfare. This corresponds to the case, for example, where the service provider's interests are aligned with the regulator's interests through proper economic mechanisms. The basic approach of social optimal pricing is to formulate the problem as an optimization problem, and design a dual-based distributed algorithm for the distributed resource allocation. Here the dual variables have the interpretation of shadow prices in economics.

We will first introduce the theory background of convex optimization and dual-based algorithms, and then illustrate the theory through two examples: single cell wireless video streaming and multi-provider resource allocation.

3.1 THEORY: DUAL-BASED OPTIMIZATION

In this section, we will cover the basics of convex optimization and dual-optimization. We will focus on the topics that are most useful in recognizing and formulating convex optimization problems in wireless networks. We follow the discussions in (11; 12), where readers can find more in-depth discussions.

3.1.1 PRELIMS

We use the notation $\mathbb{R}^n$ to denote the set of all real n-vectors. Each vector in $\mathbb{R}^n$ is called a *point* of $\mathbb{R}^n$. When $n = 1$, we will write $\mathbb{R}^1$, i.e., the set of real 1-vectors or real numbers, as $\mathbb{R}$ for brevity. The notation $f : \mathbb{R}^n \to \mathbb{R}^m$ is used to denote a function on some *subset* of $\mathbb{R}^n$ (specifically, its *domain*, which we denote $\mathcal{D}(f)$) into the set $\mathbb{R}^m$. That is, a function $f : \mathbb{R}^n \to \mathbb{R}^m$ maps every real n-vectors *in its domain* $\mathcal{D}(f)$ into an m-vector.

Convex Sets

Suppose $\boldsymbol{x}_1 \neq \boldsymbol{x}_2$ are two distinct points in $\mathbb{R}^n$. Any point $\boldsymbol{y}$ on the *line* passing through $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ can be expressed as

$$\boldsymbol{y} = \theta \boldsymbol{x}_1 + (1-\theta)\boldsymbol{x}_2, \quad \text{for some } \theta \in \mathbb{R}.$$

The parameter value $\theta = 1$ corresponds to $\boldsymbol{y} = \boldsymbol{x}_1$, and $\theta = 0$ corresponds to $\boldsymbol{y} = \boldsymbol{x}_2$. Values of θ between 0 and 1 correspond to the (closed) *line segment* between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$.

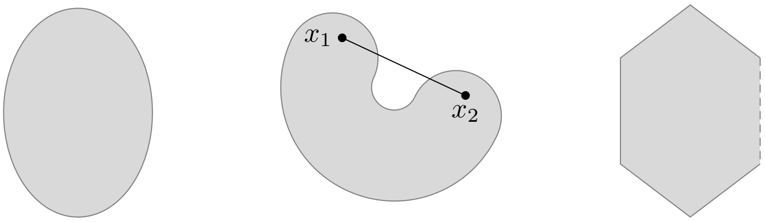

Figure 3.1: Some simple convex and nonconvex sets. (I) The ellipsoid, which includes its boundary (shown as solid curves), is convex. (II) The kidney shaped set is not convex, since the line segment between the points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is not entirely contained in the set. (III) The hexagon which contains some boundary points but not all (the dotted boundary points are not included), is not convex.

A point on the line passing through $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is referred to as an *affine combination* of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. A point on the line segment between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is referred to as a *convex combination* of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, which can be equivalently expressed as

$$\boldsymbol{y} = \theta_1 \boldsymbol{x}_1 + \theta_2 \boldsymbol{x}_2,$$

with $\theta_1 + \theta_2 = 1$ and $\theta_i \geq 0, i = 1, 2$.

A nonempty set $\mathcal{X} \subseteq \mathbb{R}^n$ is *convex* if the line segment between any two points (i.e., convex combinations of any two points) in $\mathcal{X}$ lies in $\mathcal{X}$. Specifically,

Definition 3.1 Convex Set. A nonempty set $\mathcal{X} \subseteq \mathbb{R}^n$ is *convex* if for any $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$ and any $\theta \in \mathbb{R}$ with $0 \leq \theta \leq 1$, we have

$$\theta \boldsymbol{x}_1 + (1 - \theta) \boldsymbol{x}_2 \in \mathcal{X}. \tag{3.1}$$

Geometrically, a set is convex if every point in the set can be reached by every other point, along an *inner straight path* between them, where inner means lying in the set. Obviously, any interval in $\mathbb{R}$ is a convex set. Figure 3.1 illustrates some simple convex and nonconvex sets in $\mathbb{R}^2$.

The concept of convex combination can be generalized to more than two points. Specifically, a convex combination of points $\boldsymbol{x}_1, ..., \boldsymbol{x}_k$ can be expressed as

$$\boldsymbol{y} = \theta_1 \boldsymbol{x}_1 + ... + \theta_k \boldsymbol{x}_k, \tag{3.2}$$

with $\theta_1 + ... + \theta_k = 1$ and $\theta_i \geq 0, i = 1, ..., k$. The condition for convex sets can be generalized accordingly: A nonempty set $\mathcal{X}$ is convex, if and only if for any $\boldsymbol{x}_1, ..., \boldsymbol{x}_k \in \mathcal{X}$,

$$\theta_1 \boldsymbol{x}_1 + ... + \theta_k \boldsymbol{x}_k \in \mathcal{X}, \tag{3.3}$$

when $\theta_1 + ... + \theta_k = 1$ and $\theta_i \geq 0, i = 1, ..., k$.

The sufficiency for the above condition is directly from the definition of convex sets. Next we show the necessity by backward induction. Let us introduce the intermediate vectors

$$\boldsymbol{z}_i \triangleq \frac{\theta_{i+1}\boldsymbol{x}_{i+1} + ... + \theta_k\boldsymbol{x}_k}{1 - \theta_1 - ... - \theta_i}, \quad i = 1, ..., k-1,$$

and $\boldsymbol{z}_k \triangleq \mathbf{0}$. Then we have $\boldsymbol{y} = \theta_1\boldsymbol{x}_1 + ... + \theta_k = \theta_1\boldsymbol{x}_1 + (1-\theta_1)\boldsymbol{z}_1$, and $\boldsymbol{z}_i = \tau_i\boldsymbol{x}_{i+1} + (1-\tau_i)\,\boldsymbol{z}_{i+1}$, where $\tau_i \triangleq \frac{\theta_{i+1}}{1-\theta_1-...-\theta_i} \in [0,1]$, $i = 1, ..., k-1$. To show that $\boldsymbol{y}$ lies in $\mathcal{X}$, we only need to show that $\boldsymbol{z}_1$ lies in $\mathcal{X}$, by the definition of convex sets. To show that $\boldsymbol{z}_1$ lies in $\mathcal{X}$, we need to show that $\boldsymbol{z}_2$ lies in $\mathcal{X}$. If we continue this iterative argument, then to show that $\boldsymbol{z}_{k-2}$ lies in $\mathcal{X}$, we need to show that $\boldsymbol{z}_{k-1}$ lies in $\mathcal{X}$. It is easy to see that $\boldsymbol{z}_{k-1} = \boldsymbol{x}_k$, and thus it lies in $\mathcal{X}$. Using backward induction, we conclude that $\boldsymbol{y}$ (i.e., any convex combination of points $\boldsymbol{x}_1, ..., \boldsymbol{x}_k$) lies in $\mathcal{X}$. Figure 3.2 illustrates a convex combination of four points $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3$ and $\boldsymbol{x}_4$ in $\mathbb{R}^3$, and the corresponding z_1, z_2, and z_3.

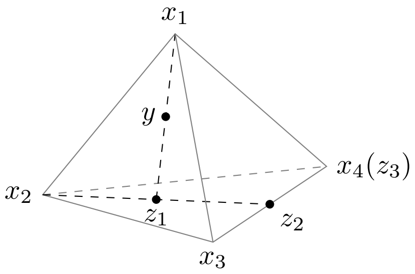

Figure 3.2: A convex combination of points $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3$ and $\boldsymbol{x}_4$ in a convex set $\mathcal{X}$. By introducing intermediate vectors $\boldsymbol{z}_i$, we have $\boldsymbol{z}_3 = \boldsymbol{x}_4 \in \mathcal{X} \Rightarrow \boldsymbol{z}_2 \in \mathcal{X} \Rightarrow \boldsymbol{z}_1 \in \mathcal{X} \Rightarrow \boldsymbol{y} \in \mathcal{X}$.

The *convex hull* of a set $\mathcal{X}$, denoted $\mathcal{H}(\mathcal{X})$, is the *smallest* convex set that contains $\mathcal{X}$. That is, it consists of the convex combinations of all points in $\mathcal{X}$. Specifically,

Definition 3.2 Convex Hull. The *convex hull* of a set $\mathcal{X}$, denoted $\mathcal{H}(\mathcal{X})$, is given by

$$\mathcal{H}(\mathcal{X}) \triangleq \{\theta_1\boldsymbol{x}_1 + ... + \theta_k\boldsymbol{x}_k \mid \theta_1 + ... + \theta_k = 1, \theta_i \geq 0, \boldsymbol{x}_i \in \mathcal{X}, i = 1, ..., k\}.$$

As the name suggests, the convex hull $\mathcal{H}(\mathcal{X})$ is always convex. Moreover, we have (i) $\mathcal{X} \subseteq \mathcal{H}(\mathcal{X})$, (ii) $\mathcal{X} = \mathcal{H}(\mathcal{X})$ if $\mathcal{X}$ is a convex set, and (iii) If $\mathcal{Y}$ is any convex set that contains $\mathcal{X}$, then $\mathcal{H}(\mathcal{X}) \subseteq \mathcal{Y}$. The last statement implies that the convex hull of a set $\mathcal{X}$ is the smallest convex set that contains $\mathcal{X}$. Figure 3.3 illustrates the convex hulls of some simple sets in $\mathbb{R}^2$.

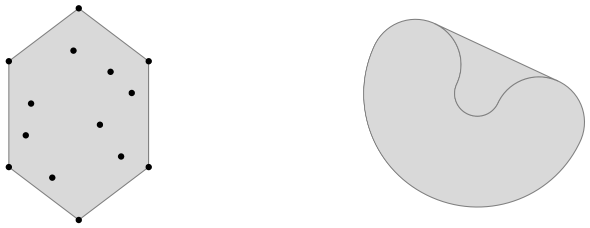

Figure 3.3: The convex hulls of two simple sets in $\mathbb{R}^2$. (I) The convex hull of a set of discrete points (shown as dots) is the pentagon (shown shaded). (II) The convex hull of the kidney shaped set in Figure 3.1 is the shaded set.

Operations Preserving Convexity of Sets

Now we describe some simple operations that preserve the convexity of sets, or allow us to construct new convex sets.

1. *Intersection*: If $\mathcal{X}_1, ..., \mathcal{X}_k$ are convex sets, then $\mathcal{X} \triangleq \mathcal{X}_1 \cap ... \cap \mathcal{X}_k$ is convex.

2. *Affine mapping*: Suppose $\mathcal{X}$ is a subset of $\mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{m\times n}$, and $\boldsymbol{b} \in \mathbb{R}^m$. Define a new set $\mathcal{Y} \subseteq \mathbb{R}^m$ by
$$\mathcal{Y} \triangleq \{\mathbf{A}\boldsymbol{x} + \boldsymbol{b} \mid \boldsymbol{x} \in \mathcal{X}\}.$$
Then if $\mathcal{X}$ is convex, so is $\mathcal{Y}$. The affine mapping operation generalizes a lot of common operations including scaling, translation, summation, projection, etc.

Convex Functions

We consider a scalar-valued function $f : \mathbb{R}^n \to \mathbb{R}$, which maps every real n-vectors in its domain $\mathcal{D}(f)$ into a real number in $\mathbb{R}$. In order to distinguish a function from a variable, we will use the notation $f(\cdot)$ to denote a function f whenever there is a need.

In the context of optimization, one of the most important properties of a function $f(\cdot)$ is its convexity (or concavity). Specifically,

Definition 3.3 Convex Function. A function $f : \mathbb{R}^n \to \mathbb{R}$ is *convex*, if $\mathcal{D}(f)$ is a convex set and if for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{D}(f)$, and $\theta \in \mathbb{R}$ with $0 \leq \theta \leq 1$, we have

$$f(\theta\boldsymbol{x} + (1-\theta)\boldsymbol{y}) \leq \theta f(\boldsymbol{x}) + (1-\theta) f(\boldsymbol{y}). \tag{3.4}$$

A function $f(\cdot)$ is *strictly convex* if the strict inequality holds in (3.4) whenever $\boldsymbol{x} \neq \boldsymbol{y}$ and $0 < \theta < 1$. We say $f(\cdot)$ is (strictly) *concave* if $-f(\cdot)$ is (strictly) convex. Note that a function can be neither convex nor concave. As a simple example, consider a scalar-valued

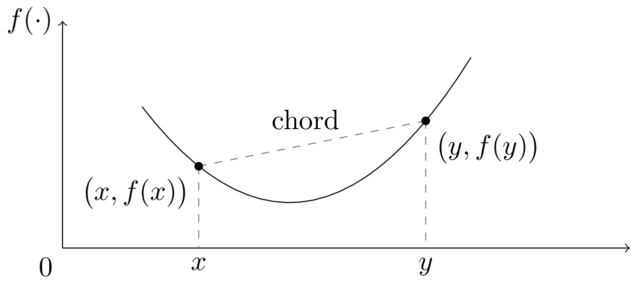

Figure 3.4: An illustration of a convex function $f(\cdot)$ on $\mathbb{R}$. The chord (shown as dots) between points $(\boldsymbol{x}, f(\boldsymbol{x}))$ and $(\boldsymbol{y}, f(\boldsymbol{y}))$ on the graph of $f(\cdot)$ lies above the graph.

function $f(x) = x^3$ on $\mathbb{R}$. We can easily find that $f(\theta x + (1-\theta)y) \geq \theta f(x) + (1-\theta)f(y)$ when $x, y \leq 0$, and $f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y)$ when $x, y \geq 0$.

Geometrically, the inequality in (3.4) means that for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{D}(f)$, the line segment between points $(\boldsymbol{x}, f(\boldsymbol{x}))$ and $(\boldsymbol{y}, f(\boldsymbol{y}))$, which is called the *chord* from $\boldsymbol{x}$ to $\boldsymbol{y}$, lies above the graph of $f(\cdot)$. Figure 3.4 illustrates a simple convex function on $\mathbb{R}$.

As with convex sets, the condition in (3.4) can be generalized to the case of more than two points: A function $f(\cdot)$ is convex, if and only if $\mathcal{D}(f)$ is convex and

$$f(\theta_1 \boldsymbol{x}_1 + ... + \theta_k \boldsymbol{x}_k) \leq \theta_1 f(\boldsymbol{x}_1) + ... + \theta_k f(\boldsymbol{x}_k), \tag{3.5}$$

for any $\boldsymbol{x}_1, ..., \boldsymbol{x}_k \in \mathcal{D}(f)$, when $\theta_1 + ... + \theta_k = 1$ and $\theta_i \geq 0, i = 1, ..., k$. The sufficiency is directly from the definition of convex functions. The necessity can be proved by a similar backward induction for the necessity of (3.3) for convex sets.

First-order Conditions for Convex Functions

For a scalar-valued function $f : \mathbb{R}^n \to \mathbb{R}$, the derivative or *gradient* of $f(\cdot)$ at a point $\boldsymbol{x} \in \mathcal{D}(f)$, denoted by $\nabla f(\boldsymbol{x})$, is an n-vector with the ith component given by

$$\nabla f(\boldsymbol{x})_i = \frac{\partial f(\boldsymbol{x})}{\partial x_i}, \ \ i = 1, ..., n, \tag{3.6}$$

provided the partial derivatives exist. Here x_i is the i-th coordinate of the vector $\boldsymbol{x}$. If the partial derivatives exist at $\boldsymbol{x}$ for all coordinates x_i, we say $f(\cdot)$ is *differentiable* at $\boldsymbol{x}$. The function $f(\cdot)$ is differentiable (everywhere in its domain) if $\mathcal{D}(f)$ is open (i.e., it contains no boundary points), and it is differentiable at every point in $\mathcal{D}(f)$.

A differentiable function $f(\cdot)$ is convex, if and only if $\mathcal{D}(f)$ is convex and

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}), \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{D}(f). \tag{3.7}$$

This inequality is called the *first-order condition* for convex functions.

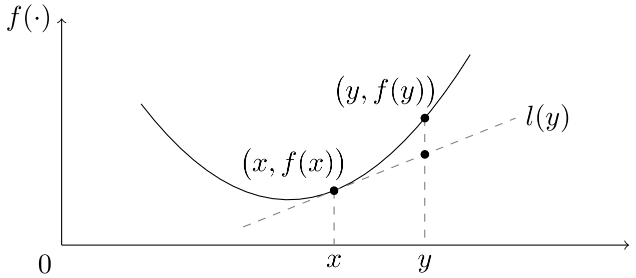

Figure 3.5: The first-order condition for a convex function $f(\cdot)$ on $\mathbb{R}$. The line $l(\boldsymbol{y}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x})$ (shown as dots) lies under the graph of $f(\cdot)$.

Geometrically, the first-order condition (3.7) means that the line passing through $(\boldsymbol{x}, f(\boldsymbol{x}))$ along the gradient of $f(\cdot)$ at $\boldsymbol{x}$, i.e., $\nabla f(\boldsymbol{x})$, lies under the graph of $f(\cdot)$. Figure 3.5 illustrates the first-order condition for a simple convex function on $\mathbb{R}$.

The first-order condition (3.7) is the most important property of convex functions, and plays an important role in convex optimization. It implies that from local information about a convex function (i.e., its value $f(\boldsymbol{x})$ and gradient $\nabla f(\boldsymbol{x})$ at a point $\boldsymbol{x}$), we can derive global information (i.e., a global underestimator of $f(\cdot)$ at any point). As one simple example, the inequality (3.7) shows that if $\nabla f(\boldsymbol{x}) = \mathbf{0}$ (i.e., $\nabla f(\boldsymbol{x})_i = 0$, $i = 1, ..., n$), then $\boldsymbol{x}$ is a global minimizer of $f(\cdot)$, since $f(\boldsymbol{y}) \geq f(\boldsymbol{x})$, $\forall \boldsymbol{y} \in \mathcal{D}(f)$.

Second-order Conditions for Convex Functions

The second derivative or *Hessian matrix* of a scalar-valued function $f(\cdot)$ at a point $\boldsymbol{x} \in \mathcal{D}(f)$, denoted by $\nabla^2 f(\boldsymbol{x})$, is an $n \times n$ matrix, given by

$$\nabla^2 f(\boldsymbol{x})_{ij} = \frac{\partial^2 f(\boldsymbol{x})}{\partial x_i \partial x_j}, \ i = 1, ..., n, j = 1, ..., n, \tag{3.8}$$

provided that $f(\cdot)$ is twice differentiable at $\boldsymbol{x}$. We say $f(\cdot)$ is twice differentiable (everywhere in its domain) if $\mathcal{D}(f)$ is open, and it is twice differentiable at every point in $\mathcal{D}(f)$.

A twice differentiable $f(\cdot)$ is convex, if and only if $\mathcal{D}(f)$ is convex and

$$\nabla^2 f(\boldsymbol{x}) \succeq 0, \quad \forall \boldsymbol{x} \in \mathcal{D}(f), \tag{3.9}$$

that is, if its Hessian matrix is positive semidefinite. This inequality is referred to as the *second-order condition* for convex functions.

For a scalar-valued function $f(\cdot)$ on $\mathbb{R}$, the inequality (3.9) reduces to the simple condition $f''(x) \geq 0$, which means that the gradient is nondecreasing. Geometrically, the second-order condition (3.9) can be interpreted as the requirement that the graph of the function have positive (upward) curvature at $\boldsymbol{x}$.

Operations Preserving Convexity of Functions

Now we describe some simple operations that preserve convexity (or concavity) of functions, or allow us to construct new convex and concave functions.

1. *Nonnegative weighted sums:* Suppose $f_1(\cdot), ..., f_k(\cdot)$ are convex, and $\theta_1, ..., \theta_k \geq 0$. Define a new function by

$$f(\boldsymbol{x}) \triangleq \theta_1 f_1(\boldsymbol{x}) + ... + \theta_k f_k(\boldsymbol{x}),$$

   with $\mathcal{D}(f) = \mathcal{D}(f_1) \cap ... \cap \mathcal{D}(f_k)$. Then $f(\cdot)$ is also convex.

2. *Composition with an affine mapping:* Suppose $f(\cdot)$ is a function on $\mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{n\times m}$, and $\boldsymbol{b} \in \mathbb{R}^n$. Define a new function by

$$g(\boldsymbol{x}) \triangleq f(\mathbf{A}\boldsymbol{x} + \boldsymbol{b}),$$

   with $\mathcal{D}(g) = \{\boldsymbol{x} \in \mathbb{R}^m | \mathbf{A}\boldsymbol{x} + \boldsymbol{b} \in \mathcal{D}(f)\}$. Then if $f(\cdot)$ is convex, so is $g(\cdot)$.

3. *Point-wise maximum:* Suppose $f_1(\cdot), ..., f_k(\cdot)$ are convex. Define a new function by their pointwise maximum

$$f(\boldsymbol{x}) \triangleq \max\{f_1(\boldsymbol{x}), ..., f_k(\boldsymbol{x})\},$$

   with $\mathcal{D}(f) = \mathcal{D}(f_1) \cap ... \cap \mathcal{D}(f_k)$. Then $f(\cdot)$ is also convex.

3.1.2 CONVEX OPTIMIZATION

A *mathematical optimization problem* usually describe the problem of finding a point over a feasible set that minimizes an objective function. It has the form

$$\begin{array}{rl} \text{minimize} & f(\boldsymbol{x}) \\ \text{subject to} & f_i(\boldsymbol{x}) \leq 0, \;\; i = 1, ..., m. \end{array} \tag{3.10}$$

The function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is called the *objective function* (or cost function), the functions $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$, $i = 1, ..., m$, are called the (inequality) *constraint functions*, and the point $\boldsymbol{x} \in \mathbb{R}^n$ is the *optimization variable* of the problem.[1] The domain of an optimization problem (3.10) is the intersection of the objective function's domain and all constraint functions' domains, denoted by $\mathcal{D}(\text{p})$ or $\mathcal{D}$ simply, i.e., $\mathcal{D} \triangleq \mathcal{D}(f) \cap \mathcal{D}(f_1) \cap ... \cap \mathcal{D}(f_m)$.

A point $\boldsymbol{x}$ is *feasible* for an optimization problem (3.10) if it satisfies all constraints $f_i(\boldsymbol{x}) \leq 0, i = 1, ..., m$. The set of all feasible points is called the *constraint set* or *feasible set* for the optimization problem (3.10), denoted by $\mathcal{C}(\text{p})$ or $\mathcal{C}$ simply, i.e.,

$$\mathcal{C} \triangleq \{\boldsymbol{x} \mid \boldsymbol{x} \in \mathcal{D}, f_i(\boldsymbol{x}) \leq 0, i = 1, ..., m\}. \tag{3.11}$$

[1] Note that any equality constraint (e.g., $h(\boldsymbol{x}) = 0$) can be represented by two inequality constraints equivalently, i.e., $h(\boldsymbol{x}) \geq 0$ and $-h(\boldsymbol{x}) \geq 0$. Therefore, we consider the inequality constraint only without loss of generality.

A feasible point $\boldsymbol{x} \in \mathcal{C}$ is called (*globally*) *optimal* (a global minimizer), or a solution of the problem (3.10), if it has the smallest objective value among all feasible points, i.e.,

$$f(\boldsymbol{x}) \leq f(\boldsymbol{z}), \ \forall \boldsymbol{z} \in \mathcal{C}. \tag{3.12}$$

A feasible point $\boldsymbol{x} \in \mathcal{C}$ is *locally optimal* (a local minimizer), if it is no worse than its feasible neighbors, that is, if there is an $\epsilon > 0$ such that

$$f(\boldsymbol{x}) \leq f(\boldsymbol{z}), \ \forall \boldsymbol{z} \in \mathcal{C} \text{ with } ||\boldsymbol{z} - \boldsymbol{x}|| \leq \epsilon, \tag{3.13}$$

where $||\boldsymbol{x}|| \triangleq \sqrt{\boldsymbol{x}^T \boldsymbol{x}}$ is the standard Euclidean norm of a vector $\boldsymbol{x}$. That is, $||\boldsymbol{z} - \boldsymbol{x}||$ is the Euclidean distance between points $\boldsymbol{z}$ and $\boldsymbol{x}$.

We are interested in a class of optimization problems called *convex optimization problems*, where the objective and constraint functions are convex. This implies the domain $\mathcal{D}$ and the constraint set $\mathcal{C}$ are both convex. A fundamental property of convex optimization problem is that: *a local minimizer is also a global minimizer. If in addition the objective function is strictly convex, then the global minimizer is unique.*

Unconstrained Convex Optimization

If there is no constraint (i.e., $m = 0$) in the problem (3.10), we say it is an *unconstrained* optimization. That is, an *unconstrained convex optimization* problem is one of the form

$$\text{minimize} \quad f(\boldsymbol{x}) \tag{3.14}$$

where the objective function $f(\cdot)$ is convex. Obviously, in an unconstrained convex optimization, the constraint set is the domain of $f(\cdot)$, i.e., $\mathcal{C} = \mathcal{D}(f)$.

The following lemma characterizes the optimality conditions for an unconstrained convex optimization, that is, the necessary and sufficient conditions for a feasible point to be (globally) optimal.

Lemma 3.4 *Suppose the objective function $f(\cdot)$ is convex and differentiable. A feasible point $\boldsymbol{x}^* \in \mathcal{C}$ is a global minimizer of $f(\cdot)$ or a solution of (3.14) if and only if*

$$\nabla f(\boldsymbol{x}^*) = \mathbf{0}, \tag{3.15}$$

that is, $\nabla f(\boldsymbol{x}^*)_i = \frac{\partial f(\boldsymbol{x}^*)}{\partial x_i} = 0$, $i = 1, ..., n$.

We first show the necessity of condition (3.15), that is, if $\boldsymbol{x}^*$ is a global minimizer of $f(\cdot)$, then $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$. To see this, we consider a point which is arbitrarily closed to $\boldsymbol{x}^*$, i.e., $\boldsymbol{x}^* - \Delta \boldsymbol{x}$ where $\Delta \boldsymbol{x} \to \mathbf{0}$. Using the first order Taylor series expansions, we have

$$f(\boldsymbol{x}^* - \Delta \boldsymbol{x}) - f(\boldsymbol{x}^*) \doteq -\nabla f(\boldsymbol{x}^*)^T \Delta \boldsymbol{x}, \text{ when } \Delta \boldsymbol{x} \to \mathbf{0}.$$

By setting $\Delta \boldsymbol{x} = \epsilon \cdot \nabla f(\boldsymbol{x}^*)$ where ϵ is positive and arbitrarily close to 0 (i.e., $\Delta \boldsymbol{x}$ equals to an infinitely scaled gradient of $f(\cdot)$ at point $\boldsymbol{x}^*$), we have $-\epsilon \cdot \nabla f(\boldsymbol{x}^*)^T \nabla f(\boldsymbol{x}^*) \geq 0$ or equivalently $\nabla f(\boldsymbol{x}^*)^T \nabla f(\boldsymbol{x}^*) \leq 0$. It follows directly that $\frac{\partial f(\boldsymbol{x}^*)}{\partial x_i} = 0$, $i = 1, ..., n$.

We then show the sufficiency of condition (3.15), that is, if $f(\cdot)$ is a convex function and $\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}$, then $\boldsymbol{x}^*$ is a global minimizer of $f(\cdot)$. Using the first-order condition (3.7) for convex functions, we have

$$f(\boldsymbol{y}) - f(\boldsymbol{x}^*) \geq \nabla f(\boldsymbol{x}^*)^T (\boldsymbol{y} - \boldsymbol{x}^*), \quad \forall \boldsymbol{y} \in \mathcal{C}.$$

If $\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}$, we have $f(\boldsymbol{y}) \geq f(\boldsymbol{x}^*)$ for any $\boldsymbol{y} \in \mathcal{C}$. So $\boldsymbol{x}^*$ is a global minimizer of $f(\cdot)$.

Note that the above optimality condition (3.15) only holds for convex objective functions. For a general objective function $f(\cdot)$, the necessary and sufficient conditions for a point $\boldsymbol{x}^*$ to be *locally* optimal (a local minimizer) are

$$\nabla f(\boldsymbol{x}^*) = \boldsymbol{0} \;\text{ and }\; \nabla^2 f(\boldsymbol{x}^*) \succeq 0. \tag{3.16}$$

For detailed discussions, please refer to (11; 12).

We now discuss the computational methods for solving an unconstrained optimization. By the optimality condition (3.15), solving the problem (3.14) is the same as finding a solution of $\nabla f(\boldsymbol{x}) = \boldsymbol{0}$, i.e., a set of n equations $\frac{\partial f(\boldsymbol{x})}{\partial x_i} = 0$, $i = 1, ..., n$. In a few special cases, it is possible to solve these n equations analytically; but more generally the problem must be solved by an iterative algorithm. That is, we want to find an algorithm that computes a sequence of feasible points $\boldsymbol{x}^{(0)}, \boldsymbol{x}^{(1)}, ...$ with $f(\boldsymbol{x}^{(k)}) \rightarrow f(\boldsymbol{x}^*)$ as $k \rightarrow \infty$. Such a sequence of points is called a *minimizing sequence* for the problem (3.14). An algorithm is said to be *iteratively descent*, if it successively generates points $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, ...,$ (from an initial point $\boldsymbol{x}^{(0)}$) such that $f(\cdot)$ is decreasing at each iteration, i.e., $f(\boldsymbol{x}^{(k+1)}) < f(\boldsymbol{x}^{(k)})$, $\forall k$.

We focus on the most popular gradient-based algorithms, which have the form

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \gamma^{(k)} \boldsymbol{d}^{(k)},$$

where $\gamma^{(k)} > 0$ is a positive scalar called the *step size* or *step length* at iteration k, and $\boldsymbol{d}^{(k)}$ is a gradient-based n-vector called the *step* or *search direction* at iteration k. The iterative descent property requires that $\nabla f(\boldsymbol{x}^{(k)})^T \boldsymbol{d}^{(k)} < 0$, otherwise, $f(\boldsymbol{x}^{(k+1)}) - f(\boldsymbol{x}^{(k)}) \geq \gamma^{(k)} \nabla f(\boldsymbol{x}^{(k)})^T \boldsymbol{d}^{(k)} \geq 0$. Two widely used gradient-based algorithms are presented below. For more algorithms, please refer to (12).

1. *Gradient Descent Method:*

   $$\boldsymbol{d}^{(k)} \triangleq -\nabla f(\boldsymbol{x}^{(k)}).$$

   That is, the search direction at each iteration k is the negative gradient at $\boldsymbol{x}^{(k)}$.

2. *Newton's Method:*

   $$\boldsymbol{d}^{(k)} \triangleq -\big(\nabla^2 f(\boldsymbol{x}^{(k)})\big)^{-1} \nabla f(\boldsymbol{x}^{(k)}).$$

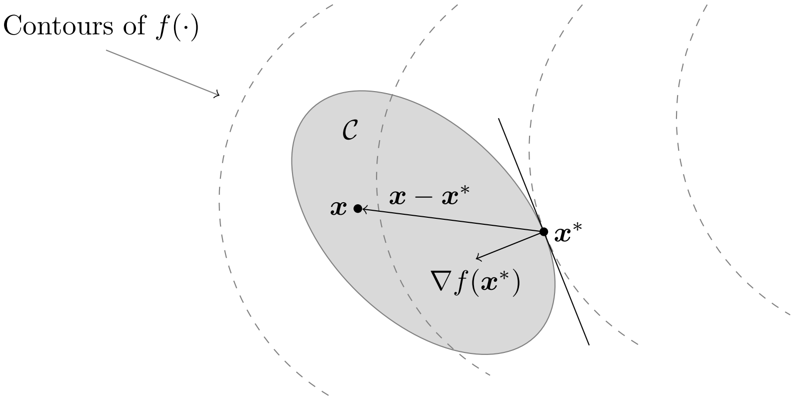

Figure 3.6: Geometric interpretation of the optimality condition in (3.18).The constraint set $\mathcal{C}$ is shown shaded. The gradient $\nabla f(\boldsymbol{x}^*)$ makes an angle less than or equal to 90 degrees with all feasible variations $\boldsymbol{x}-\boldsymbol{x}^*$.

Constrained Convex Optimization

A *constrained convex optimization* problem, or just convex optimization, is one of the form

$$\begin{array}{rl} \text{minimize} & f(\boldsymbol{x}) \\ \text{subject to} & f_i(\boldsymbol{x}) \leq 0,\ i=1,...,m, \end{array} \tag{3.17}$$

where the objective function $f(\cdot)$ and the constraint functions $f_i(\cdot)$ are convex. According to (3.11), the constraint set $\mathcal{C}$ is also convex.

The following lemma characterizes the optimality conditions for an constrained convex optimization, that is, the necessary and sufficient conditions for a feasible point to be (globally) optimal.

Lemma 3.5 *Suppose the objective function $f(\cdot)$ is convex and differentiable. A feasible point $\boldsymbol{x}^* \in \mathcal{C}$ is a global minimizer of $f(\cdot)$ or a solution of (3.17) if and only if*

$$\nabla f(\boldsymbol{x}^*)^T(\boldsymbol{x}-\boldsymbol{x}^*) \geq 0, \quad \forall \boldsymbol{x} \in \mathcal{C}. \tag{3.18}$$

To better understand the optimality condition (3.18), we illustrate it geometrically in Figure 3.6. At a minimizer $\boldsymbol{x}^*$, the gradient $\nabla f(\boldsymbol{x}^*)$ makes an angle less than or equal to 90 degrees with all feasible variations $\boldsymbol{x}-\boldsymbol{x}^*$, so that $\nabla f(\boldsymbol{x}^*)^T(\boldsymbol{x}-\boldsymbol{x}^*) \geq 0$, $\forall \boldsymbol{x} \in \mathcal{C}$.

We first show the necessity of condition (3.18), that is, if $\boldsymbol{x}^*$ is a global minimizer of a convex function $f(\cdot)$, then $\nabla f(\boldsymbol{x}^*)^T(\boldsymbol{x}-\boldsymbol{x}^*) \geq 0$, $\forall \boldsymbol{x} \in \mathcal{C}$. To see this, we consider a point $\boldsymbol{z}$ on the line segment between $\boldsymbol{x}^*$ and any feasible point $\boldsymbol{x} \in \mathcal{C}$, i.e., $\boldsymbol{z} = (1-\epsilon)\boldsymbol{x}^* + \epsilon\boldsymbol{x}$

where $0 \leq \epsilon \leq 1$. Since the constraint set $\mathcal{C}$ is convex, $\boldsymbol{z}$ is feasible, i.e., $\boldsymbol{z} \in \mathcal{C}$. Using the first order Taylor series expansions, we have

$$f(\boldsymbol{z}) - f(\boldsymbol{x}^*) \doteq \epsilon \nabla f(\boldsymbol{x}^*)^T(\boldsymbol{x} - \boldsymbol{x}^*), \text{ when } \epsilon \rightarrow 0.$$

Suppose $\nabla f(\boldsymbol{x}^*)^T(\boldsymbol{x} - \boldsymbol{x}^*) < 0$ for some $\boldsymbol{x} \in \mathcal{C}$ by contradiction. Then we have for sufficiently small $\epsilon > 0$, $f(\boldsymbol{z}) < f(\boldsymbol{x}^*)$, which violates the optimality of $\boldsymbol{x}^*$.

We then show the sufficiency of condition (3.15), that is, if $f(\cdot)$ is a convex function and $\nabla f(\boldsymbol{x}^*)^T(\boldsymbol{x} - \boldsymbol{x}^*) \geq 0$, then $\boldsymbol{x}^*$ is a global minimizer of a convex function $f(\cdot)$. Using the first-order condition (3.7) for convex functions, we have

$$f(\boldsymbol{x}) - f(\boldsymbol{x}^*) \geq \nabla f(\boldsymbol{x}^*)^T(\boldsymbol{x} - \boldsymbol{x}^*) \geq 0, \quad \forall \boldsymbol{x} \in \mathcal{C}.$$

So $\boldsymbol{x}^*$ is a global minimizer of $f(\cdot)$.

We now turn to the computational methods for solving a constrained optimization problem. Although there is a great variety of algorithms for this problem, we will restrict ourselves to a limited class of methods that generate a minimizing sequence of feasible $\boldsymbol{x}^{(k)}$ by searching along descent directions (i.e., iterative descent). Similarly, we focus on the most popular gradient-based algorithms, which have the form

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \gamma^{(k)} \boldsymbol{d}^{(k)}.$$

Two widely used gradient-based algorithms are listed below.

1. *The Conditional Gradient Method:*

$$\boldsymbol{d}^{(k)} \triangleq \overline{\boldsymbol{x}}^{(k)} - \boldsymbol{x}^{(k)},$$

where $\overline{\boldsymbol{x}}^{(k)} \triangleq \arg\max_{\boldsymbol{x} \in \mathcal{C}} \nabla f(\boldsymbol{x}^{(k)})^T(\boldsymbol{x} - \boldsymbol{x}^{(k)})$ subject to $\nabla f(\boldsymbol{x}^{(k)})^T(\boldsymbol{x} - \boldsymbol{x}^{(k)}) < 0$. If $\nabla f(\boldsymbol{x}^{(k)})^T(\boldsymbol{x} - \boldsymbol{x}^{(k)}) \geq 0$ for all $\boldsymbol{x} \in \mathcal{C}$, then $\boldsymbol{x}^{(k)}$ is the optimal solution by (3.18).

2. *Gradient Projection Method:*

$$\boldsymbol{d}^{(k)} \triangleq \overline{\boldsymbol{x}}^{(k)} - \boldsymbol{x}^{(k)},$$

where $\overline{\boldsymbol{x}}^{(k)}$ is given by $\overline{\boldsymbol{x}}^{(k)} \triangleq \left[\boldsymbol{x}^{(k)} - s^{(k)} \nabla f(\boldsymbol{x}^{(k)})\right]^+$. Here $[\cdot]^+$ denotes projection on the constraint set $\mathcal{C}$, and $s^{(k)}$ is a positive scalar.

Some Special Cases

Convex optimization techniques have found graceful properties from theoretical analysis to computational methodology, and important applications in various engineering disciplines. One of the most fundamental properties is that: *a locally optimal point is also globally optimal.* The optimality conditions (Lemmas 3.4 and 3.5) are quite clear and intuitive. Also, a convex optimization problem is easy to solve numerically by efficient computational algorithms, such as the primal-dual interior-point methods (13).

Thus, it is highly desirable to formulate a practical problem into a convex optimization. Special cases of convex optimization include Linear Programming (LP), convex Quadratic Programming (QP), Second Order Cone Programming (SOCP), and Semidefinite Programming (SDP). However, many practical problems do not have obvious convex optimization formulations. In this case, it is desirable, if possible, to convert a non-convex optimization problem to a convex one. There are many ways to achieve this. Here we will explain the method of *Geometric Programming* (GP), which is a very nice example to illustrate this transition (14; 15). GP in standard form is a non-convex optimization problem, and it can be readily turned into a convex optimization problem.

Generally, GP is a set of special optimization problems with the standard form

$$\begin{array}{rl} \text{minimize} & f(\boldsymbol{x}) \\ \text{subject to} & f_i(\boldsymbol{x}) \leq 1, \;\; i = 1, ..., m, \end{array} \tag{3.19}$$

where both the objective and constraint functions are *posynomial*, i.e.,

$$f(\boldsymbol{x}) = \sum_{k=1}^{K} d_k x_1^{a_{1k}} x_2^{a_{2k}} ... x_n^{a_{nk}},$$

and

$$f_i(\boldsymbol{x}) = \sum_{k=1}^{K} d_k^{(i)} x_1^{a_{1k}^{(i)}} x_2^{a_{2k}^{(i)}} ... x_n^{a_{nk}^{(i)}}, \quad i = 1, ..., m,$$

where $d_k \geq 0$, $d_k^{(i)} \geq 0$, $a_{jk} \in \mathbb{R}$, and $a_{jk}^{(i)} \in \mathbb{R}$, $k = 1, ..., K, i = 1, ..., m, j = 1, ..., n$.

GP in standard form is usually not a convex optimization problem, because posynomials are usually not convex functions.[2] However, with a logarithmic change of all the variables and multiplicative constants: $y_n = \log x_n$, $b_k = \log d_k$ and $b_k^{(i)} = \log d_k^{(i)}$, we can turn it into an equivalent convex optimization:

$$\begin{array}{rl} \text{minimize} & g(\boldsymbol{y}) \triangleq \log \sum_{k=1}^{K} \exp\left(\boldsymbol{a}_k^T \boldsymbol{y} + b_k\right) \\ \text{subject to} & g_i(\boldsymbol{y}) \triangleq \log \sum_{k=1}^{K} \exp\left(\boldsymbol{a}_k^{(i)T} \boldsymbol{y} + b_k^{(i)}\right) \leq 0, \;\; i = 1, ..., m, \end{array} \tag{3.20}$$

where $\boldsymbol{a}_k \triangleq (a_{1k}, ..., a_{nk})$ and $\boldsymbol{a}_k^{(i)} \triangleq (a_{1k}^{(i)}, ..., a_{nk}^{(i)})$, $k = 1, ..., K, i = 1, ..., m$.

To show that (3.20) is indeed a convex optimization problem, we need to show that the objective and inequality constraint functions are convex in $\boldsymbol{y}$. This convexity property can be readily verified through a positive-definiteness test of the Hessian. A more illuminating verification uses the fact that the objective and constraint functions in (3.20) are indeed compositions of a *log-sum-exp* function $f(\boldsymbol{x}) = \log \sum_{i=1}^{n} e^{x_i}$ (which is apparently convex) with certain affine functions.

[2] To see this, we consider a simple GP with a scalar-valued objective function on $\mathbb{R}$: $f(x) = x^3$, which is obviously non-convex.

3.1.3 DUALITY PRINCIPLE

Now we introduce the *Lagrangian duality*, which plays a central role in convex optimization. By duality principle, an optimization problem (which we refer to as the *primal problem*) can usually be converted into a (Lagrange) dual form, which is termed a (Lagrange) *dual problem*. The solution of the dual problem provides a lower bound to the solution of the primal problem. In addition if the primal problem is convex and satisfies a constraint qualification, then the value of an optimal solution of the primal problem is given by the dual problem (11).

Lagrange Dual Functions

The basic idea in Lagrangian duality is to take the constraints into account by adding the objective function with a weighted sum of the constraint functions. The weight associated with each constraint function $f_i(\boldsymbol{x})$ is referred to as the *Lagrange multiplier*, denoted by λ_i. The vector $\boldsymbol{\lambda} \triangleq (\lambda_1, ..., \lambda_m)$ is called the *dual variable* or *Lagrange multiplier vector*. The *Lagrangian function* and *dual function* for problem (3.17) are defined as follows.[3]

Definition 3.6 Lagrangian Function. The *Lagrangian function* (or just *Lagrangian*) $L : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is defined as

$$L(\boldsymbol{x}, \boldsymbol{\lambda}) \triangleq f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{x}), \tag{3.21}$$

with the domain $\mathcal{D}(L) = \mathcal{D} \times \mathbb{R}^m$, where $\mathcal{D} = \mathcal{D}(f) \cap \mathcal{D}(f_1) \cap ... \cap \mathcal{D}(f_m)$ is the domain of the optimization problem (3.17).

Definition 3.7 Dual Function. The (*Lagrange*) *dual function* $g : \mathbb{R}^m \to \mathbb{R}$ is defined as the minimum value of the Lagrangian over $\boldsymbol{x}$:

$$g(\boldsymbol{\lambda}) \triangleq \inf_{\boldsymbol{x} \in \mathcal{D}} L(\boldsymbol{x}, \boldsymbol{\lambda}) = \inf_{\boldsymbol{x} \in \mathcal{D}} \left(f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{x}) \right). \tag{3.22}$$

The dual function $g(\cdot)$ is always concave (even when the original problem (3.17) is not convex), since it is the pointwise infimum of a family of affine functions of $\boldsymbol{\lambda}$.

One of the most important properties for the dual function $g(\cdot)$ is that it yields lower bounds on the optimal value $f(\boldsymbol{x}^*)$ of the problem (3.17): for any $\boldsymbol{\lambda} \succeq \mathbf{0}$ we have

$$g(\boldsymbol{\lambda}) = \inf_{\boldsymbol{x} \in \mathcal{D}} L(\boldsymbol{x}, \boldsymbol{\lambda}) \leq \inf_{\boldsymbol{x} \in \mathcal{C}} L(\boldsymbol{x}, \boldsymbol{\lambda}) \leq L(\boldsymbol{x}^*, \boldsymbol{\lambda}) \leq f(\boldsymbol{x}^*). \tag{3.23}$$

[3] Note the following discussions are not only applicable to convex optimization problems, but also to non-convex optimization problems.

The first inequality follows because the constraint set is a subset of the domain, i.e., $\mathcal{C} \subseteq \mathcal{D}$, the second inequality follows because the optimal point $\boldsymbol{x}^*$ lies in the constraint set, i.e., $\boldsymbol{x}^* \in \mathcal{C}$, and the last inequality follows because $f_i(\boldsymbol{x}) \leq 0$ for any feasible $\boldsymbol{x} \in \mathcal{C}$.

Lagrange Dual Problems

As shown in (3.23), the Lagrange dual function $g(\cdot)$ yields a lower bound on the optimal value $f(\boldsymbol{x}^*)$ of the optimization problem (3.17), and how far the dual function $g(\cdot)$ is apart from the optimal value $f(\boldsymbol{x}^*)$ essentially depends on the dual variable $\boldsymbol{\lambda}$. Thus, a natural question is: What is the best lower bound that can be obtained from the Lagrange dual function? This leads to the following optimization problem

$$\begin{array}{rl} \text{maximize} & g(\boldsymbol{\lambda}) \\ \text{subject to} & \boldsymbol{\lambda} \succeq \mathbf{0}. \end{array} \tag{3.24}$$

The problem (3.24) is called the (*Lagrange*) *dual problem* associated with the problem (3.17), which we call the *primal problem* in this context. Obviously, the dual problem (3.24) is convex (even when the primal problem (3.17) is not convex), since the objective to be maximized is concave and the constraint set is convex. Therefore, the solution of (3.24) is given by Lemma 3.5 (suppose the dual function $g(\cdot)$ is differentiable).

Let $\boldsymbol{\lambda}^*$ denote a solution (a global maximizer) of the dual problem (3.24). For clarity, we refer to $\boldsymbol{\lambda}^*$ as *dual optimal* or *optimal Lagrange multipliers*, and $\boldsymbol{x}^*$, a solution of the primal problem (3.17), as *primal optimal*. The optimal value $g(\boldsymbol{\lambda}^*)$ of the dual problem (3.24) is, by definition, is the best lower bound on $f(\boldsymbol{x}^*)$ that can be obtained from the dual function. In particular, we have

$$g(\boldsymbol{\lambda}^*) \leq f(\boldsymbol{x}^*). \tag{3.25}$$

This property is called the *weak duality*. The difference $f(\boldsymbol{x}^*) - g(\boldsymbol{\lambda}^*)$ is called the *optimal duality gap* between the primal problem and the dual problem, which is always nonnegative.

If the optimal duality gap attains zero, that is,

$$g(\boldsymbol{\lambda}^*) = f(\boldsymbol{x}^*), \tag{3.26}$$

then we say that *strong duality* holds.

However, strong duality does not always hold, even when the primal problem is convex. There are a lot of results that establish conditions on the problem under which strong duality holds. These conditions are called *constraint qualifications* (16).

KKT Optimality Conditions

Now suppose strong duality holds. For any feasible point $\boldsymbol{x}$ of the primal problem (3.17) and $\boldsymbol{\lambda}$ of the dual problem (3.24), we have

$$f(\boldsymbol{x}) - f(\boldsymbol{x}^*) \leq f(\boldsymbol{x}) - g(\boldsymbol{\lambda}), \tag{3.27}$$

since $g(\boldsymbol{\lambda}) \leq f(\boldsymbol{x}^*)$. This means that $\boldsymbol{x}$ is an ϵ-suboptimal with $\epsilon = f(\boldsymbol{x}) - g(\boldsymbol{\lambda})$. That is, dual feasible points allow us to bound how suboptimal a given feasible point is, without knowing the exact value of $f(\boldsymbol{x}^*)$.

We refer to the gap between primal and dual objectives, i.e., $f(\boldsymbol{x}) - g(\boldsymbol{\lambda})$, as the *duality gap* associated with the primal feasible point $\boldsymbol{x}$ and dual feasible point $\boldsymbol{\lambda}$. Any primal dual feasible pair $\{\boldsymbol{x}, \boldsymbol{\lambda}\}$ localizes the optimal value of the primal and dual problems to an interval $[g(\boldsymbol{\lambda}), f(\boldsymbol{x})]$, that is,

$$g(\boldsymbol{\lambda}) \leq g(\boldsymbol{\lambda}^*) \leq f(\boldsymbol{x}^*) \leq f(\boldsymbol{x}). \tag{3.28}$$

Obviously, if the duality gap of a primal dual feasible pair $\{\boldsymbol{x}, \boldsymbol{\lambda}\}$ is zero, i.e., $g(\boldsymbol{\lambda}) = f(\boldsymbol{x})$, then $\boldsymbol{x}$ is the primal optimal, $\boldsymbol{\lambda}$ is the dual optimal and strong duality holds.

Let $\boldsymbol{x}^*$ be a primal optimal and $\boldsymbol{\lambda}^*$ be a dual optimal. By strong duality, we have

$$f(\boldsymbol{x}^*) = g(\boldsymbol{\lambda}^*) = \inf_{\boldsymbol{x} \in \mathcal{D}} L(\boldsymbol{x}, \boldsymbol{\lambda}^*) \leq L(\boldsymbol{x}^*, \boldsymbol{\lambda}^*) \leq f(\boldsymbol{x}^*). \tag{3.29}$$

The first equality states that the optimal duality gap is zero, the second equality follows the definition of the dual function, the third inequality follows because the primal optimal $\boldsymbol{x}^* \in \mathcal{C} \subseteq \mathcal{D}$, and the last inequality follows because $\boldsymbol{\lambda} \succeq \mathbf{0}$ and $f_i(\boldsymbol{x}) \leq 0$, $i = 1, ..., m$.

We can draw several interesting conclusions from (3.29). Firstly, the last inequality is indeed an equality, which implies that $\sum_{i=1}^{m} \lambda_i^* f_i(\boldsymbol{x}^*) = 0$; since each term in the sum is non-positive, we further have $\lambda_i^* f_i(\boldsymbol{x}^*) = 0$, $i = 1, ..., m$. This condition is referred to as the *complementary slackness*, which holds for any primal optimal $\boldsymbol{x}^*$ and any dual optimal $\boldsymbol{\lambda}^*$ (when strong duality holds). The complementary slackness can also be expressed as

$$\lambda_i^* > 0 \Rightarrow f_i(\boldsymbol{x}^*) = 0,$$

or, equivalently,

$$f_i(\boldsymbol{x}^*) > 0 \Rightarrow \lambda_i^* = 0.$$

Roughly speaking, this means the i-th optimal Lagrange multiplier is zero unless the i-th constraint is active at the optimum.

Secondly, the third inequality is also an equality, i.e., $\inf_{\boldsymbol{x} \in \mathcal{D}} L(\boldsymbol{x}, \boldsymbol{\lambda}^*) = L(\boldsymbol{x}^*, \boldsymbol{\lambda}^*)$, which implies that $\boldsymbol{x}^*$ minimizes the Lagrangian $L(\boldsymbol{x}, \boldsymbol{\lambda}^*)$. This means that

$$\frac{\partial L(\boldsymbol{x}^*, \boldsymbol{\lambda}^*)}{\partial \boldsymbol{x}} = \nabla f(\boldsymbol{x}^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(\boldsymbol{x}^*) = \mathbf{0}. \tag{3.30}$$

Based on above, we can obtain the necessary and sufficient conditions for a primal dual feasible pair $\{\boldsymbol{x}^*, \boldsymbol{\lambda}^*\}$ to be optimal (for the primal problem and dual problem, respectively). We refer to these conditions as the *Karush-Kuhn-Tucker* (*KKT*) conditions (17).

Lemma 3.8 Karush-Kuhn-Tucker (KKT) Conditions *A primal dual feasible pair* $\{\boldsymbol{x}^*, \boldsymbol{\lambda}^*\}$ *is optimal for the primal problem and dual problem, respectively, if and only if*

$$\begin{cases} f_i(\boldsymbol{x}^*) \leq 0, \ \lambda_i^* \geq 0, \ \lambda_i^* \cdot f_i(\boldsymbol{x}^*) = 0, \quad i = 1, ..., m \\ \nabla f(\boldsymbol{x}^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(\boldsymbol{x}^*) = \mathbf{0}. \end{cases} \tag{3.31}$$

According to Lemma 3.8, solving the primal problem (3.17) is the same as finding the primal dual feasible pairs $\{\boldsymbol{x}, \boldsymbol{\lambda}\}$ that satisfy the KKT conditions in (3.31). In a few special cases it is possible to solve the KKT conditions (and therefore, the optimization problem) analytically. More generally, the KKT conditions must be solved by an iterative algorithm.

Shadow Price

There are some interesting interpretations for the Lagrange multipliers $\lambda_i, \ i = 1, ..., m$. Now we can give a simple geometric interpretation of the Lagrange multipliers in terms of economics, where they are often be interpreted as *prices*.

To show this, we first introduce the perturbed version of the original problem (3.17)

$$\begin{array}{ll} \text{minimize} & f(\boldsymbol{x}) \\ \text{subject to} & f_i(\boldsymbol{x}) \leq u_i, \ i = 1, ..., m, \end{array} \tag{3.32}$$

where u_i is the perturbing parameter for the i-th inequality constraint. That is, When u_i is positive, it means that we have relaxed the i-th constraint; when u_i is negative, it means that we have tightened the constraint. This perturbed problem coincides with the original problem (3.17) when $\boldsymbol{u} \triangleq (u_1, ..., u_m) = \mathbf{0}$.

The optimal value of the perturbed problem (3.32) is given by

$$p^*(\boldsymbol{u}) \triangleq \inf_{\boldsymbol{x} \in \mathcal{C}(\boldsymbol{u})} f(\boldsymbol{x}),$$

where $\mathcal{C}(\boldsymbol{u}) \triangleq \{\boldsymbol{x} \mid f_i(\boldsymbol{x}) \leq u_i, i = 1, ..., m\}$ is the constraint set of the perturbed problem (3.32). Note that both the constraint set $\mathcal{C}(\boldsymbol{u})$ and the optimal value $p^*(\boldsymbol{u})$ of the perturbed problem (3.32) depend on the perturbing parameters u_i, $i = 1, ..., m$. When $\boldsymbol{u} = \mathbf{0}$, we have $\mathcal{C}(\boldsymbol{u}) = \mathcal{C}$ where $\mathcal{C}$ is the constraint set of the original problem (3.17), and $p^*(\mathbf{0}) = f(\boldsymbol{x}^*)$ where $f(\boldsymbol{x}^*) \triangleq \inf_{\boldsymbol{x} \in \mathcal{C}} f(\boldsymbol{x})$ is the optimal value of the original problem (3.17).

Suppose $\boldsymbol{x} \in \mathcal{C}(\boldsymbol{u})$ is any feasible point for the perturbed problem (3.32), i.e., $f_i(\boldsymbol{x}) \leq u_i$, $i = 1, ..., m$. For any perturbing parameters $\boldsymbol{u}$ and feasible point $\boldsymbol{x} \in \mathcal{C}(\boldsymbol{u})$, we have

$$p^*(\mathbf{0}) = f(\boldsymbol{x}^*) = g(\boldsymbol{\lambda}^*) \leq f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i^* f_i(\boldsymbol{x}) \leq f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i^* u_i.$$

The second equality follows from the strong duality, the third inequality follows from the definition of $g(\boldsymbol{\lambda}^*)$, and the last inequality follows because $f_i(\boldsymbol{x}) \leq u_i$ and $\lambda_i^* \geq 0$, $i = 1, ..., m$. Since the above formula holds for any feasible point $\boldsymbol{x} \in \mathcal{C}(\boldsymbol{u})$, we have

$$p^*(\boldsymbol{u}) \triangleq \inf_{\boldsymbol{x} \in \mathcal{C}(\boldsymbol{u})} f(\boldsymbol{x}) \geq p^*(\mathbf{0}) - \sum_{i=1}^{m} \lambda_i^* u_i.$$

Suppose now that $p^*(\boldsymbol{u})$ is differentiable at $\boldsymbol{u} = \mathbf{0}$. Then we have

$$\frac{\partial p^*(\mathbf{0})}{\partial u_i} = -\lambda_i^*.$$

Now we give a simple interpretation of the above result in terms of economics. As we view the variable $\boldsymbol{x}$ as a firm's investments on n different resources, the objective $f(\cdot)$ as the firm's cost, or $-f(\cdot)$ as the firm's profit, and each constraint $f_i(\boldsymbol{x}) \leq 0$ as a limit on some resource investments. The (negative) perturbed optimal cost function $-p^*(\boldsymbol{u})$ tells us how much more or less profit could be made if more, or less, of each resource were made available to the firm. In other words, when $\boldsymbol{u}$ is closed to $\mathbf{0}$, the Lagrange multiplier λ_i^* tells us approximately how much more profit the firm could make, for a small increase in the availability of resource i. Thus, λ_i^* can be viewed as the natural or equilibrium *price* for resource i. For this reason a dual optimal $\boldsymbol{\lambda}^*$ is sometimes called *shadow prices*.

Solving Dual Problem Using the Subgradient Method

We now consider the methods for solving dual problems, also called dual methods. There are several incentives for solving the dual problem in place of the primal: (a) The dual problem is a convex optimization, while the primal may not be convex; (b) The dual problem may have smaller dimension and/or simpler constraints than the primal; (c) If strong duality holds, the dual optimal value is exactly the primal optimal value; and (d) Even if strong duality does not hold, the dual optimal value provides a lower bound to the primal optimal value, which may be useful in designing iterative algorithms.

Of course, we should also consider some of the difficulties in solving the dual problem. The most critical ones are the following: (a) The evaluation of the dual function $g(\boldsymbol{\lambda})$ at any dual variable $\boldsymbol{\lambda}$ requires minimization of the Lagrangian $L(\boldsymbol{x}, \boldsymbol{\lambda})$ over all $\boldsymbol{x} \in \mathcal{D}$; (b) The dual function $g(\boldsymbol{\lambda})$ may not be differentiable in many types of problems; and (c) If strong duality does not hold, there is certain duality gap between the dual optimal and primal optimal.

We will discuss an important type of dual methods, namely the *subgradient method*, which is particularly suitable for solving a dual problem with nondifferentiable objective function. The basic idea of subgradient methods is to generate a minimizing sequence of dual feasible $\boldsymbol{\lambda}^{(k)}$ using subgradients rather than gradients as search direction.

Given a convex function $f : \mathbb{R}^n \to \mathbb{R}$, we say that a vector $\boldsymbol{d} \in \mathbb{R}^n$ is a *subgradient* of $f(\cdot)$ at a feasible point $\boldsymbol{x} \in \mathcal{D}(f)$ if

$$f(\boldsymbol{z}) \geq f(\boldsymbol{x}) + \boldsymbol{d}^T(\boldsymbol{z} - \boldsymbol{x}), \quad \forall \boldsymbol{z} \in \mathcal{D}(f). \tag{3.33}$$

If instead $f(\cdot)$ is a concave function, we say that $\boldsymbol{d}$ is a subgradient of $f(\cdot)$ at point $\boldsymbol{x}$ if and only if $-\boldsymbol{d}$ is a subgradient of $-f(\cdot)$ at $\boldsymbol{x}$. This means that a subgradient $\boldsymbol{d}$ of the dual function $g(\boldsymbol{\lambda})$ at a dual feasible point $\boldsymbol{\lambda} \in \mathcal{D}(g)$ satisfies:

$$g(\boldsymbol{\mu}) \leq g(\boldsymbol{\lambda}) + \boldsymbol{d}^T(\boldsymbol{\mu} - \boldsymbol{\lambda}), \quad \forall \boldsymbol{\mu} \in \mathcal{D}(g). \tag{3.34}$$

Thus, the subgradient method generates a minimizing sequence of dual feasible $\boldsymbol{\lambda}^{(k)}$ according to the following iteration

$$\boldsymbol{\lambda}^{(k+1)} = \left[\boldsymbol{\lambda}^{(k)} + \gamma^{(k)}\boldsymbol{d}^{(k)}\right]^+, \tag{3.35}$$

where $\gamma^{(k)}$ is the step-size, $\boldsymbol{d}^{(k)}$ is the subgradient of $g(\boldsymbol{\lambda})$ at point $\boldsymbol{\lambda}^{(k)}$, and $[\boldsymbol{\lambda}]^+$ denotes the projection of $\boldsymbol{\lambda}$ on the constraint set of the dual problem (3.24).

The fundamental difference between the gradient-based method and the subgradient method is that with the subgradient method, the new iterative may not improve the dual objective for all values of the step-size $\gamma^{(k)}$. That is, for some large $\gamma^{(k)}$ we may have $g(\boldsymbol{\lambda}^{(k+1)}) < g(\boldsymbol{\lambda}^{(k)})$, whereas for sufficiently small step-size $\gamma^{(k)}$ we have $g(\boldsymbol{\lambda}^{(k+1)}) \geq g(\boldsymbol{\lambda}^{(k)})$. This is shown in the following Lemma, which also provides an estimate for the range of appropriate step-sizes.

Lemma 3.9 *For every dual optimal solution $\boldsymbol{\lambda}^*$, we have $||\boldsymbol{\lambda}^{(k+1)} - \boldsymbol{\lambda}^*|| < ||\boldsymbol{\lambda}^{(k)} - \boldsymbol{\lambda}^*||$ for all step-sizes $\gamma^{(k)}$ such that*

$$0 < \gamma^{(k)} < 2 \cdot \frac{g(\boldsymbol{\lambda}^*) - g(\boldsymbol{\lambda}^{(k)})}{||\boldsymbol{d}^{(k)}||^2}. \tag{3.36}$$

Unfortunately, the above range for $\gamma^{(k)}$ requires the dual optimal value $g(\boldsymbol{\lambda}^*)$, which is usually unknown. In practice, we can use the following approximate step-size formula

$$\gamma^{(k)} = \alpha^{(k)} \cdot \frac{g^{(k)} - g(\boldsymbol{\lambda}^{(k)})}{||\boldsymbol{d}^{(k)}||^2}, \tag{3.37}$$

where $g^{(k)}$ is an approximation to the optimal dual value and $0 < \alpha^{(k)} < 2$.

There are many variations of subgradient methods that aim to accelerate the convergence of the basic method. For more details, please refer to (11; 12).

3.2 APPLICATION I: RESOURCE ALLOCATION FOR WIRELESS VIDEO STREAMING

3.2.1 BACKGROUND

With the advances of mobile computing technologies and deployments of new cellular infrastructure, video communications are becoming more important in many new business applications. However, there are still many open problems in terms of how to efficiently provision complicated video QoS requirements for mobile users. One particular challenging problem is multi-user video streaming over wireless channels, where the demand for better video quality and small transmission delays needs to be reconciled with the limited and often time-varying communication resources. There are several main technical difficulties listed below.

First, the video sources for most streaming applications are typically pre-coded stored video sequences with relative high bit rates. However, the currently deployed wireless cellular systems (e.g., (18; 19)) are designed to only support voice and lower bit rate data. In order to support video streaming over such networks, the high rate video sources need to be adapted through a variety of schemes, such as scalable video stream extraction (e.g., (20)), transcoding (e.g., (21; 22)), and summarization (e.g., (23)), before they can be accommodated by the wireless channel.

Second, different video content segments have different rate-distortion characteristics, e.g., some segment may be part of an action movie and requires a lot of bits to encode, while others maybe news anchors talking that require relatively less bits to encode. In a wireless multi-access channel, the type of multi-user content diversity in content rate-distortion characteristics should be taken into consideration while optimizing the network resource.

Third, the resource consumptions of video users are typically discrete, i.e., measured in frames instead of in bits. As a result, their utility functions (QoS as functions of allocated resources) are discrete as well, and typically do not have close form representations. Therefore most of previous work on resource allocation for elastic data traffic does not directly apply here, and a new optimization framework is needed.

Last but not least, the streaming applications have stringent delay requirements, which can be only satisfied under a carefully designed scheduling policy. This is a challenging task in a wireless network, since the transmissions of multiple users are typically tightly coupled either due to limited network resource (e.g., transmission power or bandwidth in downlink transmissions) or mutual interferences (e.g., in uplink transmissions).

Traditionally the mechanisms of content generation and the engineering of network resources are designed separately, and most of the above challenges are ignored in the network design. It is useful to consider a new coherent framework for resource allocation, source adaptation, and deadline oriented scheduling. During the resource allocation phase, the network resources are allocated to different video users by temporally treating them as "elastic data users", i.e., without considering the discrete nature of the video traffic.

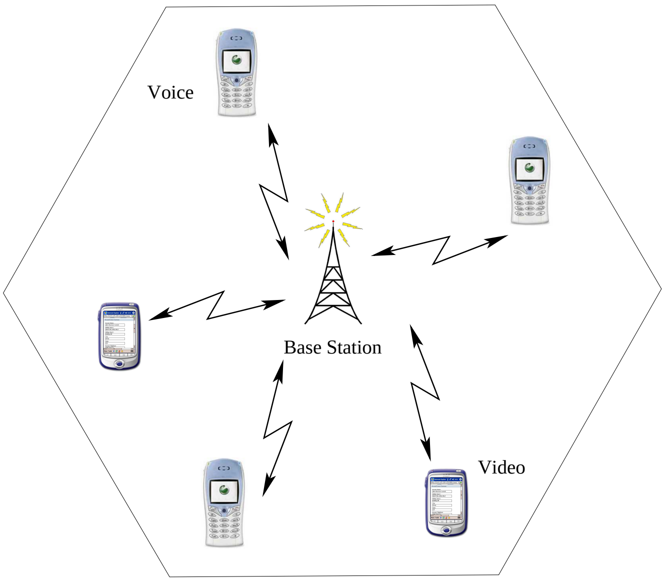

Figure 3.7: A single cell network with mixed voice and video users.

An optimal average resource allocation is achieved in a distributed fashion by exploiting the content diversity among users. Based on the average resource allocation, users perform source adaptations in a distributed fashion to select a set of video frames to be transmitted in order to match the allocated resource. And then the base station perform greedy deadline oriented scheduling by taking advantage of the variable bit rate (VBR) nature of users' traffic.

The average resource allocation can be done through a price-based approach to maximize the total network performance. We will first introduce the general solution framework, in particular, the key idea behind the average resource allocation. Then we will discuss in details the uplink and downlink streaming in details.

3.2.2 NETWORK OPTIMIZATION FRAMEWORK

We consider a single cell based on Code Division Multiple Access (CDMA). A *fixed* user population with both voice and video applications are considered, as shown in Fig. 3.7. All users communicate with the base station through one-hop transmission. A voice transmission is successful if a target Signal-to-Interference-plus-Noise Ratio (SINR) is reached at the receiver. A video users is more flexible and can adapt to the network environment in terms of the achieved SINR and the transmission rate. However, once the video frames are transmitted, stringent delay deadlines need to be satisfied in order to guarantee the normal operation of the streaming application.

The network objective is to maximize the overall quality of the video users, subject to the normal operations of voice users. We will achieve this by allocating various network resource (i.e., transmission power & transmission time), video source signal processing (i.e., adaptation by summarization) and scheduling (both "soft scheduling" in terms of deadline aware power allocation, and "hard scheduling" in a time-division-multiplexing fashion).

We will consider both uplink and downlink video streaming. In the uplink case, video users need to limit aggregate interference that they generate and affect the voice users. In the downlink case, the base station needs to limit the amount of transmission power allocated to the video users. In both cases, the optimal video streaming problem can be modeled in the framework of nonlinear constrained optimization. Two key questions that need to be answered are: 1) how to allocate resources among video users in an efficient manner (i.e., maximizing total user' quality or minimizing total users' distortion), and, 2) how to make sure that the stringent delivery deadline requirements are met for every video frame that is chosen to for transmission.

The general solution framework that answers the above two questions. This framework involves three phases:

1. *Average resource allocation.* This is achieved by solving a network utility maximization (NUM) problem. The multiuser content diversity will be fully exploited to make efficient use of the network resources. A distributed pricing-based algorithm is proposed to achieve the resultant solution.

2. *Video source adaptations.* Based on the average resource allocation results in phase 1), each video user adapts the video source by solving a localized optimization problem with video summarization.

3. *Multiuser deadline oriented scheduling.* The network decides a transmission schedule based on video users' source decisions in phase 2), in order to meet the stringent deadline constraints of the streaming applications.

In some cases we may not be able to find a feasible schedule in phase 3). This implies that although the system resource is enough in an average sense (guaranteed by phase 1)), the deadline requirements might be too stringent to satisfy. In that case, we will go back to phase 1) and re-optimize the average resource allocation, but with more stringent resource constraints (e.g., less total power in downlink transmission). This will force the users to be more conservative when doing the source adaptations in phase 2) (i.e., each user will transmit fewer frames), thus make it easier to achieve a feasible schedule in step (3).

3.2.3 AVERAGE RESOURCE ALLOCATION

A key question of resource allocation for multimedia communication is how to deal with the VBR nature of the source. We take a decoupling approach here, by first considering the resource allocation in the *average* sense without worrying about the time dependency. The time dependency will be brought back into the picture later in the source adaptation and multiuser scheduling phases.

Assume there are N video users in the cell. We characterize the QoS of a video user n by a utility function $U_n(x_n)$, which is an increasing and strictly concave function of the communication resource allocated to user n, x_n. This models various commonly used video quality measures such as the rate-PSNR function and rate-summarization distortion functions. It is well known from information theory (24) that the rate-distortion functions for a variety of sources are convex, and in practice, the operational rate distortion functions are usually convex as well. Thus the utility functions (defined as negative distortion) are concave. For the average resource allocation phase, we assume that $U_n\left(x_n\right)$ is continuous in x_n.

The average resource allocation is achieved by solving the following NUM problem, where $X_{\max}$ denotes the total limited resource available to the video users (i.e., total transmission power in the downlink case and total transmission time in the uplink case),

$$\max_{\{x_n \geq 0, 1 \leq n \leq N\}} \sum_n U_n\left(x_n\right), \text{ s.t. } \sum_n x_n \leq X_{\max}. \tag{3.38}$$

Solving Problem (3.38) directly requires a centralized computation due to the coupling resource constraint. However, a distributed solution is often more desirable, since the base station typically does not know the utility functions of individual video users. Here we use the dual decomposition technique as introduced in Section 3.1.3, where the base station sets a price on the resource, and each mobile user determines its average resource request depending on the announced price and its own source utility characteristic.

For our problem, the dual-based decomposition works as follows. First, we relax the constraint in (3.38) with a dual variable λ and obtain the following Lagrangian

$$L\left(\boldsymbol{x}, \lambda\right) \triangleq \sum_n U_n\left(x_n\right) - \lambda\left(\sum_n x_n - X_{\max}\right), \tag{3.39}$$

where $\boldsymbol{x} = (x_{\,n}, 1 \leq n \leq N)$. Then Problem (3.38) can be solved at two levels. At the lower level, each video user solves the following problem,

$$\max_{x_n \geq 0} \left\{U_n\left(x_n\right) - \lambda x_n\right\}, \tag{3.40}$$

which corresponds to maximizing the surplus (i.e., utility minus payment) based on price λ. Denote the optimal solution of (3.40) as $x_n\left(\lambda\right)$, which is unique since the utility function is continuous, increasing, and strictly concave. The video users then feedback the values of $x_n\left(\lambda\right)$ to the base station. At the higher level, the base station adjusts λ to solve the following problem

$$\min_{\lambda \geq 0} g\left(\lambda\right) \triangleq \sum_n g_n\left(\lambda\right) + \lambda X_{\max}, \tag{3.41}$$

where $g_n(\lambda)$ is the maximum value of (3.40) for a given value of λ. The dual function $g(\lambda)$ is non-differentiable in general, and (3.41) can be solved using the sub-gradient searching method,

$$\lambda^{l+1} = \max\left\{0, \lambda^l + \alpha^l \left(\sum_n x_n\left(\lambda^l\right) - X_{\max}\right)\right\}, \tag{3.42}$$

where l is the search iteration index and α^l is a small step size at iteration l.

The two level optimizations together solve the *dual* problem of the original NUM problem (3.38) (which we call the *primal* problem). This enables us to obtain a distributed solution. Base station controls the resource price according to (3.42), and each video user n chooses the average resource request x_n to maximize its surplus according to (3.40) in a distributed fashion. This avoids centralized computation and makes the solution scalable in a large network.

Given the assumption on the utility functions, we have the property of strong duality which implies zero duality gap. In other words, given the optimal dual solution λ^*, the corresponding $x_n(\lambda^*)$ for all n are the optimal solution of the primal problem (3.38).

The complete distributed algorithm is given in Algorithm 1.

Algorithm 1 Dual-based Optimization Algorithm to solve Problem (3.38)

```
Initialization: set iteration index $l = 0$, and choose $0 < \epsilon \ll 1$ as the stopping criterion.
Base station announces an arbitrary initial price $\lambda^0 > 0$.
repeat
    for all video user $n$ do
        Locally determine the resource consumption $x_n\left(\lambda^l\right) = \arg\max_{x_n}\{U_n(x_n) - \lambda^l x_n\}$.
        Send the value of $x_n\left(\lambda^l\right)$ to the base station.
    end for
    Base station announces a new price $\lambda^{l+1} = \max\left\{0, \lambda^l + \alpha^l\left(\sum_n x_n\left(\lambda^l\right) - X_{\max}\right)\right\}$.
    $l = l + 1$.
until $|\lambda^l - \lambda^{l-1}| < \epsilon$.
```

Algorithm 1 converges under properly chosen step sizes, as stated in the following proposition (for proof, see (25)).

Proposition 3.10 *If the step-sizes in (3.42) satisfy* $\lim_{l\to\infty} \alpha^l = 0$ *and* $\sum_l \alpha^l \to \infty$ *(e.g.,* $\alpha^l = 1/l$*), then Algorithm 1 converges to the optimal solution of Problem (3.38).*

So far we have not specified how Problem (3.40) is solved in Algorithm 1. Since the utility functions in video communications typically do not have closed form representations, Problem (3.40) needs to be solved by using various source adaptation techniques. This is different from, for example, congestion control in the Internet, where each source determines the transmission rate as a closed form function of the network congestion price. To solve Problem (3.40), we need to design intelligent source adaptation and opportunistic deadline oriented scheduling algorithms, with details shown in (26).

Next we give two concrete examples of Problem (3.40) for wireless uplink and downlink streaming.

3.2.4 WIRELESS UPLINK STREAMING

In a wireless CDMA network, different users transmit using different spreading codes. These codes are mathematically orthogonal under synchronous reception. However, the orthogonality is partially destroyed when the transmissions are asynchronous, such as in the uplink transmissions. The received SINR in that case is determined by the users' transmission power, the spreading factors (defined as the ratio of the bandwidth and the achieved rate), the modulation scheme used, and the background noise. The maximum constrained resource of the video users can be expressed as the maximum received power at the base station, derived based on a physical layer model similar as the one used in (27).

We consider the uplink transmission in a single CDMA cell with M voice users and N video streaming users. The total bandwidth W is fixed and shared by all users. Each voice user has a QoS requirement represented in bit error rates (BER) (or frame error rates (FER)), which can be translated into a target SINR at the base station, γ_{voice}. Each voice user also has a target rate constraint R_{voice}. Assuming perfect power control, each voice user achieves the same received power at the base station, P^r_{voice}. The total received power at the base station from all video users is denoted as $P^{r,all}_{video}$. The background noise n_0 is fixed and includes both thermal noise and inter-cell interferences.

In order to support the successful transmission of all voice users, we need to satisfy

$$\frac{W}{R_{voice}} \frac{G_{voice} P^r_{voice}}{n_0 W + (M-1) P^r_{voice} + P^{r,all}_{video}} \geq \gamma_{voice}. \tag{3.43}$$

Here W/R_{voice} is the spreading factor, and coefficient G_{voice} reflects the fixed modulation and coding schemes used by all voice users (e.g., $G_{voice} = 1$ for BPSK and $G_{voice} = 2$ for QPSK). For each voice user, the received interference comes from the other $M-1$ voice users and all video users. From (3.43), we can solve for the maximum allowed value of $P^{r,all}_{video}$, denoted as $P^{r,\max}_{video}$

$$P^{r,\max}_{video} = \left(\frac{W G_{voice}}{R_{voice} \gamma_{voice}} - (M-1) \right) P^r_{voice} - n_0 W, \tag{3.44}$$

which is assumed to be fixed given fixed number of voice users M.

The network objective is to choose the transmission power of each video user during a time segment $[0, T]$, such that the total video's utility is maximized, i.e.,

$$\begin{aligned} \max_{\{P_n(t), 1 \leq n \leq N\}} & \sum_{n=1}^{N} U_n \left(\int_0^T R_n \left(\boldsymbol{P}\left(t\right) \right) \right) dt \\ \text{s.t.} & \sum_{n=1}^{N} h_n P_n\left(t\right) \leq P_{video}^{r,\max}, \forall t \in [0, T] \\ & 0 \leq P_n\left(t\right) \leq P_n^{\max}, 1 \leq n \leq N, \end{aligned} \tag{3.45}$$

where $P_n(t)$ is the transmission power of video user n at time t, $\boldsymbol{P}(t)$ is the vector of all video users' transmission power at time t, $P_n^{\max}$ is the maximum peak transmission power of user n, and h_n is the fixed channel gain from the transmitter of user n to the base station. $R_n\left(t\right)$ is the rate achieved by user n at time t, and depends on all video users' transmission power, the channel gains, the background noise, and interference from voice users. A user n's utility function is defined on the video summarization quality of its transmitted sequence during $[0, T]$.

Problem (3.45) is not a special case of Problem (3.38), since 1) Problem (3.45) optimizes over N functions $\left(P_n\left(t\right), 1 \leq n \leq N\right)$, whereas Problem (3.38) optimizes over N variables $\left(x_n, 1 \leq n \leq N\right)$, and 2) the objective function in Problem (3.45) is coupled across users, whereas the objective in Problem (3.38) is fully decoupled. This makes (3.45) difficult to solve in a distributed fashion.

In order to solve Problem (3.45), we will resort to the framework described in Section 3.2.2, where we will perform average resource allocation (in terms of average transmission power), source adaptation (to match the average resource allocation), and the deadline scheduling (to determine the exact power allocation functions by deadline aware water-filling).

To simplify the problem and make the solution tractable, we consider the case where video users transmit in a TDM fashion. This is motivated by (28), where the authors showed that in order to achieve maximum total rate in a CDMA uplink, it is better to transmit weak power users in groups and strong power users one by one. Since video users typically need to achieve much higher rate than voice users (thus transmit at much higher power), it is reasonable to avoid simultaneous transmissions among video users, and thus avoid large mutual interference. A more important motivation for TDM transmission here is to exploit the temporal variation of the video contents, i.e., content diversity. Under such a TDM transmission scheme, the constrained resource to be allocated to the video users becomes the total transmission time of length T. The total number of bits that can be transmitted by user n is determined by the transmission time allocated to it, $t_n \in [0, T]$, and the maximum rate it can achieve while it is allowed to transmit. Let us denote this rate as R_n^{TDM}, and it can be calculated by,

$$R_n^{TDM} = W \log_2 \left(1 + \frac{\min \{ h_n P_n^{\max}, P_{video}^{r,\max} \}}{n_0 W + M P_{voice}^r} \right). \tag{3.46}$$

Under the assumption of TDM transmission, Problem (3.45) can be written as follows

$$\max_{\{t_n \geq 0, 1 \leq n \leq N\}} \sum_{n=1}^{N} \tilde{U}_n (t_n), \text{ s.t. } \sum_{n=1}^{N} t_n \leq T, \tag{3.47}$$

where the new utility function $\tilde{U}_n$ is defined as

$$\tilde{U}_n (t_n) = U_n \left(R_n^{TDM} t_n \right), \tag{3.48}$$

i.e., a user n's total transmitted data during time $[0, T]$ is determined by the product of R_n^{TDM} and the active transmission time t_n. Now Problem (3.47) is a special case of Problem (3.38), where we replace U_n by $\tilde{U}_n$, x_n by t_n and $X_{\max}$ by T. As a result, the optimal transmission time allocation per user can be found based on the discussions in Section using Algorithm 1.

Once the transmission time allocations are determined, each user locally adapts its source using summarization, which leads to the best sequence of video frames that fit into the transmission time allocation t_n. The transmission of each frame needs to meet a certain delivery deadline, after which the frame becomes useless. This requires the base station to determine a transmission schedule for all users, and the details can be found in Section III of (26).

3.2.5 WIRELESS DOWNLINK STREAMING

Different from the uplink case, transmissions in the downlink are orthogonal to each other, thus it is desirable to allow simultaneous transmissions of multiple video users. The resource constraint in the downlink case is the maximum peak transmission power at the base station. The objective here is to determine the transmission power functions, $P_n(t)$, of each user n during time $t \in [0, T]$, such that the total user utility (measured in video quality) is maximized.

Following the framework described in Section 3.2.2, the first step is to perform average resource allocation. For the downlink case, we will allocate the transmission power to each user, subject to the total transmission power constraint (for video users) at the base station, $P_{\max}^{base}$. Since there is no mutual interference, the transmissions of the voice users need not be taken into consideration when determining the achievable rates of the video users.

At this stage, we will temporality assume that each user n will transmit at a fixed power level P_n throughout the time segment $[0, T]$. The problem we want to solve is:

$$\max_{\{P_n \geq 0, 1 \leq n \leq N\}} \sum_{n=1}^{N} U_n (P_n), \text{ s.t. } \sum_{n=1}^{N} P_n \leq P_{\max}^{base}. \tag{3.49}$$

Problem (3.49) is a special case of Problem (3.38), and can be solved using Algorithm 1. Assuming that user n is allocated a constant transmission power P_n^*, its total throughput within $[0, T]$ is given by

$$TW \log_2 \left(1 + \frac{h_n P_n^*}{n_0 W}\right), \tag{3.50}$$

where h_n is the channel gain from base station to the mobile receiver, and n_0 is the background noise density at the receiver end. The user can determine its best video summary sequence based on this achieved throughput.

Due to the difference in frame sizes and locations, transmitting at constant power levels is not optimal in terms of meeting the frame delivery deadlines. We can further perform an energy-efficient water-filling power allocation to improve the performance of Problem (3.49). For details, see Section IV of (26).

3.3 APPLICATION II: WIRELESS SERVICE PROVIDER COMPETITION

3.3.1 BACKGROUND

Due to the deregulation of the telecommunication industry, future wireless users are likely to freely choose a provider (or providers) offering the best tradeoff of parameters. This is already happening with some public Wi-Fi connections, where users can connect to wireless access points of their choice, with usage-based payments and no contracts. Despite the common presence of a free public Wi-Fi network, some users may still choose more expensive providers who offer better quality of service.

Here we consider a situation where wireless service providers compete to sell limited wireless resources (e.g., frequency bands, time slots, transmission power) to users who are free to choose provider(s). We investigate how providers set prices for the resource, and how users choose the amount of resource they purchase and from which providers. The focus of our study is to characterize the outcome of this interaction. We consider the general case where different users have different utility functions and experience different channel conditions to different service providers.

A proper model for this system is a multi-leader-follower game. The providers announce the wireless resource prices in the first stage, and the users announce their demand for the resource in the second stage. A user's choice is based on providers' prices and its channel conditions. The providers select their prices to maximize their revenues, keeping in mind the impact of their prices on the demand of the users. As in (29; 30), we assume that users pay for the allocated resources instead of the received services. This turns out to be crucial in achieving the globally optimal resource allocation. However, in this chapter we will first look at the corresponding social welfare optimization problem, as well as a distributed primal-dual algorithm that can achieve the optimal solution of the problem. In

Section 5.3, we will revisit this problem and see how to analyze the game theoretical interactions between the competitive providers. A surprising result there is that the equilibrium of the game is actually the same as the optimal solution of the social welfare optimization problem studied here.

3.3.2 SYSTEM MODEL AND ASSUMPTIONS

We consider a set $\mathcal{J} = \{1, \ldots, J\}$ of service providers and a set $\mathcal{I} = \{1, \ldots, I\}$ of users. Provider $j \in \mathcal{J}$ has a total of Q_j resource. A user $i \in \mathcal{I}$ can obtain resource from one or more providers, with a demand vector $\boldsymbol{q_i} = [q_{i1} \ \cdots \ q_{iJ}]$ and q_{ij} represents the demand from user i to provider j. We use $\boldsymbol{q} = [\boldsymbol{q}_1 \cdots \boldsymbol{q}_I]$ to denote the demand vector of all users.

User i's utility function would be $u_i\left(\sum_{j=1}^{J} q_{ij} c_{ij}\right)$, where c_{ij} is the *channel quality offset* for the channel between user i and the base station of provider j (see Example 3.11 and Assumption 2), and u_i is an increasing and concave utility function. The communication can be both downlink or uplink, as long as users do not interfere with each other by using orthogonal resources.

Under this model, a user is allowed to purchase from several providers at the same time. For this to be feasible, a user's device might need to have several wireless interfaces. Mathematically, the solution of this model gives an upper bound on best performance of any situation where users are constrained to purchase from one provider alone.

Next we give a concrete example of how our model is mapped into a physical wireless system.

Example 3.11 Consider wireless providers operating on orthogonal frequency bands W_j, $j \in \mathcal{J}$. Let q_{ij} be be the fraction of time that user i is allowed to transmit exclusively on the frequency band of provider j, with the constraint $\sum_{i \in \mathcal{I}_j} q_{ij} = 1$, $j \in \mathcal{J}$. Furthermore, assume that each user has a peak power constraint P_i. We can then define $c_{ij} = W_j \log(1 + \frac{P_i |h_{ij}|^2}{\sigma_{ij}^2 W_j})$, where h_{ij} is the channel gain and σ_{ij}^2 is the Gaussian noise variance for the channel between user i and network j. In this case, a user's payoff is the difference between its utility function (in terms of total rate) and payments, $v_i = u_i\left(\sum_{j=1}^{J} q_{ij} c_{ij}\right) - \sum_{j=1}^{J} p_j q_{ij}$.

Although the c_{ij} channel quality offset factor represents channel capacity in Example 3.11, it can be any increasing function of the channel strength depending on the specific application scenario.

We make the following assumptions throughout this section:

Assumption 1 (Utility functions) *For every user $i \in \mathcal{I}$, $u_i(x)$ is differentiable, increasing, and strictly concave in x. This is a standard way to model elastic data applications in network literature (see, e.g., (31)).*

Assumption 2 (Channel quality offsets and channel gains) *Channel quality offsets c_{ij} are drawn independently from continuous, possibly different utility distributions. In particular $Pr(c_{ij} = c_{kl}) = 0$ for any $i, k \in \mathcal{I},\ j, l \in \mathcal{J}$. The channel quality offset accounts for the effect that buying the same amount of resource from different providers typically has different effects on a user's quality of service. As Example 3.11 shows, channel quality offset c_{ij} may be a function of the channel gain h_{ij} between user i and provider j. In this case the assumption is fulfilled if channel gains are drawn from independent continuous probability distribution (e.g., Rayleigh, Rician, distance-based path-loss model).*

Assumption 3 (Atomic and price-taking users) *The demand for an atomic user is not infinitely small and can have an impact on providers' prices. Precise characterization of this impact is one of the focuses here. On the other hand, users are price-takers by the assumption of the two-stage game, and do not strategically influence prices.*

3.3.3 SOCIAL WELFARE OPTIMIZATION

Problem Formulation

Now let us formulate the social welfare maximization problem, which aims at maximizing the sum of payoffs of all participants, (users and providers). The social welfare problem is equivalent to maximizing the sum of users' utility functions, as we assume that the resource does not have value for the providers.[4]

The key result here is the uniqueness of the optimal solution of the social welfare maximization problem in terms of users' demands. For clarity of exposition, we define the following notation.

Definition 3.12 (Effective resource). Let $\boldsymbol{x} = [x_1 \ \cdots \ x_I]$ be the vector of *effective resources*, where $x_i(\boldsymbol{q}_i) = \sum_{j=1}^{J} q_{ij} c_{ij}$ is a function of user i's demand $\boldsymbol{q}_i = [q_{i1} \ldots q_{iJ}]$.

The social welfare optimization problem (SWO) is:

$$\textbf{SWO} : \max \ u(\boldsymbol{x}) = \sum_{i=1}^{I} u_i\left(x_i\right) \tag{3.51}$$

$$\text{subject to} \ \sum_{j=1}^{J} q_{ij} c_{ij} = x_i \ i \in \mathcal{I} \tag{3.52}$$

$$\sum_{i=1}^{I} q_{ij} = Q_j, \ j \in \mathcal{J} \tag{3.53}$$

$$\text{over } q_{ij}, x_i \geq 0 \ \forall i \in \mathcal{I}, j \in \mathcal{J}. \tag{3.54}$$

[4] Later, we will see that users need to pay providers for the resources, and the providers care about their revenue. However, the payments between users and providers are not considered in the social welfare optimization as they will cancel out each other.

For clarity we expressed the SWO in terms of two different variables: effective resource vector $\boldsymbol{x}$ and demand vector $\boldsymbol{q}$, even though the problem can be expressed entirely in terms of $\boldsymbol{q}$. In particular, a vector $\boldsymbol{q}$ uniquely determines a vector $\boldsymbol{x}$ through equations (3.52), i.e. we can write $\boldsymbol{x}$ as $\boldsymbol{x}(\boldsymbol{q})$. With some abuse of notation we will write $u(\boldsymbol{q})$ when we mean $u(\boldsymbol{x}(\boldsymbol{q}))$.

Uniqueness of Optimal Solution

Lemma 3.13 *The SWO problem has a unique optimal solution $\boldsymbol{x}^*$.*

Proof. Since $u_i(x_i)$ is strictly concave in x_i, then $u(\boldsymbol{x}) = \sum_{i=1}^{I} u_i(x_i)$ is strictly concave in $\boldsymbol{x}$. The feasible region defined by constraints (3.52)-(3.54) is convex. Hence, $u(\boldsymbol{x})$ has a unique optimal solution $\boldsymbol{x}^*$ subject to constraints (3.52)-(3.54). □

Even though $u_i(\cdot)$'s are strictly concave in x_i, they are not strictly concave in the demand vector $\boldsymbol{q_i}$. Hence, SWO is non-strictly concave in $\boldsymbol{q}$. It is well-known that a non-strictly concave maximization problem might have several different global optimizers (several different demand vectors $\boldsymbol{q}$ in our case) (see e.g. (32; 33)). In particular, one can choose c_{ij}'s, Q_j's, and $u_i(\cdot)$'s in such a way that a demand maximizing vector $\boldsymbol{q}^*$ of SWO is not unique. However, we can show that such cases arise with zero probability whenever channel offsets factors c_{ij}'s are independent random variables drawn from continuous distributions (see Assumption 2).

In the remainder of this section, we show that SWO has a unique maximizing demand vector $\boldsymbol{q}^*$ with probability 1. We begin by proving Lemma 3.15, which is an intermediate result stating that any two maximizing demand vectors of SWO must have different non-zero components. We then use it to prove the main result in Theorem 3.16.

To make our argument precise, we first define the support set of a demand vector $\boldsymbol{q}_i$ as follows.

Definition 3.14 **(Support set).** The support set $\hat{\mathcal{J}}_i(\boldsymbol{q}_i)$ of a demand vector $\boldsymbol{q}_i$ contains the indices of its non-zero entries:

$$\hat{\mathcal{J}}_i(\boldsymbol{q}_i) = \{j \in \mathcal{J} : q_{ij} > 0\}.$$

Given a demand vector $\boldsymbol{q}$, the ordered *collection* of support sets $\hat{\mathcal{J}}_1, \hat{\mathcal{J}}_2, \ldots, \hat{\mathcal{J}}_I$ is denoted by $\{\hat{\mathcal{J}}_i\}_{i=1}^{I}$. The support set contains providers that user i has strictly positive demand from.

Lemma 3.15 *Let $\boldsymbol{q}^*$ be an optimal solution of SWO (a maximizing demand vector) and $\{\hat{\mathcal{J}}_i\}_{i=1}^I$ be the corresponding collection of support sets. Then, $\boldsymbol{q}^*$ is almost surely*[5] *the unique maximizing demand vector corresponding to $\{\hat{\mathcal{J}}_i\}_{i=1}^I$.*

Proof. For a maximizing demand vector $\boldsymbol{q}^*$, equations (3.52)-(3.54) hold, with $\sum_{j=1}^J q_{ij}^* c_{ij} = x_i^*$. To prove the lemma, we will uniquely construct $\boldsymbol{q}^*$ from $\boldsymbol{x}^*$ and $\{\hat{\mathcal{J}}_i\}_{i=1}^I$.

We can divide the users into two categories. The *decided* users purchase from only one provider ($|\hat{\mathcal{J}}_i| = 1$), and the *undecided* users from several ($|\hat{\mathcal{J}}_i| > 1$). It is also possible that some users have zero demand to all providers, but without loss of generality, we treat such users as decided. Recall that for all users we have $x_i^* = \sum_{j=1}^J q_{ij}^* c_{ij}$. For a *decided* user i who purchases only from provider $\bar{\jmath}$, this reduces to $x_i^* = q_{i\bar{\jmath}}^* c_{i\bar{\jmath}}$, and the corresponding unique demand vector is $\boldsymbol{q}_i^* = [0 \cdots 0 \frac{x_i^*}{c_{i\bar{\jmath}}} 0 \cdots 0]$.

For undecided users, finding the unique $\boldsymbol{q}_i^*$ is less straightforward as there is more than one $\boldsymbol{q}_i$ such that $\sum_{j \in \hat{\mathcal{J}}_i} q_{ij} c_{ij} = x_i^*$. To show that the demand of undecided users is unique, we construct the bipartite graph representation (BGR) $\mathcal{G}$ of the undecided users' support sets as follows. We represent undecided users by circles, and providers of undecided users as squares. We place an edge (i, j) between a provider node j and a user node i if $j \in \hat{\mathcal{J}}_i$.

We give an example of a BGR in Fig. 3.8, where $\hat{\mathcal{J}}_1 = \{a, b, c\}$, $\hat{\mathcal{J}}_2 = \{b, d\}$, $\hat{\mathcal{J}}_3 = \{d, e, f\}$ and $\hat{\mathcal{J}}_4 = \{b, g\}$.

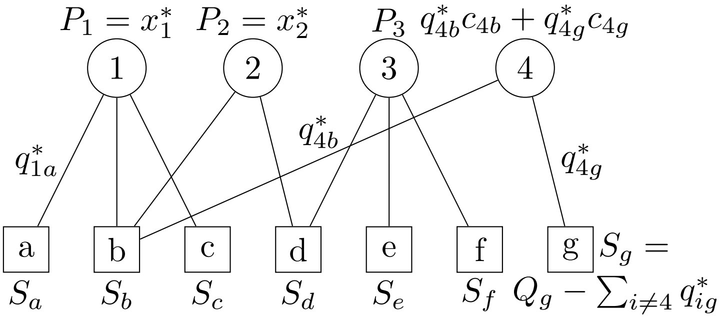

Figure 3.8: Bipartite graph representation

The BGR has the following properties[6] (see Fig. 3.8):

(a) The sum of effective resource on all the edges connected to user i is the optimal effective resource $x_i^* = \sum_{j \in \hat{\mathcal{J}}_i} q_{ij}^* c_{ij} = P_i$. Borrowing from coding theory and with some abuse of terminology, we call P_i the *check-sum* of user node i.

[5] This holds on the probability space defined by the distributions of c_{ij}'s.

[6] Fig. 3.8 shows a connected graph, but this need not be the case.

(b) The sum of all edges connected to provider node j equals to the difference between the supply Q_j and the demand from decided users who connect to provider j: $\sum_{i:(i,j)\in\mathcal{G}} q_{ij}^* = Q_j - \sum_{i:(i,j)\notin\mathcal{G}} q_{ij}^* = S_j$. We call S_j the *check-sum* of provider node j.

(c) With probability 1, the BGR does not contain any loops. For the proof of this point, see (34).

As it is the case in Fig. 3.8, the number of undecided users is smaller than the number of providers. This is a direct consequence of Property (c).

We can use the BGR to uniquely determine the demands of undecided users. Here we use Fig. 3.8 as an illustrative example. Consider the leaf node (a node with only one edge) g and edge q_{4g}^*. The BGR implies that user 4 is the only undecided customer of provider g. Since the demands of all decided users have been determined, then we know that $q_{4g}^* = S_g = Q_g - \sum_{i\neq 4} q_{ig}^*$. We can then remove edge q_{4g}^* and node g from the BGR, and update the check-sum value of node 4 to $P_4 = x_4^* - q_{4g}^* c_{4g}$. Now consider node 4 and edge q_{4b}^*. Since edge q_{4b}^* is now the only edge connecting with user node 4, we have $q_{4b}^* c_{4b} = P_4$ and hence $q_{4b}^* = P_4/c_{4b}$. Next we can consider node a, e, or f, and so on. Property (c) is crucial in this procedure since it guarantees that we can always find a leaf node in the reduced graph.

In each step of the algorithm, we determine the unique value of q_{ij}^* associated with the edge of one leaf. We can show that this value is independent of the order in which we pick the leaf nodes. So, we can construct unique demand vector $\boldsymbol{q}_i^*$ for each undecided user i. Together with the unique demand vectors of the decided users, we have found the unique maximizing demand vector $\boldsymbol{q}^*$ of SWO with support sets $\{\hat{\mathcal{J}}_i\}_{i=1}^I$. □

Theorem 3.16 *The SWO problem has a unique maximizing solution $\boldsymbol{q}^*$ with probability 1.*

To intuitively understand Theorem 3.16, assume there exist two maximizing demand vectors of SWO which. By Lemma 3.15, these two demand vectors have different supports sets. The support set of a non-trivial convex combination of any two non-negative vectors is the union of support sets of these two vectors. Hence, all convex combinations of two maximizing demand vectors of SWO, which are also maximizing demand vectors, have the same support. This is a contradiction to Lemma 3.15.

Given an optimal demand vector $\boldsymbol{q}^*$ of the SWO problem, there exists a unique corresponding Lagrangue multiplier vector $\boldsymbol{p}^*$, associated with the resource constraints of J providers (35). This $\boldsymbol{p}^*$ actually can be interpreted as the prices announced by the providers, which will be useful for understanding the following primal-dual algorithm.

3.3.4 PRIMAL-DUAL ALGORITHM

The previous analysis assumes that a centralized decision maker can perform network optimization with complete network information. This may not be true in practice. In this section we present a distributed primal-dual algorithm where providers and users only know local information and make local decisions in an iterative fashion. We will show that the primal-dual algorithm converges to a set containing the optimal solution of SWO. We can further show that this set contains only the unique optimal solution in most cases, regardless of the values of the updating rates. We first present the algorithm, and then the proof of its convergence.

Primal-dual algorithm

In this section, we will consider a continuous-time algorithm, where all the variables are functions of time. For compactness of exposition, we will sometimes write q_{ij} and p_j when we mean $q_{ij}(t)$ and $p_j(t)$, respectively. Their time derivatives $\frac{\partial q_{ij}}{\partial t}$ and $\frac{\partial p_j}{\partial t}$ will often be denoted by $\dot{q}_{ij}$ and $\dot{p}_j$. We denote by $\boldsymbol{q}^*$ and $\boldsymbol{p}^*$ the unique maximizer of SWO and the corresponding Lagrange multiplier vector, respectively.

To simplify the notation, we denote by $f_{ij}(t)$ or simply f_{ij} the marginal utility of user i with respect to q_{ij} when his demand vector is $\boldsymbol{q_i}(t)$:

$$f_{ij} = \frac{\partial u_i(\boldsymbol{q_i})}{\partial q_{ij}} = c_{ij}\frac{\partial u_i(x)}{\partial x}\Big|_{x=x_i=\sum_{j=1}^{J} q_{ij}c_{ij}}. \tag{3.55}$$

We will use f^*_{ij} to denote the value of $f_{ij}(t)$ evaluated at $\boldsymbol{q}^*_i$, the maximizing demand vector of user i. So, f^*_{ij} is a constant that is equal to a user's marginal utility at the global optimal solution of the SWO problem, as opposed to $f_{ij}(t)$ which indicates marginal utility at a particular time t. We also define $\nabla u_i(\boldsymbol{q_i}) = [f_{i1} \cdots f_{iJ}]^T$ and $\nabla u_i(\boldsymbol{q^*_i}) = [f^*_{i1} \cdots f^*_{iJ}]^T$, where all the vectors are column vectors.

We define $(x)^+ = \max(0, x)$ and

$$(x)^+_y = \begin{cases} x & y > 0 \\ (x)^+ & y \leq 0. \end{cases}$$

Another way to think of this notation is $(x)^+_y = x(1 - \mathbb{1}_{(-\infty,0]}(x)\mathbb{1}_{(-\infty,0]}(y))$, where $\mathbb{1}$ is the indicator function, i.e., $\mathbb{1}_A(x) = 1$ if $x \in A$, and 0 otherwise.

Motivated by the work in (36), we consider the following standard *primal-dual variable update algorithm*:

$$\dot{q}_{ij} = k^q_{ij}\left(f_{ij} - p_j\right)^+_{q_{ij}}, \; i \in \mathcal{I}, j \in \mathcal{J} \tag{3.56}$$

$$\dot{p}_j = k^p_j \left(\sum_{i=1}^{I} q_{ij} - Q_j\right)^+_{p_j}, \; j \in \mathcal{J}. \tag{3.57}$$

Here k_{ij}^p, k_j^p are the constants representing update rates. The update rule ensures that, when a variables of interest (q_{ij} or p_j) is already zero, it will not become negative even when the direction of the update (i.e. quantity in the parenthesis) is negative. The tuple $(\boldsymbol{q}(t), \boldsymbol{p}(t))$ controlled by equations (3.56) and (3.57) will be referred to as the *solution trajectory* of the differential equations system defined by (3.56) and (3.57).

The motivation for the proposed algorithm is quite natural. A provider increases its price when the demand is higher than its supply and decreases its price when the demand is lower. A user decreases his demand when a price is higher than his marginal utility and increases it when a price is lower. In essence, the algorithm is following the natural direction of market forces.

One key observation is that these updates can be implemented in a distributed fashion. The users only need to know the prices proposed by the providers. The providers only need to know the demand of the users for their own resource, and not for the resource of other providers. In particular, only user i needs to know his own channel offset parameters c_{ij}, $j \in \mathcal{J}$.

The first step to prove the algorithm's convergence is to construct a lower-bounded La Salle function $V(\boldsymbol{q}(t), \boldsymbol{p}(t))$ and show that its value is non-increasing for any solution trajectory $(\boldsymbol{q}(t), \boldsymbol{p}(t))$ that satisfies (3.56) and (3.57). This will ensure that $(\boldsymbol{q}(t), \boldsymbol{p}(t))$ converge to a set of values that keeps $V(\boldsymbol{q}(t), \boldsymbol{p}(t))$ constant.

Convergence of the primal-dual algorithm

We consider the following La Salle function:

$$V(\boldsymbol{q}(t), \boldsymbol{p}(t)) = V(t) = \sum_{i,j} \frac{1}{k_{ij}^q} \int_0^{q_{ij}(t)} (\beta - q_{ij}^*) d\beta + \sum_j \frac{1}{k_j^p} \int_0^{p_j(t)} (\beta - p_j^*) d\beta. \quad (3.58)$$

It can be shown that $V(\boldsymbol{q}(t), \boldsymbol{p}(t)) \geq V(\boldsymbol{q}^*, \boldsymbol{p}^*)$, i.e., V is bounded from below. This ensures that if the function V is non-increasing, it will eventually reach a constant value (which may or may not be the global minimum $V(\boldsymbol{q}^*, \boldsymbol{p}^*)$).

The derivative of V *along the solution trajectories of the system*, denoted by $\dot{V} = \frac{\partial V}{\partial t}$, is given by:

$$\dot{V}(t) = \sum_{i,j} \frac{\partial V}{\partial q_{ij}} \dot{q}_{ij} + \sum_j \frac{\partial V}{\partial p_j} \dot{p}_j.$$

Lemma 3.17 *The value of the La Salle function V is non-increasing along the solution trajectory, defined by (3.56) and (3.57), i.e. $\dot{V}(t) \leq 0$.*

The key idea of proving Lemma 3.17 is to manipulate the expression for $\dot{V}$ and show that it can be reduced to the following form:

$$\dot{V} \leq \sum_i \left(\sum_j (q_{ij}(t) - q_{ij}^*)(f_{ij}(t) - f_{ij}^*) \right) + \sum_{i,j} (q_{ij}(t) - q_{ij}^*)(f_{ij}^* - p_j^*). \quad (3.59)$$

Using concavity of $u_i's$ and properties of the global optimal solution $(\boldsymbol{q}^*, \boldsymbol{p}^*)$, we can show that individual elements of the summations in (3.59) are non-positive.

Combining Lemma 3.17 and the La Salle's *invariance principle* (Theorem 4.4 of (37)) we can prove the following:

Proposition 3.18 *The pair* $(\boldsymbol{q}(t), \boldsymbol{p}(t))$ *converges to the invariant set* $V_L = \{\boldsymbol{q}(t), \boldsymbol{p}(t) : \dot{V}(\boldsymbol{q}(t), \boldsymbol{p}(t)) = 0\}$ *as* $t \to \infty$.

It is clear that the invariant set V_L contains the solution trajectory that has the value of the unique maximizer of SWO $(\boldsymbol{q}^*(t), \boldsymbol{p}^*(t)) = (\boldsymbol{q}^*, \boldsymbol{p}^*)$ for all t, since $\dot{V}(\boldsymbol{q}^*, \boldsymbol{p}^*) = 0$. However, it may contain other points as well. When the trajectory $(\boldsymbol{q}(t), \boldsymbol{p}(t))$ enters the invariant set, it either reaches its minimum (i.e., by converging to the unique equilibrium point $(\boldsymbol{q}^*, \boldsymbol{p}^*)$), or it gets stuck permanently in some limit cycle. In either case, the trajectory will be confined to a subset of $V_L = \{(\boldsymbol{q}(t), \boldsymbol{p}(t)) : \dot{V}(\boldsymbol{q}(t), \boldsymbol{p}(t)) = 0\}$.

The good news is that we can indeed show that the invariant set V_L contains only the equilibrium point $(\boldsymbol{q}^*, \boldsymbol{p}^*)$. This can be done in two steps. First, we show that the set V_L has only one element for the majority of the network scenarios, without any restrictions on the variable update rates. Second, we provide a sufficient condition on the update rates so that the global convergences to the unique equilibrium point is also guaranteed in the remaining scenarios. For details, see (34).

3.3.5 NUMERICAL RESULTS

For numerical results, we extend the setup from Example 3.11, where the resource being sold is the fraction of time allocated to exclusive use of the providers' frequency band, i.e., $Q_j = 1$ for $j \in \mathcal{J}$. We take the bandwidth of the providers to be $W_j = 20$MHz, $j \in \mathcal{J}$. User i's utility function is $a_i \log(1 + \sum_{j=1}^J q_{ij} c_{ij})$, where we compute the spectral efficiency c_{ij} from the Shannon formula $\frac{1}{2} W \log(1 + \frac{E_b/N_0}{W} |h_{ij}|^2)$, q_{ij} is the allocated time fraction, E_b/N_0 is the ratio of transmit power to thermal noise, and a_i is the individual willingness to pay factor taken to be the same across users. The channel gain amplitudes $|h_{ij}| = \frac{\xi_{ij}}{d_{ij}^{\alpha/2}}$ follow Rayleigh fading, where ξ_{ij} is a Rayleigh distributed random variable with parameter 1, and $\alpha = 3$ is the outdoor power distance loss. We choose the parameters so that the c_{ij} of a user is on average around 3.5Mbps when the distance is 50m, and around 60Mbps when the distance is 5m. The average signal-to-noise ratio $E_b/(N_0 d^\alpha)$ at 5m is around 25dB. We

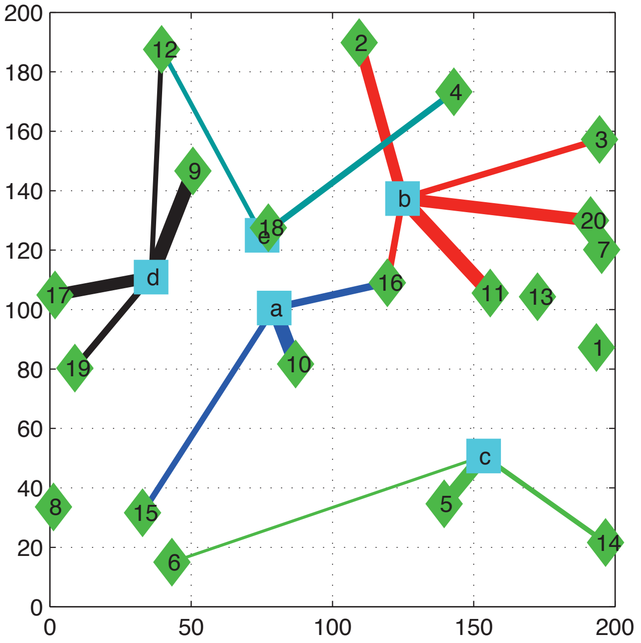

Figure 3.9: Example of equilibrium user-provider association

assume perfect modulation and coding choices such that the communication rates come from a continuum of values. The users are uniformly placed in a 200m by 200m area. We want to emphasize that the above parameters are chosen for illustrative purposes only. Our theory applies to any number of providers, any number of users, any type of channel attenuation models, and arbitrary network topologies.

We first consider a single instantiation with 20 users and 5 providers. In Fig. 3.9, we show the user-provider association at the equilibrium for a particular realization of channel gains, where the thickness of the link indicates the amount of resource purchased. The users are labeled by numbers (1-20), and the providers are labeled by letters (*a*-*e*). This figure shows two undecided users (12 and 16), and that certain users (1,7,13, and 8) do not purchase any resource at equilibrium. Fig. 3.10 shows the evolution of the mismatch between supply and demand as well as the prices of the five providers. The equilibrium prices reflect the competition among users: in Fig. 3.9 we see that provider *b* has the most customers, so it is not surprising that its price is the highest, as seen in Fig. 3.10.

We next consider the convergence time of the discrete time version of the primal-dual algorithm. We fix the number of providers to be 5, and change the number of users from 20 to 100. For each parameter, we run 200 experiments with randomly generated user and provider locations and plot the average speed of convergence. The convergence is defined

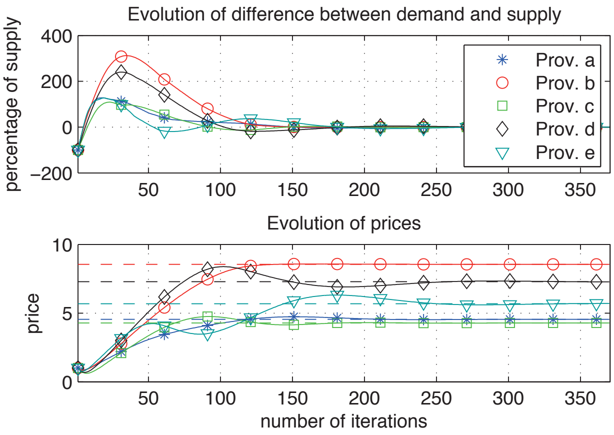

Figure 3.10: Evolution of the primal-dual algorithm

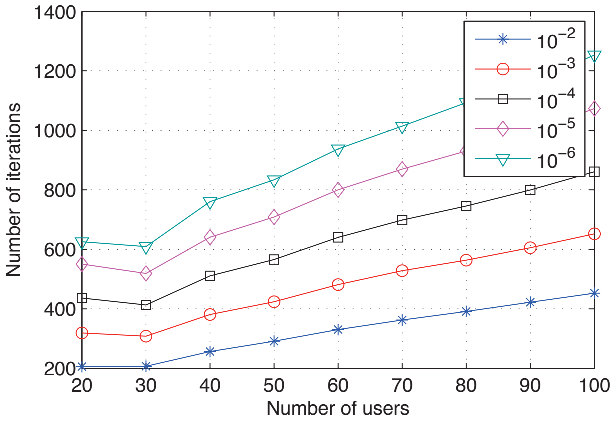

Figure 3.11: Average time to convergence, varying ϵ

as the number of iterations after which the difference between supply and demand is no larger than ϵQ_j. Fig. 3.11 shows the average convergence time for different values of ϵ. In general, 200 to 400 iterations are needed for convergence with $\epsilon = 10^{-2}$, and 100-200 more iterations to get to 10^{-3}.

3.4 CHAPTER SUMMARY

Pricing can facilitate a wireless network operator to achieve the social optimality in a distributed fashion. In this chapter, we discuss the theory and applications of such social optimal pricing.

We start by introducing the basic concepts of convex sets and convex functions, as well as several operations that preserve the convexity of sets and functions. This helps us to define the convex optimization, which concerns the minimization of convex functions over convex sets. Although convexity is the watershed between "easy" and "complex" optimization problems, the convexity (or the non-convexity) of an optimization problem may be changed through proper transformations. To illustrate this example, we discuss the Geometric Programming, which is a class of non-convex optimization that can be easily converted to convex optimization (and thus have nice solutions structures). Then we move on with the key theory of this chapter: duality-based distributed algorithm for solving a convex problem. We first introduce the Lagrange dual problem formulation of a primal optimization problem (not necessarily convex), and characterize the KKT necessary conditions under which a primal dual feasible solution pair is optimal for both the primal and dual problems. Most importantly, the dual variables have the nice interpretations of shadow prices (hence we are talking about it here in this pricing book), and dual problem can be solved in a distributed fashion through the subgradient method using shadow prices as coordinating signals. For more detailed discussions of the duality-based optimization, please see (38).

To better illustrate the idea of dual-based distributed optimization, we show two applications in wireless networks. The first application concerns the resource allocation and scheduling for wireless video streaming in a single cell network. The problem is rather complicated as the video source optimization is often discrete and thus not convex. However, we decompose the network optimization problem into phrases: average resource allocation, video source adaptation, and multiuser deadline oriented scheduling. We show that it is possible to only consider the "semi-elastic" nature of today's video sources to perform average resource allocation via the dual-based resource allocation. We make the discussions more concrete by considering different formulations in both wireless uplink and downlink streaming. In the second application, we consider the resource allocation among multiple wireless service providers serving overlapping areas. As users have the flexibility of purchasing from one or more service providers, the social welfare maximization turns out to be convex but not strictly convex. Despite the fact that in general there can be multiple global optimal solutions for non strictly convex optimization problem, we show that in our problem the global optimal solution is unique with probability 1. Then we design a primal-dual algorithm, which is a generalization of the dual-based algorithm, such that the users and providers can coordinate in a distributed fashion and converge to the unique global optimal solution. For more details especially mathematical proofs related to the two applications, see (26; 34).

CHAPTER 4
Monopoly and Price Discriminations

In this chapter, we will move away from social optimal pricing and look at the issue of profit maximization instead. In particular, we will look at case where a single service provider dominates the market. The provider can charge a single optimized price to all consumers, or he can charge different prices based on the consumer types if he knows such information. Such price differentiation may often significantly improve the provider's profit.

We first introduce the theory of monopoly pricing and price discriminations, and illustrate the theory through two examples. In the first example, we consider the revenue management in cognitive underlay networks, where the spectrum owner can control the demand elasticity of the users by adjusting total available bandwidth and tolerable interference level. In the second example, we study how to design an incentive-compatible pricing menu under incomplete information, in order to achieve the same maximum revenue under price differentiation with complete information. We also discuss how to optimize the price differentiation parameters when the service provider is constrained in terms of the number of prices it can select.

4.1 THEORY: MONOPOLY PRICING

In this section, we cover the basic concepts of monopoly pricing, i.e., the pricing theory in a monopoly market. Our discussions follow closely those in (9; 39), where readers can find more in-depth discussions.

4.1.1 WHAT IS MONOPOLY?

Before discussing the basic theory in monopoly pricing, we need to define what "monopoly" means. Etymology suggests that a "*monopoly*" is a *single seller*, the only firm in its industry. But such a vague answer may cause serious confusions. Consider Apple Inc., which is obviously the only firm that sells iPhone; however, however, Apple is *not* the only firm that sells mobile phones. Thus, whether Apple is a single seller depends on how narrowly you define the market.

In order to avoid such confusions, we will use a different definition relying on the *monopoly power* or *market power*, a widely used concept in economics. As defined in many economic literatures (e.g., (9)), monopoly power or market power is the ability of a firm to affect market prices through its actions. A firm with monopoly power is referred to as a *monopoly* or *monopolist*.[1] More specifically,

Definition 4.1 **Monopoly Power.** A firm has monopoly power, if and only if (i) it faces a downward-sloping demand curve for its product, and (ii) it has no supply curve.

The first condition implies that a monopolist is never perfectly competitive. That is, he is able to set the market price so as to shape the demand. The second condition implies that a monopolist never faces a going market price. In fact, the market price is a consequence of the monopolist's actions, rather than a datum to which he must react. By this definition, Apple is obviously a monopoly (in the iPhone market), since it can lower the price of iPhone to increase the sales of iPhones (i.e., the demand curve for iPhone slops downward). The competitive soy farmer who can increase/decrease his output and still sell it all at the going market price is not a monopoly (in the soy market).

In what follows, we will study how a monopolist chooses price and quantity, and what is the profit consequences of these choices.

4.1.2 PROFIT MAXIMIZATION BASED ON DEMAND ELASTICITY

Let P denote the market price a monopolist chooses. Let $Q \triangleq D(P)$ denote the downward-sloping demand curve the monopolist faces (or the best quantity the monopolist chooses). A key question is: *how should the monopolist choose a market price to maximize his profit*? We will show that the answer depends greatly on the demand curve the monopolist faces. In particular, it depends on the *price elasticity* of demand defined in Section 2.2.

We first consider the monopolist's total revenue $\pi(P)$ under a particular market price P. Formally, we have

$$\pi(P) \triangleq P \cdot Q, \quad \text{where} \;\; Q = D(P). \tag{4.1}$$

It is easy to check that $\pi(P)$ is a concave function of P, and therefore the optimal price P^* that maximizes $\pi(P)$ is given by the first-order condition:

$$\frac{\mathrm{d}\pi(P)}{\mathrm{d}P} = Q + P \cdot \frac{\mathrm{d}Q}{\mathrm{d}P} = 0, \tag{4.2}$$

which leads to the following optimality condition:

$$\frac{P \cdot \triangle Q}{Q \cdot \triangle P} + 1 = 0, \tag{4.3}$$

where $\triangle Q$ and $\triangle P$ are small changes in quality and price, respectively.

Next we show that the above revenue maximization problem is closely related to the following problem: *how much does the monopolist have to lower his price to sell one more*

[1] The only seller in its industry is obviously a good example of a monopoly, but note that it is not the only example.

product? The answer actually leads to the price elasticity of demand defined in Section 2.2. Recall that the price elasticity of demand is defined as the change in demand that results from one unit increase in price, and given by the formula:

$$\eta \triangleq \frac{\triangle Q/Q}{\triangle P/P} = \frac{P \cdot \triangle Q}{Q \cdot \triangle P}, \tag{4.4}$$

or equivalently,

$$\triangle P = \frac{P \cdot \triangle Q}{Q \cdot \eta}, \tag{4.5}$$

which shows how much the market price must change to sell additional $\triangle Q$ of product. Note that to sell an extra product, the change in price $\triangle P$ must be negative, which can be confirmed by the fact that η is always negative. Thus, we can also write the absolute value of $\triangle P$ as $|\triangle P| = \frac{P \cdot \triangle Q}{Q \cdot |\eta|} = -\frac{P \cdot \triangle Q}{Q \cdot \eta}$.

Now we consider the consequences of selling an additional product. That is, how much the monopolist's total revenue changes by selling an additional product. Specifically, there are two factors affecting the monopolist's revenue π. On one hand, the monopolist gains an additional revenue $P \cdot \triangle Q$ by selling an additional unit $\triangle Q$ of product at price P. On the other hand, the monopolist suffers a revenue loss $|\triangle P| \cdot Q$, since the price for the previous Q products is decreased by $|\triangle P|$. Thus, the net change in the monopolist's revenue is

$$\triangle \pi \triangleq P \cdot \triangle Q - |\triangle P| \cdot Q. \tag{4.6}$$

Substitute (4.5) into (4.6), we can rewrite the revenue change as

$$\triangle \pi = P \cdot \triangle Q - \frac{P \cdot \triangle Q}{Q \cdot |\eta|} \cdot Q = P \cdot \triangle Q \cdot \left(1 - \frac{1}{|\eta|}\right), \tag{4.7}$$

which shows how much the monopolist's revenue changes by selling an additional unit $\triangle Q$ of product. Note that as we set $\triangle Q = 1$, this is essentially the monopolist's *marginal revenue* (MR), i.e., the change in his revenue by selling one extra product.

The formula (4.7) shows that $\triangle \pi < 0$ if $|\eta| < 1$. This implies that a monopolist would never lower the price (or increase the quantity equivalently) when $|\eta| < 1$. In other words, a monopolist must operate on a market price or quantity such that $|\eta| \geq 1$. If in addition there is no other cost, the optimal price or quantity satisfies $\triangle \pi = 0$, which implies that $|\eta| = 1$ or $1 + \eta = 0$. This is obviously equivalent to the first-order condition in (4.2). Note that if there is certain cost (e.g., the producing cost), the optimal price or quantity satisfies $\triangle \pi = \triangle C$, i.e., the change in revenue equals to the change in cost.

When $|\eta| > 1$, we say that the demand curve is *elastic*; when $|\eta| < 1$ we say that the demand curve is *inelastic*. An immediate observation is that a profit-maximization monopolist will increase the price whenever the demand curve is inelastic. Thus, our conclusion for a monopolist's operation is given in the following theorem (9).

Theorem 4.2 *A monopolist always operates on the elastic portion of the demand curve.*

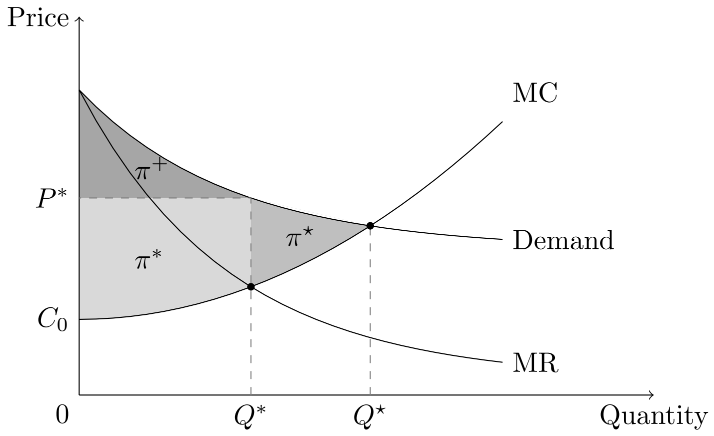

Figure 4.1: Increasing the monopolist's profit by eliminating consumer surplus. When charging a single monopoly price P^* to all consumers, the monopolist's profit is shown by the shaded area π^* and consumer surplus is shown by the shaded area π^+. Suppose the monopolist charges each consumer the most that he would be willing to pay for each product that he buys, the monopolist's profit is now $\pi^* + \pi^+$, and the consumers get zero surplus.

4.2 THEORY: FIRST, SECOND, AND THIRD DEGREE PRICE DISCRIMINATIONS

The analysis of monopoly pricing in Section 4.1 assumes that the monopolist will sell all of products at a single price. This section deals with a monopolist that can engage in *price discrimination*, i.e., charging different prices for the same product. A common goal of price discrimination is to raise the monopolist's revenue by reducing consumer surplus.

Basically, with price discrimination, the monopolist can either charge different prices to a single consumer (for different units of products), or charge uniform but different prices to different groups of consumers. In this section, we will discuss the motivations, conditions, and means of price discrimination.

4.2.1 AN ILLUSTRATIVE EXAMPLE

We first show that it is possible for a monopolist to improve his revenue and eliminate consumer surplus by price discrimination.

Consider a simple example with a monopolist facing a downward-sloping demand curve $Q = D(P)$ and a production cost $C(Q)$. The demand curve (D), marginal revenue (MR) and marginal cost (MC) are illustrated in Figure 4.1. The marginal revenue function intersects the marginal cost function when the monopolist chooses the monopoly quantity

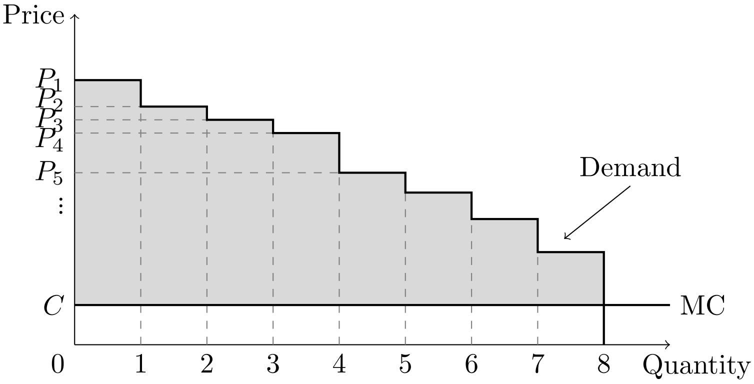

Figure 4.2: Under first-degree price discrimination, the consumer is charged a price P_1 for the first product he purchases, P_2 for the second product he purchases, and so on.

Q^* and sells each product at the monopoly price P^*. The monopolist's total profit is equal to the area labeled π^*, and the consumer surplus is equal to the area labeled π^+.

Note that by charging the single monopoly price P^* to all consumers, the monopolist does not collect all the consumer surplus, since consumers are still left with a surplus of π^+. Now suppose that the monopolist can charge different prices for different units of products. Then, he can collect the consumer surplus π^+ by charging the demand price $P(D)$ for each successive unit of product, i.e., charging each consumer the most that he would be willing to pay for each additional product that he buys. Suppose in addition that the monopolist increases the monopoly quantity Q^* to $Q^\star$, i.e., the interaction of the marginal cost and the demand curve. Then, he can not only collect the consumer surplus in the area π^+, but also the surplus in the area $\pi^\star$. In this case, the monopolist collects all the social surplus $\pi^* + \pi^+ + \pi^\star$, which is also the maximum social surplus the monopolist can achieve.[2] This is essentially the first-degree price discrimination, which will be discussed soon later.

The above example shows that a monopolist can increase its profit by charging different prices to a consumer or to different consumers. Next we will show that the amount of additional profit the monopolist can extract from consumers depends on the information he has about the consumers. As a result, there are three types of price discriminations: first-, second-, and third-degree, which will be discussed in the following sections.

4.2.2 FIRST DEGREE PRICE DISCRIMINATION

With the **first-degree price discrimination**, or *perfect price discrimination*, the monopolist charges each consumer the most that he would be willing to pay for each product

[2] Here the social surplus is defined as the total profits of all involved players. Since the payment cancels out each other, the social surplus is related only to the quantity.

that he buys (9). It requires that the monopolist knows exactly the maximum price that every consumer is willing to pay for each product, i.e., the full knowledge about every consumer demand curve. In this case, the monopolist captures all the market surplus, and the consumer gets zero surplus.

Figure 4.2 illustrates the first-degree price discrimination, where the consumer is willing to pay a maximum price P_1 for the first product, P_2 for the second product, and so on. When the price is between $[P_2, P_1]$, the total demand is 1; while when the price is between $[P_4, P_3]$, the total demand is 3; and so on. The demand curve can be represented by the downward stepped curve shown in the figure. Under the first-degree price discrimination, the consumer is charged his maximum willingness to pay for successive products, i.e., P_1 for the first product, P_2 for the second product, and so on. Obviously, the monopolist capture all the market surplus (shown shaded).

In practice, however, it is difficult or even impossible for the monopolist to obtain the complete demand information. Thus, the first-degree price discrimination is primarily theoretical and seldom exists in reality.

4.2.3 SECOND DEGREE PRICE DISCRIMINATION

With the **second-degree price discrimination**, or *declining block pricing*, the monopolist offers a *bundle of prices* to the consumers, with different prices for different blocks of units (9). Recall that in the first-degree price discrimination, a different price is set for every different unit. In this sense, the second-degree price discrimination can be viewed as a more limited version of the first-degree price discrimination.

Figure 4.3 illustrates a second-degree price discrimination, where the monopolist offers a bundle of prices $\{P_1, P^*, P_2\}$ with $P_1 > P^* > P_2$ to the consumer. The price P_1 is for the first block (the first Q_1 units) of products, P^* is for the second block (from Q_1 to Q^*) of products, and P_2 is for the third block (from Q^* to Q_2). That is, the consumer pays P_1 for each unit (of product) up to Q_1 units, P^* for each unit between Q_1 and Q^* units, and P_2 for each unit between Q^* and Q_2 units. In effective, we can view the monopolist offers a discount $\frac{P^*}{P_1}$ for the purchasing quantity above Q_1, and an additional discount $\frac{P_2}{P^*}$ for the purchasing quantity above Q^*. Without price discrimination, the monopolist's maximum profit is $P^*Q^* - C(Q^*)$. With second-degree price discrimination, the monopolist's profit is shown shaded, which is obviously larger than $P^*Q^* - C(Q^*)$. Furthermore, the monopolist's profit with the second-degree price discrimination is less than that with the first-degree price discrimination, and the gap is shown as summation of the blank regions denoted by δ. We can see that if the number of prices the monopolist offers is sufficiently large, and the region of each block is close sufficiently to the demand curve, then the second-degree price discrimination converges effectively to the first-degree price discrimination (the blank regions vanish).

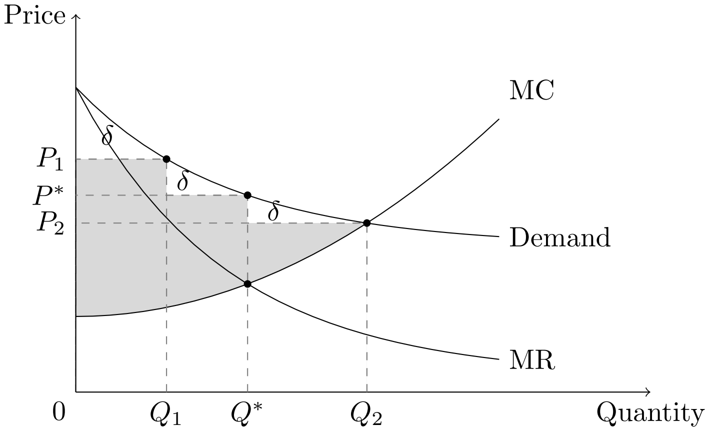

Figure 4.3: Under second-degree price discrimination, the consumer is charged a price P_1 for the first block (from 0 to Q_1) of products, P^* for the second block (from Q_1 to Q^*), and P_2 for the third block (from Q^* to Q_2).

It is worth noting that the second-degree price discrimination does not require the monopolist to know the complete information of every consumer demand curve. For example, the second-degree price discrimination illustrated in Figure 4.3 requires some particular points on the consumer demand curve only, i.e., (Q_1, P_1), (Q_2, P_2), and (Q^*, P^*). Obviously, the more information the monopolist knows, the higher profit he can absorb from the consumer.

4.2.4 THIRD DEGREE PRICE DISCRIMINATION

A monopolist that performs the first- or second-degree price discrimination knows something about the demand curve of every *individual* consumer, and benefits from this information by charging the consumer different prices. A natural question is whether (and how, if so) the monopolist discriminates the price to increase his profit, if he has no information on the individual demand curve (but knows from experience that different groups of consumers have different total demand curves)?

The answer is YES, and it actually leads to the third and the most common form of price discrimination, the **third-degree price discrimination**, or *multi-market price discrimination* (9). Simply speaking, third-degree price discrimination usually occurs when a monopolist faces two (or more) *identifiably* different groups of consumers having different (downward-sloping) total demand curves, and knows the total demand curve of every group but not the individual demand curve of every consumer. In this case, the monopolist can potentially increase his profit by setting different prices for different groups.

To apply third-degree price discrimination, the monopolist first uses some characteristic of consumers to segment consumers into groups. Then he picks different prices for the different groups that maximize his profit. In this process, it is implicitly assumed that the monopolist is able to sort consumers into groups (i.e., identify the type of each consumer), and thus consumers in the group with a higher price cannot purchase in the lower-priced market. A simple example of this kind of price discrimination is that the Disney Park offers different ticket prices to children, adults, and elders.

To show how a monopolist discriminates among groups of consumers, we consider a simple scenario, where the monopolist sorts consumers into two groups (two markets). The total demand curves of different markets are different. The monopolist needs to decide the price P_i for each market $i \in \{1, 2\}$ (and therefore the sales $Q_i = D_i(P_i)$ in each market i). To maximize his own profit, the monopolist must decide:

- Whether to charge the same price or different prices in different markets?
- Which market should get the lower price if the firm charges different prices?
- What the relation between the prices of two markets?

Under prices P_1 and P_2, the monopolist's total profit $\pi(P_1, P_2)$ is given by

$$\pi(P_1, P_2) \triangleq P_1 \cdot Q_1 + P_2 \cdot Q_2 - C(Q_1 + Q_2), \qquad \text{where } \; Q_i = D_i(P_i). \tag{4.8}$$

It is easy to check that $\pi(P_1, P_2)$ is a concave function of vector (P_1, P_2), and therefore the optimal price vector (P_1^*, P_2^*) that maximizes $\pi(P_1, P_2)$ is given by the first-order condition:

$$\frac{\mathrm{d}\pi(P_1, P_2)}{\mathrm{d}P_i} = Q_i + P_i \cdot \frac{\mathrm{d}Q_i}{\mathrm{d}P_i} - C'(Q_1 + Q_2) \cdot \frac{\mathrm{d}Q_i}{\mathrm{d}P_i} = 0, \quad i = 1, 2, \tag{4.9}$$

which leads to the following optimality condition:

$$C'(Q_1 + Q_2) = P_i + Q_i \cdot \frac{\mathrm{d}P_i}{\mathrm{d}Q_i} = P_i \cdot \left(1 - \frac{1}{|\eta_i|}\right), \quad i = 1, 2, \tag{4.10}$$

where $\eta_i \triangleq \frac{P_i}{Q_i} \frac{\mathrm{d}Q_i}{\mathrm{d}P_i}$ is the price elasticity of market i. Note that the right-hand side in (4.10) is in fact the marginal revenue in each market i. Thus, the above condition suggests that under the optimality, the marginal revenue in each market i equals to the marginal cost. This further leads to

$$P_1 \cdot \left(1 - \frac{1}{|\eta_1|}\right) = P_2 \cdot \left(1 - \frac{1}{|\eta_2|}\right). \tag{4.11}$$

Intuitively, we can see from (4.10) and (4.11) that under the optimality, the monopolist sets an optimal price vector (P_1^*, P_2^*) such that the marginal revenues in all market are the same, and all equal to the marginal cost.

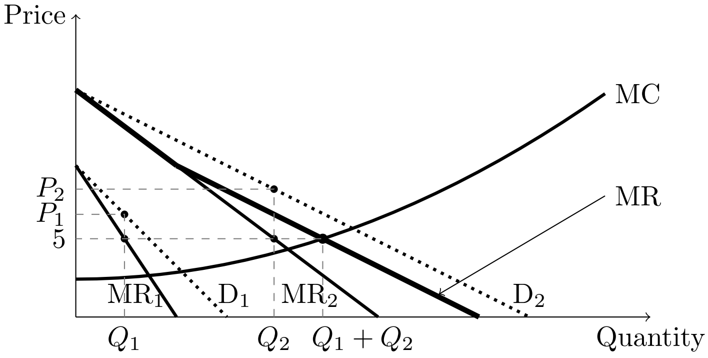

Figure 4.4: Under third-degree price discrimination, the monopolist charges a higher price P_1 from market 1 (with a lower price elasticity), and a lower price P_2 from market 2 (with a higher price elasticity), so that he achieves the same marginal revenues from both markets.

The optimality conditions in (4.10) and (4.11) provider answers to the above three questions. First, we have $P_1 \neq P_2$ as long as $\eta_1 \neq \eta_2$. That is, the monopolist will charge different prices when different groups of customers have different elasticities. Second, we have: (i) $P_1 < P_2$ if $|\eta_1| > |\eta_2|$, and (ii) $P_1 > P_2$ if $|\eta_1| < |\eta_2|$. That is, the market with the higher price elasticity will get a lower price. Third, the relation between the prices of two markets is given by (4.11).

Figure 4.4 provides a graphic interpretation for the above optimal solution. The MR1 and MR2 curves are the marginal revenue curves in both markets, and the D1 and D2 curves are the demand curves in both markets. The MR curve is obtained by summing MR1 and MR2 horizontally. That is, for any price, read the corresponding quantities of MR1 and MR2, and then add these to get the corresponding quantity on MR.

The monopolist, to maximize his profit, can equalize his marginal cost and both marginal revenues by choosing the quantity where his marginal cost curve MC crosses the MR curve (see from (4.10) and (4.11)). This means that he produces a total of $Q_1 + Q_2$ products, so that his marginal cost is 5 per unit of product. He sells Q_1 of these products in market 1 and Q_2 in market 2, such that his marginal revenue is 5 per unit in each market. Once the monopolist chooses Q_1 and Q_2, the price for each market are impliedly given by the inverse demand function, i.e., $P_i = D_i^{-1}(Q_i)$, $i = 1, 2$. From Figure 4.4, we can find that the prices for different markets are different, and the market with the higher price elasticity (market 2 in Figure 4.4) gets the lower price (P_2 in Figure 4.4).

We can now summarize the necessary conditions to make the third-degree price discrimination profitable (9).

- **Monopoly power** – The firm must have the monopoly power (or market power) to affect market price, which means we don't see price discrimination in perfectly competitive markets.

- **Market segmentation** – The firm must be able to split the market into different groups of consumers, and also be able to identify the type of each consumer.

- **Elasticity of demand** – There must be a different price elasticity of demand for different markets. This allows the firm to charge a higher price to the market with a relatively inelastic demand and a lower price to those with a relatively elastic demand. The firm will then be able to extract more consumer surplus which will lead to additional profit.

4.3 APPLICATION I: REVENUE MANAGEMENT WITH POWER-INTERFERENCE ELASTICITY

4.3.1 BACKGROUND

Recent advances in cognitive radio technology have enabled wireless devices and networks to locate and exploit under-utilized spectrum. Most existing studies in this field have focused on how to improve the network performance by allowing primary (licensed) and secondary (unlicensed) users to efficiently and flexibly share the spectrum. However, in practice a primary user may not have enough incentives to share the spectrum with the secondary users in a fear of potential degradation of its Quality of Services.

In this application, we consider a scenario where a primary user can collect revenue from the secondary users. This provides the necessary incentive for the primary user to share the spectrum with the secondary users. We consider a spectrum underlay network where the secondary users can transmit simultaneously with the primary user, subject to total bandwidth and tolerable interference constraints. The primary user collects the revenue through charging the secondary users proportional to their generated interference. The primary user can maximize its revenue by adjusting the bandwidth and tolerable interference constraints, as long as certain physical constraints are not violated.

The key for the primary user to maximize revenue is to consider the new concept of *power-interference elasticity*. In the existing literature, Shenker (40) was the first to introduce the concept of elasticity into networking, where he refereed *elasticity* as the applications' ability to adapt their sending rates according to the available resource. Yuksel and Kalyanaraman (41) further developed this idea and defined a *utility-bandwidth elasticity*, which is always nonnegative. The authors also calculated the corresponding optimal pricing to maximize total utility. Marbach and Berry (42) proposed an optimal pricing scheme by price discriminating users with respect to the *power-price elasticity*, but the market equilibrium is not necessarily reached. By contrast, the *power-interference elasticity* introduced here is about the relationship between the demand (power) and the environment (interference), instead of the utility and the resource (bandwidth).

4.3.2 SYSTEM MODEL

We consider a primary licensed user who owns a wireless spectrum of bandwidth $\bar{B}$. The primary user is able to tolerate a maximum interference temperature $\bar{P}/\bar{B}$. The primary user can allow the secondary users to share a total spectrum of bandwidth B with a total generated interference power at the primary user's receiver as P, as long as $B \leq \bar{B}$ and $P/B \leq \bar{P}/\bar{B}$.

There exists a set $\mathcal{M} = \{1, \ldots, M\}$ of secondary users, who transmit to the same secondary base station. We focus on the case where the secondary base station is co-located with the primary user's receiver. In this case, the total power received from the secondary users at the base station is the same as the total tolerable interference at the primary user's receiver.

A user $i's$ valuation of the resource is characterized by a utility function $U_i\left(\theta_i, \gamma_i(\boldsymbol{p})\right)$, which is increasing, strictly concave, and twice continuously differentiable of its signal-to-interference plus noise ratio (SINR) at the base station

$$\gamma_i\left(\boldsymbol{p}\right) = \frac{p_i}{n_0 + p_{-i}/B}, \tag{4.12}$$

where n_0 is the background noise power density, p_i is the power received from user i at the secondary base station, $p_{-i} = \sum_{j \neq i, j \in \mathcal{M}} p_j$, $\boldsymbol{p} = (p_i, p_{-i})$, and θ_i is a user-dependent parameter. We assume that the users choose spread spectrum transmission over the entire allowed bandwidth B, and thus the interference of a user i is the total received power from other secondary users scaled by the bandwidth. To simplify the notations, we denote $U_i'\left(\theta, \gamma\right) = \partial U_i\left(\theta, \gamma\right)/\partial \gamma$ and $U_i''\left(\theta, \gamma\right) = \partial^2 U_i\left(\theta, \gamma\right)/\partial \gamma^2$.

The key constraint is that the total power allocation satisfies

$$\sum_{i \in \mathcal{M}} p_i = p_i + p_{-i} \leq P, \tag{4.13}$$

where P is the total tolerable interference at the primary user's receiver.

The power allocation is performed in a distributed fashion via pricing. The primary user announces a positive unit price π. Each secondary user i submits the demand $p_i^*(p_{-i}, \pi)$ to maximize his surplus

$$\begin{aligned} p_i^*(\pi, p_{-i}, \theta_i) &= \arg\max_{\hat{p}_i \geq 0} S_i\left(\pi, \hat{p}_i, p_{-i}, \theta_i\right) \\ &= \arg\max_{\hat{p}_i \geq 0} U_i\left(\theta_i, \gamma_i\left(\hat{p}_i, p_{-i}\right)\right) - \pi \hat{p}_i. \end{aligned} \tag{4.14}$$

It is clear that secondary users play a noncooperative game here, since a user decision p_i^* depends on the other users' choice p_{-i}. Details of such game theoretical analysis for a finite system can be found in (30).

Here, we focus on a *large system limit* where there are many secondary users sharing a large bandwidth. Mathematically, we focus on the asymptotic behavior as P, B, M go to infinity, while keeping P/M and P/B fixed. We can show that if the utilities are asymptotically sublinear with respect to γ_i (i.e., $\lim_{\gamma_i\to\infty}\frac{1}{\gamma_i}U_i\left(\theta_i,\gamma_i\right)=0$ for all i) and θ_i is continuously distributed in a nonnegative interval $\left[\underline{\theta},\bar{\theta}\right]$, then we can always find a price *market clearing price* π^* such that $\sum_{i=1}^{M}p_i^*\left(p_{-i}^*,\pi^*\right)=P$. More importantly, a user i's SINR at the market equilibrium is

$$\gamma_i\left(p_i\right)=\frac{p_i}{n_0+P/B}, \tag{4.15}$$

i.e., the interference experienced by any secondary user i is a user-independent constant P/B.

The sublinear requirement can be satisfied by many common utility functions, e.g., $\theta\ln\left(\gamma\right)$, $\theta\ln\left(1+\gamma\right)$, $\theta\gamma^{\alpha}$ $\left(\alpha\in(0,1)\right)$, and any upper-bounded utility such as $1-e^{-\theta\gamma}$. The user-independent property of the interference makes the large system limit analytically more attractable than the finite system. In (30) we showed that this large system limit can be reached with moderate number of users (less than 20) in practice.

Next will restrict our study to revenue management at the market clearing price π^*. The results can be easily generalized to the case where the primary user can choose a price that does not clear the market, in which case the primary user may increase the revenue by further price discriminating among users (8).

4.3.3 ELASTICITIES IN A LARGE SYSTEM

To simplify the notation, let us write $I=P/B$. A user i's surplus is a function of the price π, interference I, and power allocation p:

$$S\left(\pi,p,I,\theta\right)=U\left(\theta,\frac{p}{n_0+I}\right)-\pi p. \tag{4.16}$$

Here we consider a generic user and omit the user index i.

The *power demand function* $p^o\left(\pi,I,\theta\right)$ (i.e., optimal choice of power for a user to maximize its surplus) is

$$\begin{aligned} p^o\left(\pi,I,\theta\right) &= \arg\max_{\hat{p}\geq 0} S\left(\pi,\hat{p},I,\theta\right) \\ &= \begin{cases} \left(n_0+I\right)g_\theta^{-1}\left(\pi\left(n_0+I\right)\right), & U'\left(\theta,0\right)>\pi\left(n_0+I\right) \\ 0, & U'\left(\theta,0\right)\leq\pi\left(n_0+I\right) \end{cases}, \end{aligned} \tag{4.17}$$

where $g_\theta\left(\gamma\right)=U'\left(\theta,\gamma\right)$ and the superscript -1 denotes the inverse function. The corresponding *SINR demand function* is $\gamma^o\left(\pi,I,\theta\right)=p^o\left(\pi,I,\theta\right)/\left(n_0+I\right)$.

Proposition 4.3 *In a large system,*

$$\frac{\partial\gamma^o\left(\pi,I,\theta\right)}{\partial I}=\frac{\pi}{U''\left(\theta,\gamma^o\left(\pi,I,\theta\right)\right)}<0 \tag{4.18}$$

for $\pi < U'(\theta, 0)/(n_0 + I)$.

Proposition 4.3 shows that a user will always choose a smaller SINR when the interference increases. This is, however, not the case for power demand $p^o(\pi, I, \theta)$.

To facilitate further discussion, we first introduce the *power-price elasticity* used in economics:

Definition 4.4 Power-price elasticity in a large system is

$$e_\pi(p^o(\pi, I, \theta)) = \frac{\partial p^o(\pi, I, \theta)/p^o(\pi, I, \theta)}{\partial \pi/\pi}. \tag{4.19}$$

Since the utility function is concave in γ, thus the power demand curve has a negative slope, and the power-price elasticity is always negative (43). The elasticity characteristic $L_{e_\pi(p^o(\pi,I,\theta))}$ is defined as (8):

$$L_{e_\pi(p^o(\pi,I,\theta))} = \begin{cases} \text{elastic}, & e_\pi \in (-\infty, -1) \\ \text{unitary elastic}, & e_\pi = -1 \\ \text{inelastic}, & e_\pi \in (-1, 0) \end{cases} \tag{4.20}$$

Next we define the *power-interference elasticity* which is new in this subsection:

Definition 4.5 Power-interference elasticity in a large system is

$$e_I(p^o(\pi, I, \theta)) = \frac{\partial p^o(\pi, I, \theta)/p^o(\pi, I, \theta)}{\partial I/(n_0 + I)}. \tag{4.21}$$

The power-interference elasticity shows how the power demand changes with respect to the change of interference. Although it is possible to give a similar definition of the elasticity characteristic of $e_I(p^o(\pi, I, \theta))$ as in (4.20), we are more interested in the sign of (4.21) since it is not necessarily negative as shown below.

Proposition 4.6 *In the large system limit,*

$$e_I(p^o(\pi, I, \theta)) = e_\pi(p^o(\pi, I, \theta)) + 1 \tag{4.22}$$

for $\pi < U'(\theta, 0)/(n_0 + I)$.

Proposition 4.6 shows the simple relationship between the power-interference and the power-price elasticities. Moreover, the sign of $e_I(p^o(\pi, I, \theta))$ depends simply on the elasticity characteristic of $e_\pi(p^o(\pi, I, \theta))$.

Table 4.1: Elasticity of power demand in a large system

$U(\theta,\gamma)$	Power-price elasticity		Power-interference elasticity	
$\theta\ln(\gamma)$	-1	unitary elastic	0	zero
$\theta\ln(\gamma+1)$	$-\frac{\theta}{\theta-\pi(n_0+I)}<-1$	elastic	$-\frac{\pi(n_0+I)}{\theta-\pi(n_0+I)}<0$	negtive
$\theta\gamma^{\alpha}\ (\alpha\in(0,1))$	$-\frac{1}{1-\alpha}<-1$	elastic	$-\frac{\alpha}{1-\alpha}<0$	negtive
$1-e^{-\theta\gamma}$	$-\frac{1}{\ln\left(\frac{\theta}{\pi(n_0+I)}\right)}<0$	depends	$-\frac{1-\ln\left(\frac{\theta}{\pi(n_0+I)}\right)}{\ln\left(\frac{\theta}{\pi(n_0+I)}\right)}<1$	depends

Table 4.1 shows both elasticities for some common utility functions, when $\pi < U'(\theta,0)/(n_0+I)$.

Similarly, we can define the aggregated power-price elasticity (aggregate power-interference elasticity, respectively) in a large system as $e_{A\pi}\left(\sum_{i=1}^{M} p_i^o(\pi,I,\theta_i)\right)$ $\left(e_{AI}\left(\sum_{i=1}^{M} p_i^o(\pi,I,\theta_i)\right)\right.$, respectively) by substituting $p^o(\pi,I,\theta)$ in Definition 4.4 (Definition 4.5, respectively) with $\sum_{i=1}^{M} p_i^o(\pi,I,,\theta_i)$. It is easy to show that

$$e_{AI}\left(\sum_{i=1}^{M} p_i^o(\pi,I,\theta_i)\right) = e_{A\pi}\left(\sum_{i=1}^{M} p_i^o(\pi,I,\theta_i)\right)+1. \tag{4.23}$$

4.3.4 REVENUE MAXIMIZATION

First consider the impact of B and P on secondary users' total utility $\sum_{i\in\mathcal{M}} U_i(\theta_i,\gamma_i)$.

Theorem 4.7 *In the large system limit, the secondary users' total utility is maximized at the market clearing price. Moreover, the total utility and the active users' SINRs are increasing in P and B.*

This shows that allowing more bandwidth or more tolerable interference to the secondary users will increase the secondary users' QoS. This is intuitive, as more bandwidth means less interference and more tolerable interference (from the primary user's point of view) means higher transmission power (for the secondary users), and both will lead to high SINRs of the secondary users.

The impact of B and P on the revenue of the primary user, however, is more complicated.

Theorem 4.8 *In the large system limit,*

$$\partial R/\partial B \begin{cases} >0, & e_{AI}\left(\sum_{i=1}^{M} p_i^o(\pi^*,I,\theta_i)\right)<0 \\ =0, & e_{AI}\left(\sum_{i=1}^{M} p_i^o(\pi^*,I,\theta_i)\right)=0 \\ <0, & e_{AI}\left(\sum_{i=1}^{M} p_i^o(\pi^*,I,\theta_i)\right)>0 \end{cases}, \tag{4.24}$$

where π^* is the market clearing price.

Theorem 4.9 *In the large system limit,*

$$\partial R/\partial P \begin{cases} > 0, & e_{AI}\left(\sum_{i=1}^{M} p_i^o\left(\pi^*, I, \theta_i\right)\right) < 0 \\ = 0, & e_{AI}\left(\sum_{i=1}^{M} p_i^o\left(\pi^*, I, \theta_i\right)\right) = 0 \\ < 0, & e_{AI}\left(\sum_{i=1}^{M} p_i^o\left(\pi^*, I, \theta_i\right)\right) > 0 \end{cases},$$

where π^* is the market clearing price. If n_0 is negligible compared with interference for all i, then revenue R does not change with P, i.e., $\partial R/\partial P = 0$.

Theorems 4.8 and 4.9 show that the aggregated power-interference elasticity is important for the primary user's revenue maximization decision. If it is negative, the primary user should increase P and B until it becomes zero, or the resource is exhausted, or the interference temperature is reached. If it is positive, the manger should decrease P and B until it becomes zero, or the last user is indifferent in joining or quiting the system (but is still active), or the interference temperature is reached. Finally, if it is zero, nothing needs to be done since the revenue is already maximized.

If we assume that the primary user is the only seller in the spectrum market of a certain time period at certain geographic area, and the secondary users can not transfer the usage rights among themselves, then the primary user can further improve the revenue if he can separate the users into groups by the individual power-interference elasticities. The improvement can be achieved by a third-degree price discrimination as introduced in Section 4.2.4. One thing to notice is that the secondary users' behavior in our problem depends heavily on the interference level, thus the primary user should be careful in assigning resources for different groups, i.e., the values of P, B, and the ratio P/B. This is rather unique for our problem.

Theorem 4.9 also shows that when n_0 is very small compared with interference, the primary user could decrease P (and thus decrease interference temperature) while keeping the revenue unchanged. However this will eventually breaks down when P/B is close to n_0.

4.4 APPLICATION II: DIFFERENTIAL PRICING WITH INCOMPLETE INFORMATION AND LIMITED PRICE CHOICES

4.4.1 SYSTEM MODEL

We consider a network with a total of S divisible resource (which can be in the form of rate, bandwidth, power, time slot, etc.). The resource is allocated by a monopolistic service provider to a set $\mathcal{I} = \{1, \ldots, I\}$ of user groups. Each group $i \in \mathcal{I}$ has N_i homogeneous users[3] with the same utility function:

$$u_i(s_i) = \theta_i \ln(1 + s_i), \tag{4.25}$$

where s_i is the allocated resource to one user in group i and θ_i represents the willingness to pay of group i. The logarithmic utility function is commonly used to model the proportionally fair resource allocation in communication networks (see (44) for detailed explanations). Without loss of generality, we assume that $\theta_1 > \theta_2 > \cdots > \theta_I$. In other words, group 1 contains users with the highest valuation, and group I contains users with the lowest valuation.

We consider two types of information structures:

1. **Complete information**: the service provider knows each user's utility function. Though the complete information is a very strong assumption, it is the most frequently studied scenario the network pricing literature. The significance of studying the complete information is two-fold. It serves as the benchmark of practical designs and provides important insights for the incomplete information analysis.

2. **Incomplete information**: the service provider knows the total number of groups I, the number of users in each group $N_i, i \in \mathcal{I}$, and the utility function of each group $u_i, i \in \mathcal{I}$. It does not know which user belongs to which group. Such assumption in our discrete setting is analogous to that the service provider knows only the users' types distribution in a continuum case. Such statistical information can be obtained through long term observations of a stationary user population.

The interaction between the service provider and users can be characterized as a two-stage Stackelberg model shown in Fig. 4.5. The service provider publishes the pricing scheme in Stage 1, and users respond with their demands in Stage 2. The users want to maximize their surpluses by optimizing their demands according to the pricing scheme. The service provider wants to maximize its revenue by setting the right pricing scheme to induce desirable demands from users. Since the service provider has a limited total resource, he must guarantee that the total demand from users is no larger than what he can supply.

Next we will discuss how the service provider chooses different pricing schemes under different information scenarios and complexity requirements.

4.4.2 COMPLETE PRICE DIFFERENTIATION UNDER COMPLETE INFORMATION

We first consider the complete information case. Since the service provider knows the utility and the identity of each user, it is possible to maximize the revenue by charging a different

[3] A special case is N_i=1 for each group, *i.e.,* all users in the network are different.

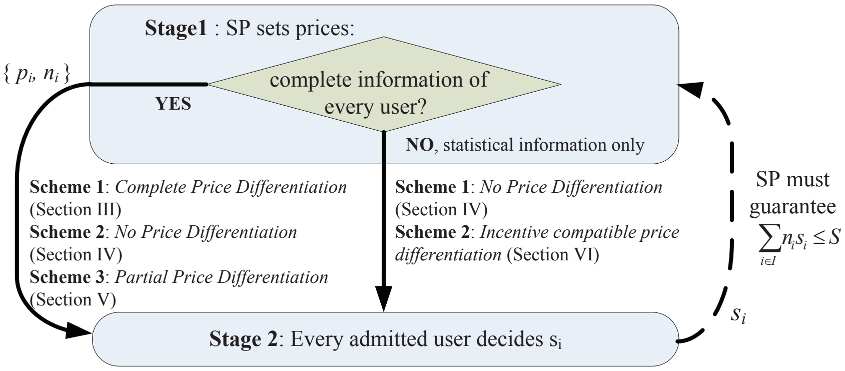
Stage1 : SP sets prices:
complete information of every user?
{ p_i, n_i }
YES
NO, statistical information only
Scheme 1: Complete Price Differentiation (Section III)
Scheme 2: No Price Differentiation (Section IV)
Scheme 3: Partial Price Differentiation (Section V)
Scheme 1: No Price Differentiation (Section IV)
Scheme 2: Incentive compatible price differentiation (Section VI)
SP must guarantee $\sum_{i\in I} n_i s_i \leq S$
s_i
Stage 2: Every admitted user decides s_i

Figure 4.5: A two-stage leader and follower model

price to each group of users. The analysis will be based on backward induction, starting from Stage 2 and then moving to Stage 1.

User's Surplus Maximization Problem in Stage 2

If a user in group i has been admitted into the network and offered a linear price p_i in Stage 1, then it solves the following surplus maximization problem,

$$\underset{s_i \geq 0}{\text{maximize}}\ u_i(s_i) - p_i s_i, \tag{4.26}$$

which leads to the following unique optimal demand

$$s_i(p_i) = \left(\frac{\theta_i}{p_i} - 1\right)^+, \quad \text{where } (\cdot)^+ \triangleq \max(\cdot, 0). \tag{4.27}$$

Remark 4.10 The analysis of the Stage 2 user surplus maximization problem is the same for all pricing schemes. The result in (4.27) will be also used in Sections 4.4.3, 4.4.4 and 4.4.5.

Service Provider's Pricing and Admission Control Problem in Stage 1

In Stage 1, the service provider maximizes its revenue by choosing the price p_i and the admitted user number n_i for each group i subject to the limited total resource S. The key idea is to perform a Complete Price differentiation (CP) scheme, i.e., charging each group with a different price.

$$
\begin{aligned}
CP: \quad & \underset{\boldsymbol{p}\geq 0, \boldsymbol{s}\geq 0, \boldsymbol{n}}{\text{maximize}} && \sum_{i\in\mathcal{I}} n_i p_i s_i && (4.28)\\
& \text{subject to} && s_i = \left(\frac{\theta_i}{p_i} - 1\right)^+ , \quad i \in \mathcal{I}, && (4.29)\\
& && n_i \in \{0, \ldots, N_i\} \ , \quad i \in \mathcal{I}, && (4.30)\\
& && \sum_{i\in\mathcal{I}} n_i s_i \leq S. && (4.31)
\end{aligned}
$$

where $\boldsymbol{p} \triangleq \{p_i, i \in \mathcal{I}\}$, $\boldsymbol{s} \triangleq \{s_i, i \in \mathcal{I}\}$, and $\boldsymbol{n} \triangleq \{n_i, i \in \mathcal{I}\}$. We use bold symbols to denote vectors in the sequel. Constraint (4.29) is the solution of the Stage 2 user surplus maximization problem in (4.27). Constraint (4.30) denotes the admission control decision, and constraint (4.31) represents the total limited resource in the network.

CP Problem is not straightforward to solve, since it is a non-convex optimization problem with a non-convex objective function (summation of products of n_i and p_i), a coupled constraint (4.31), and integer variables $\boldsymbol{n}$. However, it is possible to convert it into an equivalent convex formulation through a series of transformations, and thus the problem can be solved efficiently.

First, we can remove the $(\cdot)^+$ sign in constraint (4.29) by realizing the fact that there is no need to set p_i higher than θ_i for users in group i; users in group i already demand zero resource and generate zero revenue when $p_i = \theta_i$. This means that we can rewrite constraint (4.29) as

$$
p_i = \frac{\theta_i}{s_i + 1} \text{ and } s_i \geq 0, i \in \mathcal{I}. \tag{4.32}
$$

Plugging (4.32) into (4.28), then the objective function becomes $\sum_{i\in\mathcal{I}} n_i \frac{\theta_i s_i}{s_i+1}$. We can further decompose the CP Problem in the following two subproblems:

1. *Resource allocation*: for a fixed admission control decision $\boldsymbol{n}$, solve for the optimal resource allocation $\boldsymbol{s}$.

$$
\begin{aligned}
CP_1: \quad & \underset{\boldsymbol{s}\geq 0}{\text{maximize}} && \sum_{i\in\mathcal{I}} n_i \frac{\theta_i s_i}{s_i + 1}\\
& \text{subject to} && \sum_{i\in\mathcal{I}} n_i s_i \leq S. && (4.33)
\end{aligned}
$$

Denote the solution of CP_1 as $\boldsymbol{s}^* = (s_i^*(\boldsymbol{n}), \forall i \in \mathcal{I})$. We further maximize the revenue of the integer admission control variables $\boldsymbol{n}$.

2. *Admission control*:

$$
\begin{aligned}
CP_2: \quad & \underset{\boldsymbol{n}}{\text{maximize}} && \sum_{i\in\mathcal{I}} n_i \frac{\theta_i s_i^*(\boldsymbol{n})}{s_i^*(\boldsymbol{n}) + 1} && (4.34)\\
& \text{subject to} && n_i \in \{0, \ldots, N_i\} \ , \quad i \in \mathcal{I}
\end{aligned}
$$

Let us first solve CP_1 Subproblem in $\boldsymbol{s}$. Note that it is a convex optimization problem. By using Lagrange multiplier technique, we can get the first-order necessary and sufficient condition:

$$s_i^*(\lambda) = \left(\sqrt{\frac{\theta_i}{\lambda}} - 1\right)^+, \tag{4.35}$$

where λ is the Lagrange multiplier corresponding to the resource constraint (4.33).

Meanwhile, we note the resource constraint (4.33) must hold with equality, since the objective is strictly increasing function in s_i. Thus, by plugging (4.35) into (4.33), we have

$$\sum_{i \in \mathcal{I}} n_i \left(\sqrt{\frac{\theta_i}{\lambda}} - 1\right)^+ = S. \tag{4.36}$$

This weighted water-filling problem (where $\frac{1}{\sqrt{\lambda}}$ can be viewed as the water level) in general has no closed-form solution for λ. However, we can efficiently determine the optimal solution λ^* by exploiting the special structure of our problem. Note that since $\theta_1 > \cdots > \theta_I$, then λ^* must satisfy the following condition:

$$\sum_{i=1}^{K^{cp}} n_i \left(\sqrt{\frac{\theta_i}{\lambda^*}} - 1\right) = S, \tag{4.37}$$

for a group index threshold value K^{cp} satisfying

$$\frac{\theta_{K^{cp}}}{\lambda^*} > 1 \text{ and } \frac{\theta_{K^{cp}+1}}{\lambda^*} \leq 1. \tag{4.38}$$

In other words, only groups with index no larger than K_{cp} will be allocated the positive resource. This property leads to the simple Algorithm 2 to compute λ^* and group index threshold K^{cp}: we start by assuming $K^{cp} = I$ and compute λ. If (4.38) is not satisfied, we decrease K^{cp} by one and recompute λ until (4.38) is satisfied. Since $\theta_1 > \lambda(1) = (\frac{n_1}{s+n_1})^2\theta_1$, Algorithm 2 always converges and returns the unique values of K^{cp} and λ^*. The total complexity is $\mathcal{O}(I)$, i.e., linear in the number of user groups (not the number of users).

With K^{cp} and λ^*, the solution of the resource allocation problem can be written as

$$s_i^* = \begin{cases} \sqrt{\frac{\theta_i}{\lambda^*}} - 1, & i = 1, \ldots, K^{cp}; \\ 0, & \text{otherwise.} \end{cases} \tag{4.39}$$

For the ease of discussions, we introduce a new notion of the *effective market*, which denotes all the groups allocated non-zero resource. For resource allocation subproblem CP_1, the threshold K^{cp} describes the size of the effective market. All groups with indices no larger than K^{cp} are *effective groups*, and users in these groups as *effective users*. An example of the effective market is illustrated in Fig. 4.6.

Algorithm 2 Solving the Resource Allocation Problem CP_1

```
function $CP(\{n_i, \theta_i\}_{i\in\mathcal{I}}, S)$
    $k \leftarrow I$, $\lambda(k) \leftarrow \left(\frac{\sum_{i=1}^{k} n_i\sqrt{\theta_i}}{S+\sum_{i=1}^{k} n_i}\right)^2$
    while $\theta_k \le \lambda(k)$ do
        $k \leftarrow k-1$, $\lambda(k) \leftarrow \left(\frac{\sum_{i=1}^{k} n_i\sqrt{\theta_i}}{S+\sum_{i=1}^{k} n_i}\right)^2$
    end while
    $K^{cp} \leftarrow k$, $\lambda^* \leftarrow \lambda(k)$
    return $K^{cp}$, $\lambda^*$
end function
```

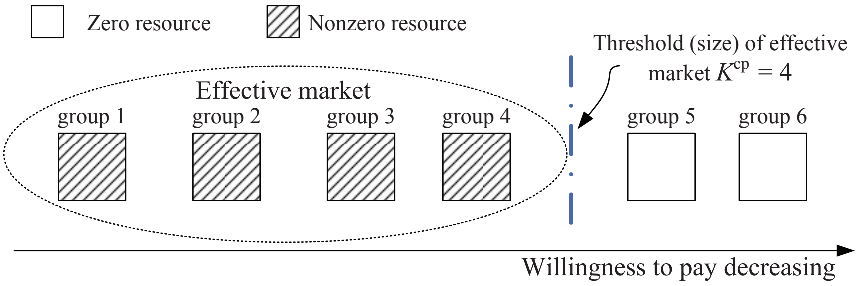

Figure 4.6: A 6-group example for effective market: the willingness to pays decrease from group 1 to group 6. The effective market threshold can be obtained by Algorithm 2, and is 4 in this example.

Now let us solve the admission control subproblem CP_2. Denote the objective (4.34) as $R_{cp}(\boldsymbol{n})$, by (4.39), then $R_{cp}(\boldsymbol{n}) \triangleq \sum_{i=1}^{K^{cp}} n_i\left(\sqrt{\frac{\theta_i}{\lambda^*(\boldsymbol{n})}} - 1\right)\sqrt{\theta_i\lambda^*(\boldsymbol{n})}$. We first relax the integer domain constraint of n_i as $n_i \in [0, N_i]$. Since (4.37), by taking the derivative of the objective function $R_{cp}(\boldsymbol{n})$ with respect to n_i, we have

$$\frac{\partial R_{cp}(\boldsymbol{n})}{\partial n_i} = \left(\sum_{i=1}^{K^{cp}} n_i\left(\sqrt{\frac{\theta_i}{\lambda^*(\boldsymbol{n})}} - 1\right)\right)\frac{\partial\sqrt{\theta_i\lambda^*(\boldsymbol{n})}}{\partial n_i}, \tag{4.40}$$

Also from (4.37), we have $\lambda^*=\left(\frac{\sum_{i=i}^{K^{cp}} n_i\sqrt{\theta_i}}{S+\sum_{i=1}^{K^{cp}} n_i}\right)^2$, thus $\frac{\partial\sqrt{\lambda^*(\boldsymbol{n})}}{\partial n_i} > 0$, for $i = 1,\ldots,K^{cp}$, and $\frac{\partial\sqrt{\lambda^*(\boldsymbol{n})}}{\partial n_i} = 0$, for $i = K^{cp}+1,\ldots,I$. This means that the objective $R_{cp}(\boldsymbol{n})$ is strictly increasing in n_i for all $i = 1,\ldots,K^{cp}$, thus it is optimal to admit all users in the effective market. The admission decisions for the groups not in the effective market is irrelevant to the optimization, since those users consume zero resource. Therefore, one of the optimal solutions of CP_1 Subproblem is $n_i^* = N_i$ for all $i \in \mathcal{I}$. Solving CP_1 and CP_2 Subproblems leads to the optimal solution of CP Problem:

Theorem 4.11 *There exists an optimal solution of CP Problem that satisfies the following conditions:*

- *All users are admitted: $n_i^* = N_i$ for all $i \in \mathcal{I}$.*
- *There exist a value λ^* and a group index threshold $K^{cp} \leq I$, such that only the top K^{cp} groups of users receive positive resource allocations,*

$$s_i^* = \begin{cases} \sqrt{\frac{\theta_i}{\lambda^*}} - 1, & i = 1, \ldots, K^{cp}; \\ 0, & \text{otherwise.} \end{cases}$$

with the prices

$$p_i^* = \begin{cases} \sqrt{\theta_i \lambda^*}, & i = 1, \ldots, K^{cp}; \\ \theta_i, & \text{otherwise.} \end{cases}$$

The values of λ^* and K^{cp} can be computed as in Algorithm 2 by setting $n_i = N_i$, for all $i \in \mathcal{I}$.

Theorem 4.11 provides the right economic intuition: service provider maximizes its revenue by charging a higher price to users with a higher willingness to pay. It is easy to check that $p_i > p_j$ for any $i < j$. The small willingness to pay users are excluded from the markets.

4.4.3 SINGLE PRICING SCHEME

We just showed that the CP scheme is the optimal pricing scheme to maximize the revenue under complete information. However, such a complicated pricing scheme is of high implementational complexity. Here we study the single pricing scheme. It is clear that the scheme will in general suffer a revenue loss comparing with the CP scheme. We will try to characterize the impact of various system parameters on such revenue loss.

Let us first formulate the Single Pricing (SP) problem.

$$\begin{aligned} SP: \quad & \underset{p \geq 0,\ \boldsymbol{n}}{\text{maximize}} && p \sum_{i \in \mathcal{I}} n_i s_i \\ & \text{subject to} && s_i = \left(\frac{\theta_i}{p} - 1 \right)^+, \quad i \in \mathcal{I} \\ & && n_i \in \{0, \ldots, N_i\} \quad , \quad i \in \mathcal{I} \\ & && \sum_{i \in \mathcal{I}} n_i s_i \leq S. \end{aligned}$$

Comparing with CP Problem in Section 4.4.2, here the service provider charges a single price p to all groups of users. After a similar transformation as in Section 4.4.2, we can show that the optimal single price satisfies the following the weighted water-filling condition

$$\sum_{i \in \mathcal{I}} N_i \left(\frac{\theta_i}{p} - 1 \right)^+ = S.$$

Thus we can obtain the following solution that shares a similar structure as complete price differentiation.

Theorem 4.12 *There exists an optimal solution of SP Problem that satisfies the following conditions:*

- *All users are admitted: $n_i^* = N_i$, for all $i \in \mathcal{I}$.*
- *There exist a price p^* and a group index threshold $K^{sp} \leq I$, such that only the top K^{sp} groups of users receive positive resource allocations,*

$$s_i^* = \begin{cases} \frac{\theta_i}{p^*} - 1, & i = 1, 2, \dots, K^{sp}, \\ 0, & \text{otherwise,} \end{cases}$$

with the price

$$p^* = p(K^{sp}) = \frac{\sum_{i=1}^{K^{sp}} N_i \theta_i}{S + \sum_{i=1}^{K^{sp}} N_i}.$$

The value of K^{sp} and p^* can be computed as in Algorithm 3.

Algorithm 3 Search the threshold of the SP Problem

```
function $SP(\{N_i, \theta_i\}_{i \in \mathcal{I}}, S)$
    $k \leftarrow I$, $p(k) \leftarrow \frac{\sum_{i=1}^{k} N_i \theta_i}{S + \sum_{i=1}^{k} N_i}$
    while $\theta_k \leq p(k)$ do
        $k \leftarrow k - 1$, $p(k) \leftarrow \frac{\sum_{i=1}^{k} N_i \theta_i}{S + \sum_{i=1}^{k} N_i}$
    end while
    $K^{sp} \leftarrow k$, $p^* \leftarrow p(k)$
    return $K^{sp}$, $p^*$
end function
```

4.4.4 PARTIAL PRICE DIFFERENTIATION UNDER COMPLETE INFORMATION

For a service provider facing thousands of user types, it is often impractical to design a price choice for each user type. The reasons behind this, as discussed in (45), are mainly high system overheads and customers' aversion. However, the single pricing scheme may suffer a considerable revenue loss compared with the complete price differentiation. How to achieve a good tradeoff between the implementational complexity and the total revenue? In reality, we usually see that the service provider offers only a few pricing plans for the entire users population; we term it as the *partial price differentiation* scheme. In this section, we will answer the following question: if the service provider is constrained to maintain a limited number of prices, $p^1, \ldots, p^J$, $J \leq I$, then what is the optimal pricing strategy and the maximum revenue? Concretely, the Partial Price differentiation (PP) problem is formulated as follows.

$$
\begin{aligned}
PP: \quad \underset{n_i, p_i, s_i, p^j, a_i^j}{\text{maximize}} \quad & \sum_{i \in \mathcal{I}} n_i p_i s_i \\
\text{subject to} \quad & s_i = \left(\frac{\theta_i}{p_i} - 1\right)^+, \; \forall\, i \in \mathcal{I}, && (4.41)\\
& n_i \in \{0, \ldots, N_i\}, \; \forall\, i \in \mathcal{I}, && (4.42)\\
& \sum_{i \in \mathcal{I}} n_i s_i \leq S, && (4.43)\\
& p_i = \sum_{j \in \mathcal{J}} a_i^j p^j, && (4.44)\\
& \sum_{j \in \mathcal{J}} a_i^j = 1, \; a_i^j \in \{0,1\}, \forall\, i \in \mathcal{I}. && (4.45)
\end{aligned}
$$

Here $\mathcal{J}$ denotes the set $\{1, 2, \ldots, J\}$. Since we consider the complete information scenario in this section, the service provider can choose the price charged to each group, thus constraints (4.41) – (4.43) are the same as in CP Problem. Constraints (4.44) and (4.45) mean that p_i charged to each group i is one of J choices from the set $\{p^j, j \in \mathcal{J}\}$. For convenience, we define *cluster* $\mathcal{C}^j \triangleq \{i \,|\, a_i^j = 1\}$, $j \in \mathcal{J}$, which is a set of groups charged with the same price p^j. We use superscript j to denote clusters, and subscript i to denote groups through this section. We term the binary variables $\boldsymbol{a} \triangleq \{a_i^j,\; j \in \mathcal{J},\; i \in \mathcal{I}\}$ as the *partition*, which determines which cluster each group belongs to.

PP Problem is a combinatorial optimization problem, and is more difficult than the previous CP and SP Problems. On the other hand, we notice that this PP Problem formulation includes the CP scheme ($J = I$) and the SP scheme scenario ($J = 1$) as special cases. The insights we obtained from solving these two special cases in Sections 4.4.2 and 4.4.3 will be helpful to solve the general PP problem.

To solve PP Problem, we decompose and tackle it in three levels. In the lowest level-3, we determine the pricing and resource allocation for each cluster, given a fixed partition and fixed resource allocation among clusters. In level-2, we compute the optimal resource allocation among clusters, given a fixed partition. In level-1, we optimize the partition among groups.

Level-3: Pricing and resource allocation in each cluster

For a fixed partition $\boldsymbol{a}$ and a cluster resource allocation $\boldsymbol{s} \triangleq \{s^j\}_{j\in\mathcal{J}}$, we focus the pricing and resource allocation problems within each cluster $\mathcal{C}^j$, $j \in \mathcal{J}$:

$$\begin{aligned} \text{Level-3:} \quad \underset{n_i, s_i, p^j}{\text{maximize}} \quad & \sum_{i\in C^j} n_i p^j s_i \\ \text{subject to} \quad & s_i = \left(\frac{\theta_i}{p^j} - 1\right)^+, \quad \forall\, i \in \mathcal{C}^j, \\ & n_i \le N_i, \quad \forall\, i \in \mathcal{C}^j, \\ & \sum_{i\in\mathcal{C}^j} n_i s_i \le s^j. \end{aligned}$$

Level-3 Subproblem coincides with the SP scheme discussed in Section 4.4.3, since all groups within the same cluster $\mathcal{C}^j$ are charged with a single price p^j. We can then directly apply the results in Theorem 3 to solve the Level-3 problem. We denote the effective market threshold[4] for cluster $\mathcal{C}^j$ as K^j, which can be computed in Algorithm 3. An illustrative example is shown in Fig. 4.7, where the cluster contains four groups (group 4, 5, 6 and 7), and the effective market contains groups 4 and 5, thus $K^j = 5$. The service provider obtains the following maximum revenue obtained from cluster $\mathcal{C}^j$:

$$R^j(s^j, \boldsymbol{a}) = \frac{s^j \sum_{i\in C^j,\, i\le K^j} N_i\theta_i}{s^j + \sum_{i\in C^j,\, i\le K^j} N_i}. \tag{4.46}$$

Level-2: Resource allocation among clusters

For a fixed partition $\boldsymbol{a}$, we then consider the resource allocation among clusters.

$$\begin{aligned} \text{Level-2:} \quad \underset{s^j\ge 0}{\text{maximize}} \quad & \sum_{j\in\mathcal{J}} R^j(s^j, \boldsymbol{a}) \\ \text{subject to} \quad & \sum_{j\in\mathcal{J}} s^j \le S \end{aligned}$$

We will show in Section 4.4.4 that subproblems in Level-2 and Level-3 can be transformed into a complete price differentiation problem under proper technique conditions. Let us denote the its optimal value as $R_{pp}(\boldsymbol{a})$.

[4] Note that we do not assume that the effective market threshold equals to the number of effective groups, e.g., there are 2 effective groups in Fig. 5, but threshold $K^j = 5$. Later we will prove that there is unified threshold for PP Problem. Then by this result, the group index threshold actually coincides with the number of effective groups.

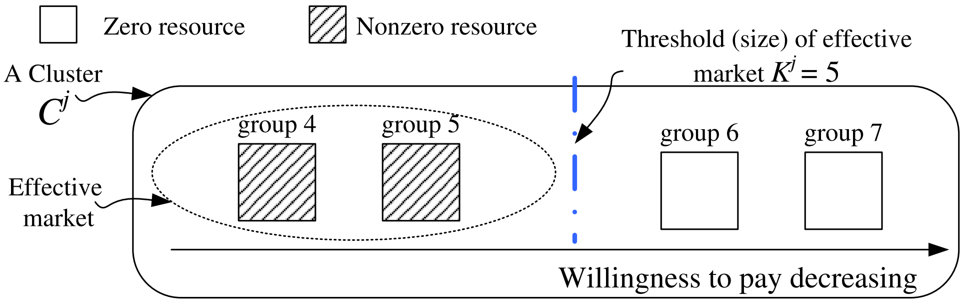

Figure 4.7: An illustrative example: the cluster contains four groups, group 4, 5, 6 and 7; and the effective market contains group 4 and 5, thus $K^j = 5$

Level-1: cluster partition

Finally, we solve the cluster partition problem.

$$\begin{aligned} \text{Level-1:} \quad & \underset{a_i^j \in \{0,1\}}{\text{maximize}} \quad R_{pp}(\boldsymbol{a}) \\ & \text{subject to} \quad \sum_{j \in \mathcal{J}} a_i^j = 1, \ i \in \mathcal{I}. \end{aligned}$$

This partition problem is a combinatorial optimization problem. The size of its feasible set is $S(I,J) = \frac{1}{J!} \sum_{t=1}^{J} (-1)^{J+t} C(J,t) t^I$, *Stirling number of the second kind* (46, Chap.13), where $C(J,t)$ is the binomial coefficient. Some numerical examples are given in the third row in Table 4.2. If the number of prices J is given, the feasible set size is exponential in the total number of groups I. For our problem, however, it is possible to reduce the size of the feasible set by exploiting the special problem structure. More specifically, the group indices in each cluster should be consecutive at the optimum. This means that the size of the feasible set is $C(I-1, J-1)$ as shown in the last row in Table 4.2, and thus is much smaller than $S(I,J)$.

Table 4.2: Numerical examples for feasible set size of the partition problem in Level-1

Number of groups	$I = 10$		$I = 100$		$I = 1000$
Number of prices	$J = 2$	$J = 3$	$J = 2$	$J = 3$	$J = 2$
$S(I,J)$	511	9330	6.33825×10^{29}	8.58963×10^{46}	5.35754×10^{300}
$C(I-1, J-1)$	9	36	99	4851	999

Next we discuss how to solve the three level subproblems. A route map for the whole solving process is given in Fig. 4.8.

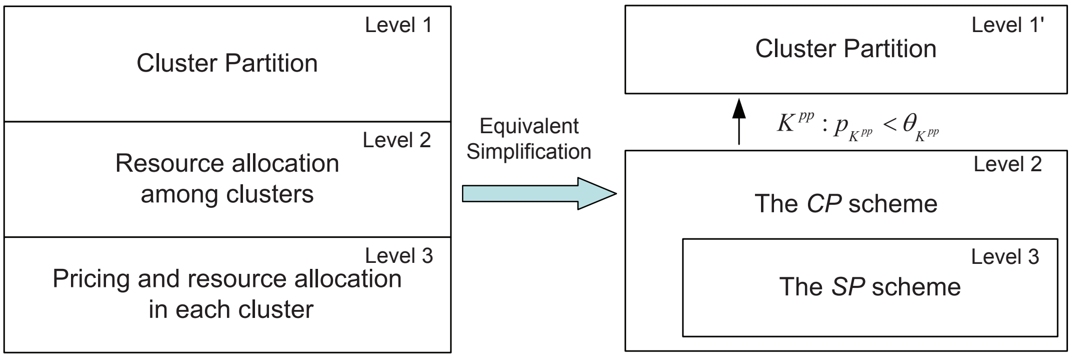
Level 1
Cluster Partition
Level 2
Resource allocation among clusters
Level 3
Pricing and resource allocation in each cluster
Equivalent Simplification
Level 1'
Cluster Partition
$K^{pp}: p_{K^{pp}} < \theta_{K^{pp}}$
Level 2
The *CP* scheme
Level 3
The *SP* scheme

Figure 4.8: Decomposition and simplification of the general *PP* Problem: The three-level decomposition structure of *PP* Problem is shown in the left hand side. After simplifications in Section 4.4.4 and 4.4.4, the problem will be reduced to structure in right hand side.

Solving Level-2 and Level-3

The optimal solution (4.46) of the Level-3 problem can be equivalently written as

$$R^j(\boldsymbol{s}, \boldsymbol{a}) = \frac{s^j \sum_{i \in C^j,\, i \le K^j} N_i \theta_i}{s^j + \sum_{i \in C^j,\, i \le K^j} N_i} \overset{(a)}{=} \frac{s^j N^j \theta^j}{s^j + N^j}, \tag{4.47}$$

$$\text{where} \begin{cases} N^j & = \sum_{i \in C^j,\, i \le K^j} N_i, \\ \theta^j & = \sum_{i \in C^j,\, i \le K^j} \frac{N_i \theta_i}{N^j}. \end{cases} \tag{4.48}$$

The equality (a) in (4.47) means that each cluster $\mathcal{C}^j$ can be equivalently treated as a group with N^j homogeneous users with the same willings to pay θ^j. We name this equivalent group as a *super-group* (SG). We summarize the above result as the following lemma.

Lemma 4.13 *For every cluster C^j and total resource s^j, $j \in \mathcal{J}$, we can find an equivalent super-group which satisfies conditions in (4.48) and achieves the same revenue under the SP scheme.*

Based on Lemma 4.13, Level-2 and level-3 subproblems together can be viewed as *CP* Problem for super-groups. Since a cluster and its super-group from a one-to-one mapping, we will use the two words interchangeably in the sequel.

However, simply combining Theorems 4.11 and 4.12 to solve Level-2 and Level-3 Subproblems for a fixed partition $\boldsymbol{a}$ can result in a very high complexity. This is because the effective markets within each super-group and between super-groups are coupled together. An illustrative example of this coupling effective market is shown in Fig. 4.9, where K^c is the threshold between clusters and has three possible positions (i.e., between group 2 and group 3, between group 5 and group 6, or after group 6); and K_1 and K_2 are thresholds within cluster $\mathcal{C}^1$ and $\mathcal{C}^2$, which have two or three possible positions, respectively. Thus, there are $(2 \times 3) \times 3 = 18$ possible thresholds possibilities in total.

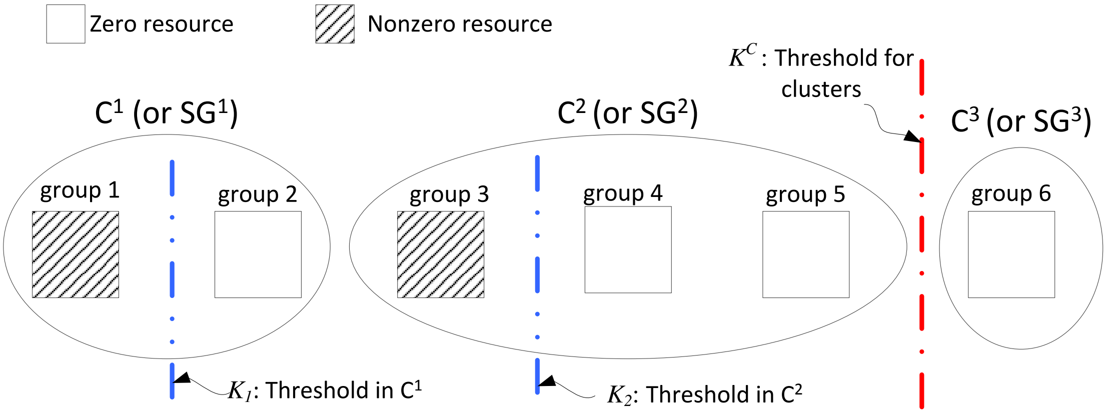

Figure 4.9: An example of coupling thresholds.

The key idea to resolve this coupling issue is to show that the situation in Fig. 4.9 can not be an optimal solution of PP Problem. The results in Sections 4.4.2 and 4.4.3 show that there is a unified threshold at the optimum in both the CP and SP cases, e.g., Fig. 4.6. Next we will show that a unified single threshold also exists in the PP case.

Lemma 4.14 *At any optimal solution of the PP scheme, the group indices of the effective market is consecutive.*

The intuition is that the resource should be always allocated to high willingness to pay users at the optimum. Thus, it is not possible to have Fig. 4.9 at an optimal solution, where high willingness to pay users in group 2 are allocated zero resource while low willingness to pay users in group 3 are allocated positive resources.

Based on Lemma 4.14, we know that there is a unified effective market threshold for PP Problem, denoted as K^{pp}. Since all groups with indices larger than K^{pp} make zero contribution to the revenue, we can ignore them and only consider the partition problem for the first K^{pp} groups. Given a partition that divides the K^{pp} groups into J clusters (super-groups), we can apply the CP result in Section 4.4.2 to compute the optimal revenue in

Level-2 based on Theorem 4.11.

$$
\begin{aligned}
R_{pp}(\boldsymbol{a}) &= \sum_{j=1}^{J} N^j \theta^j - \frac{\left(\sum_{j=1}^{J} N^j \sqrt{\theta^j}\right)^2}{S + \sum_{j=1}^{J} N^j} \\
&= \sum_{i=1}^{K^{pp}} N_i \theta_i - \frac{\left(\sum_{j=1}^{J} N^j \sqrt{\theta^j}\right)^2}{S + \sum_{i=1}^{K^{pp}} N_i}. \qquad (4.49)
\end{aligned}
$$

Solving Level-1

We first consider solving Level-1 with a given effective market threshold K^{pp}. Based on the previous results, we first simplify Level-1 Subproblem, and prove the theorem below.

Theorem 4.15 *For a given threshold K^{pp}, the optimal partition of Level-1 Subproblem is the solution of the following optimization problem.*

$$
\begin{aligned}
\textit{Level-1}' \quad & \underset{a_i^j, N^j, \theta^j}{\textit{minimize}} && \sum_{j \in \mathcal{J}} N^j \sqrt{\theta^j} \\
& \textit{subject to} && N^j = \sum_{i \in \mathcal{K}^{pp}} N_i a_i^j, \;\; j \in \mathcal{J}, \\
& && \theta^j = \sum_{i \in \mathcal{K}^{pp}} \frac{N_i a_i^j}{N^j} \theta_i \;\; j \in \mathcal{J}, \\
& && \sum_{j \in \mathcal{J}} a_i^j = 1, \;\; a_i^j \in \{0, 1\}, i \in \mathcal{K}^{pp} \; j \in \mathcal{J}, \\
& && \theta_{K^{pp}} > p^J = \sqrt{\theta^J(\boldsymbol{a})\lambda(\boldsymbol{a})}. \qquad (4.50)
\end{aligned}
$$

where $\mathcal{K}^{pp} \triangleq \{1, 2, \ldots, K^{pp}\}$, $\theta^J(\boldsymbol{a})$ is the value of average willingness to pay of the Jth group for the partition $\boldsymbol{a}$, and $\lambda(\boldsymbol{a}) = \left(\frac{\sum_{j \in \mathcal{J}} N^j \sqrt{\theta^j}}{S + \sum_{i=1}^{K^{pp}} N_i}\right)^2$.

Level-1′ Problem is still a combinatorial optimization problem with a large feasible set of $\boldsymbol{a}$ (similar as the original Level-1). The following result can help us to reduce the size of the feasible set.

Theorem 4.16 *For any effective market size K^{pp} and number of prices J, an optimal partition of PP Problem involves consecutive group indices within clusters.*

The intuition is that high willingness to pay users should be allocated positive resources with priority. It implies that groups with similar willingness to pays should be partitioned in the same cluster, instead of in several far away clusters. Or equivalently, the group indices within each cluster should be consecutive.

We define $\mathcal{A}$ as the set of all partitions with consecutive group indices within each cluster, and $v(\boldsymbol{a}) = \sum_{j\in\mathcal{J}} N^j\sqrt{\theta^j}$ is the value of objective of Level-1$'$ Problem for a partition $\boldsymbol{a}$. Algorithm 4 finds the optimal solution of Level-1$'$. The main idea for this algorithm is to enumerate every possible partition in set $\mathcal{A}$, and then check whether the threshold condition (4.50) can be satisfied. The main part of this algorithm is to enumerate all partitions in set $\mathcal{A}$ of $C(K^{pp}-1, J-1)$ feasible partitions. Thus the complexity of Algorithm 4 is no more than $\mathcal{O}((K^{pp})^{J-1})$.

Algorithm 4 Solve the Level-1$'$ Problem with fixed K^{pp}

```
function Level-1($K^{pp}$, $J$)
    $k \leftarrow K^{pp}$
    $v^* \leftarrow \sqrt{\sum_{i=1}^{k} N_i\theta_i}$, $\boldsymbol{a}^* = \mathbf{0}$
    for $\boldsymbol{a} \in \mathcal{A}$ do
        if $\theta_k > \sqrt{\theta^J(\boldsymbol{a})\lambda(\boldsymbol{a})}$ then
            if $v(\boldsymbol{a}) < v^*$ then
                $v^* \leftarrow v(\boldsymbol{a})$, $\boldsymbol{a}^* \leftarrow \boldsymbol{a}$
            end if
        end if
    end for
    return $\boldsymbol{a}^*$
end function
```

Now we search the optimal effective market threshold K^{pp}. We know the optimal market threshold K^{pp} is upper-bounded, i.e., $K^{pp} \leq K^{cp} \leq I$. Thus we can first run Algorithm 2 to calculate the effective market size for the CP scheme K^{cp}. Then, we search the optimal K^{pp} iteratively using Algorithm 4 as an inner loop. We start by letting $K^{pp} = K^{cp}$ and run Algorithm 4. If there is no solution, we decrease K^{pp} by one and run Algorithm 4 again. The algorithm will terminate once we find an effective market threshold where Algorithm 4 has an optimal solution. Once the optimal threshold and the partition of the clusters are determined, we can further run Algorithm 2 to solve the joint optimal resource allocation and pricing scheme. The pseudo code is given in Algorithm 5 as follows.

In Algorithm 5, it invokes two functions: CP($\{N_i\theta_i\}_{i\in\mathcal{I}}$, S) as described in Algorithm 2 and and Level-1(k, J) as in Algorithm 4. CP($\{N_i\theta_i\}_{i\in\mathcal{I}}$, S) returns a vector with two elements: CP($\{N_i\{\theta_i\}_{i\in\mathcal{I}}$, S)_1 denotes the first element K^{cp}, and CP($\{N_i\theta_i\}_{i\in\mathcal{I}}$, S)_2 denotes the second element λ^* in CP Problem.

The above analysis leads to the following theorem:

Theorem 4.17 *The solution obtained by Algorithm 5 is optimal for PP Problem.*

Algorithm 5 Solve Partial Price Differentiation Problem

```
$p_i \leftarrow \theta_i$
$k \Leftarrow \mathrm{CP}(\{N_i, \theta_i\}_{i\in\mathcal{I}},\ S)\_1$, $\boldsymbol{a}^* \Leftarrow \text{Level-1}(k,J)$
while $\boldsymbol{a}^* == \mathbf{0}$ do
    $k \leftarrow k-1$, $\boldsymbol{a}^* \Leftarrow \text{Level-1}(k,J)$
end while
for $j \leftarrow 1, J$ do
    $N^j \leftarrow \sum_{i=1}^{k} N_i a_i^j$, $\theta^j \leftarrow \sum_{i=1}^{k} \frac{N_i a_i^j}{N^j}\theta_i$
end for
$\lambda \Leftarrow \mathrm{CP}(\{N^j, \theta^j\}_{i\in\mathcal{J}},\ S)\_2$
for $i \leftarrow 1, k$ do
    $p_i \leftarrow \sum_{j=1}^{J} a_i^j \sqrt{\theta^j \lambda}$
end for
return $\{p_i\}_{i\in\mathcal{I}}$
```

4.4.5 PRICE DIFFERENTIATION UNDER INCOMPLETE INFORMATION

In Sections 4.4.2, 4.4.3, and 4.4.4, we discuss various pricing schemes with different implementational complexity level under complete information, the revenues of which can be viewed as the benchmark of practical pricing designs. In this section, we further study the incomplete information scenario, where the service provider does not know the group association of each user. The challenge for pricing in this case is that the service provider needs to provide the right incentive so that a group i user does not want to pretend to be a user in a different group. It is clear that the CP scheme in Section 4.4.2 and the PP scheme in Section 4.4.4 cannot be directly applied here. The SP scheme in Section 4.4.3 is a special case, since it does not require the user-group association information in the first place and thus can be applied in the incomplete information scenario directly. On the other hand, we know that the SP scheme may suffer a considerable revenue loss compared with the CP scheme. Thus it is natural to ask whether it is possible to design an incentive compatible differentiation scheme under incomplete information. In this section, we design a quantity-based price menu to incentivize the users to make the right self-selection and achieve the same maximum revenue of the CP scheme under complete information under proper technical conditions. We name it as the Incentive Compatible Complete Price differentiation ($ICCP$) scheme.

In the $ICCP$ scheme, the service provider publishes the quantity-based price menu, which consists of several step functions of resource quantities. Users are allowed to freely choose their quantities. The aim of this price menu is to make the users *self-differentiated*, so that to mimic the same result (the same prices and resource allocations) of the CP scheme under complete information. Based on Theorem 4.11, there are only K (without confusion, we remove the superscript "cp" to simplify the notation) effective groups of users receiving non-zero resource allocations, thus there are K steps of unit prices, $p_1^* > p_2^* > \cdots > p_K^*$ in the price menu. These prices are exactly the same optimal prices that the service provider would charge for K effective groups as in Theorem 4.11. Note that for the $K+1, \ldots, I$ groups, all the prices in the menu are too high for them, then they will still demand zero resource. The quantity is divided into K intervals by $K-1$ thresholds, $s_{th}^1 > s_{th}^2 > \cdots > s_{th}^{K-1}$. The $ICCP$ scheme can specified as follows:

$$p(s) = \begin{cases} p_1^* & \text{when } s > s_{th}^1 \\ p_2^* & \text{when } s_{th}^1 \geq s > s_{th}^2 \\ \vdots & \\ p_K^* & \text{when } s_{th}^{K-1} \geq s > 0. \end{cases} \tag{4.51}$$

A four-group example is shown in Fig. 4.10.

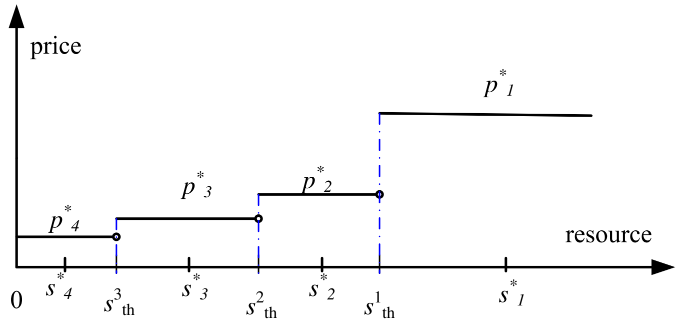

Figure 4.10: A four-group example of the $ICCP$ scheme: where the prices $p_1^* > p_2^* > p_3^* > p_4^*$ are the same as the CP scheme. To mimic the same resource allocation as under the CP scheme, one necessary (but not sufficient) condition is $s_{th}^{j-1} \geq s_j^*$ for all j, where s_j^* is the optimal resource allocation of the CP scheme.

Note that in contrast to the usual "volume discount", here the price is non-decreasing in quantity. This is motivated by the resource allocation in Theorem 4.11, that a user with a higher θ_i is charged a higher price for a larger resource allocation. Thus the observable quantity can be viewed as an indication of the unobservable users' willingness to pay, and help to realize price differentiation under incomplete information.

The key challenge in the $ICCP$ scheme is to properly set the quantity thresholds so that users are perfectly segmented through self-differentiation. This is, however, not always possible. Next we derive the necessary and sufficient conditions to guarantee the perfect segmentation.

Let us first study the self-selection problem between two groups: group i and group q with $i<q$. Later on we will generalize the results to multiple groups. Here group i has a higher willingness to pay, but will be charged with a higher price p_i^* in the CP case. The incentive compatible constraint is that a high willingness to pay user can not get more surplus by pretending to be a low willingness to pay user, i.e., $\max_s U_i(s;p_i^*) \geq \max_s U_i(s;p_q^*)$, where $U_i(s;p)=\theta_i\ln(1+s)-ps$ is the surplus of a group i user when it is charged with price p.

Without confusion, we still use s_i^* to denote the optimal resource allocation under the optimal prices in Theorem 4.11, i.e., $s_i^*=\arg\max_{s_i\geq 0} U_i(s_i;p_i^*)$. We define $s_{i\to q}$ as the quantity satisfying

$$\begin{cases} U_i(s_{i\to q};p_q^*)=U_i(s_i^*;p_i^*) \\ s_{i\to q}<s_i^* \end{cases}. \tag{4.52}$$

In other words, when a group i user is charged with a lower price p_q^* and demands resource quantity at $s_{i\to q}$, it achieves the same as the maximum surplus under the optimal price of the CP scheme p_i^*, as showed in Fig. 4.11. Since the there two solutions of the first equation of (4.52), we constraint $s_{i\to q}$ to be the one that is smaller than s_i^*.

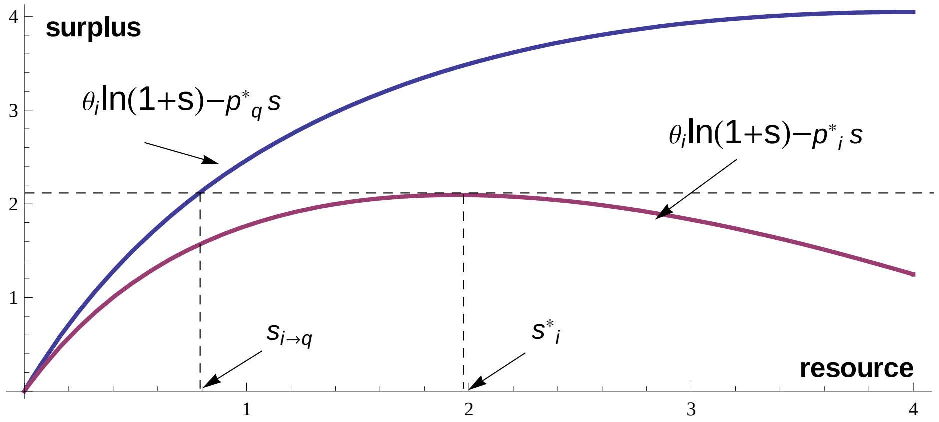

Figure 4.11: When the threshold $s_{th}^{q-1}<s_{i\to q}$, the group i user can not obtain $U(s_i^*,p_i^*)$ if it chooses the lower price p_q at a quantity less than s_{th}^{q-1}. Therefore it will automatically choose the high price p_i^* to maximize its surplus.

To maintain the group i users' incentive to choose the higher price p_i^* instead of p_q^*, we must have $s_{th}^{q-1}\leq s_{i\to q}$, which means a group i user can not obtain $U_i(s_i^*,p_i^*)$ if it chooses a quantity less than s_{th}^{q-1}. In other words, it will automatically choose the higher (and the desirable) price p_i^* to maximize its surplus. On the other hand, we must have $s_{th}^{q-1}\geq s_q^*$ in order to maintain the optimal resource allocation and allow a group q user to choose the right quantity-price combination (illustrated in Fig. 4.10).

Therefore, it is clear that the *necessary and sufficient* condition that the *ICCP* scheme under incomplete information achieves the same maximum revenue of the *CP* scheme under complete information is

$$s_q^* \leq s_{i \to q}, \ \forall \, i < q, \forall \, q \in \{2, \ldots, K\}. \tag{4.53}$$

By solving these inequalities, we can obtain the following theorem.

Theorem 4.18 *There exist unique thresholds* $\{t_1, \ldots, t_{K-1}\}$*, such that the ICCP scheme achieves the same maximum revenue as in the complete information case if*

$$\sqrt{\frac{\theta_q}{\theta_{q+1}}} \geq t_q \quad for \, q = 1, \ldots, K-1.$$

Moreover, t_q *is the unique solution of the equation*

$$t^2 \ln t - (t^2 - 1) + \frac{t \sum_{k=1}^{q} N_k + N_{q+1}}{S + \sum_{k=1}^{K^{cp}} N_k}(t-1) = 0$$

over the domain $t > 1$*.*

We want to mention that the condition in Theorem 4.18 is necessary and sufficient for the case of $K = 2$ effective groups[5]. For $K > 2$, Theorem 4.18 is sufficient but not necessary. The intuition of Theorem 4.18 is that users need to be sufficiently different to achieve the maximum revenue.

The following result immediately follows Theorem 4.18.

Corollary 4.19 *The* t_q*s in Theorem 4.18 satisfy* $t_q < t_{root}$ *for* $q = 1, \ldots, K-1$*, where* $t_{root} \approx 2.21846$ *is the larger root of equation* $t^2 \ln t - (t^2 - 1) = 0$*.*

The Corollary 4.19 means that the users do not need to be extremely different to achieve the maximum revenue.

When the conditions in Theorem 4.18 are not satisfied, there may be revenue loss by using the pricing menu in (4.51). Since it is difficult to explicitly solve the parameterized transcend equation (4.52), we are not able to characterize the loss in a closed form yet.

[5] There might be other groups who are not allocated positive resource under the optimal pricing.

4.4.6 NUMERICAL RESULTS

We provide a numerical example to quantitatively study two key questions regarding the performance comparison of different algorithms:

- When is price differentiation most beneficial?
- What is the best tradeoff of partial price differentiation?

Definition 4.20 (**Revenue gain**) We define the revenue gain G of one pricing scheme as the ratio of the revenue difference (between this pricing scheme and the single pricing scheme) normalized by the revenue of single pricing scheme.

We consider a three-group example and three different sets of parameters as shown in Table 4.3. To limit the dimension of the problem, we set the parameters such that the total number of users and the average willingness to pay (i.e., $\bar{\theta} = \sum_{i=1}^{3} N_i\theta_i/(\sum_{i=1}^{3} N_i)$) of all users are the same across three different parameter settings. This ensures that the SP scheme achieves the same revenue in three different cases when resource is abundant. Figure 4.12 illustrates how the differentiation gain changing changes in resource S.

Table 4.3: Parameter settings of a three-group examples

	θ_1	N_1	θ_2	N_2	θ_3	N_3	$\bar{\theta}$
Case 1	9	10	3	10	1	80	2
Case 2	3	33	2	33	1	34	2
Case 3	2.2	80	1.5	10	1	10	2

Fig. 4.12 shows that the revenue gain is large only when the high willingness to pay users are minorities (e.g. case 1) in the effective market and the resource is limited but not too small ($100 \leq S \leq 150$ in all three cases). When resource S is large enough (e.g., ≥ 150), the gain will gradually diminish to zero as the resource increases. For each curve in Fig. 4.12, there are two peak points. Each peak point represents a change of the effective market threshold in the SP scheme, i.e., when the resource allocation to a group becomes zero.

Now let us consider a five-group example with parameters shown in Table 4.4 to illustrate the tradeoff of partial price differentiation. Note that high willingness to pay users are minorities here. Figure 4.13 shows the revenue gain G as a function of total resource S under different PP schemes (including CP scheme as a special case).

We enlarge Fig. 4.13 within the range of $S \in [0, 50]$, which is the most complex and interesting part due to several peak points. Similar as Fig. 4.12, we observe $I - 1 = 4$ peak

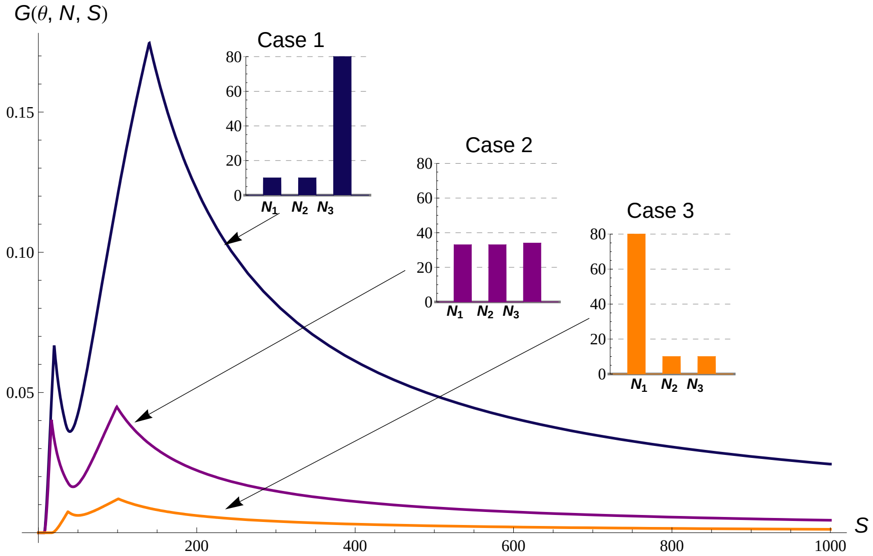

Figure 4.12: An example of the revenue gain of the three-group market with the same average willingness to pay

Table 4.4: Parameter setting of a five-group example

group index i	1	2	3	4	5
θ_i	16	8	4	2	1
N_i	2	3	5	10	80

points for each curve in Fig. 4.13. Each peak point again represents a change of effective market threshold of the single pricing scheme.

As the resource S increases from 0, all gains in Fig. 4.13 first overlap with each other, then the two-price scheme (blue curve) separates from the others at $S = 3.41$, after that the three-price scheme (purple curve) separates at $S = 8.89$, and finally the four-price scheme (dark yellow curve) separates at near $S = 20.84$. These phenomena are due to the threshold structure of the PP scheme. When the resource is very limited, the effective markets under all pricing scheme include only one group with the highest willingness to pay, and all pricing schemes coincide with the SP scheme. As the resource increases, the effective market enlarges from two groups to finally five groups.

Figure 4.13 provides the service provider a global picture of choosing the most proper pricing scheme according to achieve the desirable financial target under a certain parameter setting. For example, if the total resource $S = 100$, the two-price scheme seems to be a sweet

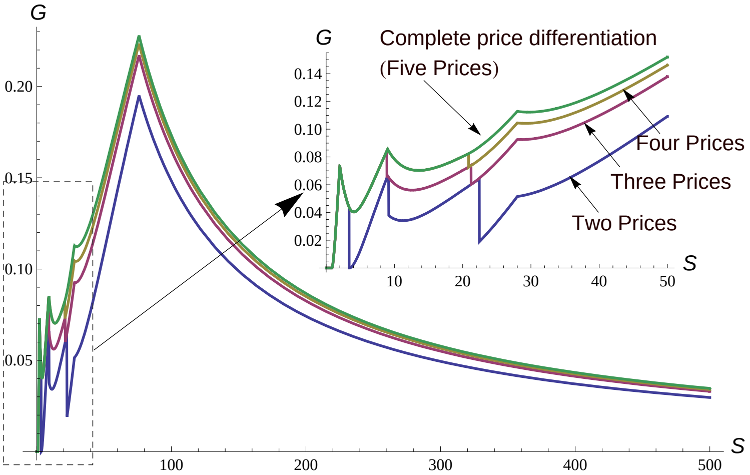

Figure 4.13: Revenue gain of a five-group example under different price differentiation schemes

spot, as it achieves a differential gain of 14.8% comparing to the SP scheme and is only 2.4% worse than the CP scheme with five prices.

4.5 CHAPTER SUMMARY

This chapter discusses how a monopoly maximizes its profit through proper pricing mechanisms.

We start by discussing how a monopoly should choose a single profit maximizing price based on the demand elasticity. The key result is that a monopoly will always operate on the elastic portion of the demand curve. Then we look at the monopolist's options when it can charge different prices to the same customer or different customer groups. This leads to three types of price discrimination schemes. In the first-degree price discrimination, the monopoly knows the complete demand information of all customers and performs perfect price discrimination. This unrealistic case provides a theoretical benchmark for other schemes. In the second-degree price discrimination, the monopolist offers a bundle of prices corresponding to different demand quantities, and let the customers choose their best choices. This is often applied when the monopolist knows only limited information of the consumers' demands. In the last third-degree price discrimination, which is one of the most commonly used ones, the monopolist segments the market into several groups, and charges different prices for different groups. This applies for the case that the monopoly knows the total demand for each group, but not the individual demand information. Both second and third degree price discrimination induce profit loss comparing to the first degree, and this is inevitable due to the lack of proper information. For more details about the theory, please see (9; 39).

We then introduce two networking examples to illustrate the ideas. We first consider a revenue maximization problem in cognitive radio networks, where a primary user wants to control the available bandwidth and maximum interference temperature for secondary users to maximize the revenue. We define the concept of power-interference elasticity, which illustrates how the power demand changes with the interference level in the system. We show that such an elasticity determines how the primary user should maximize the revenue. In the second example, we study the revenue-maximizing problem for a monopoly service provider under both complete and incomplete network information. Under complete information, our focus is to investigate the tradeoff between the total revenue and the implementational complexity (measured in the number of pricing choices available for users). Among the three pricing differentiation schemes we proposed (*i.e.,* complete, single, and partial), the partial price differentiation is the most general one and includes the other two as special cases. By exploiting the unique problem structure, we designed an algorithm that computes the optimal partial pricing scheme in polynomial time, and numerically quantifies the tradeoff between implementational complexity and total revenue. Under incomplete information, designing an incentive-compatible differentiation pricing scheme is difficult in general. We show that when the users are significantly different, it is possible to design a quantity-based pricing scheme that achieves the same maximum revenue as under complete information. For more details about these two examples, please see (47; 48).

CHAPTER 5

Oligopoly Pricing

In Chapter 4, we considered how a single decision maker, the monopolist, chooses the price(s) to maximize the profit. In this chapter, we consider a more complicated and yet more common situation, where many self-interested individuals (including firms, consumers, etc.) make *interdependent* interactions, that is, the payoff of each individual depends not only on his own choices, but also on the choices of other individuals. The well-known theoretic tool used for analyzing such economic settings is *game theory*. After introducing the basic concepts of game theory following (49; 50; 51; 52), we will look at multiple classical market competition models, including Cournot competition based on output quantities, Bertrand competition based on pricing, and Hoteling model that captures the location information in the competition.

In terms of applications, we first revisit the multiple base station model discussed in Chapter 3. We will study how multiple base stations compete in the market by pricing their resources to attract customers and maximize their own revenues. In the second example, we examine how two secondary wireless service providers compete by leasing resources from spectrum owners and provide services to the same group of customers.

5.1 THEORY: GAME THEORY

5.1.1 WHAT IS A GAME?

A game is a formal representation of a situation in which a number of individuals interact in a setting of *strategic interdependence*. By that, we mean that each individual's welfare depends not only on his own choices but also on the choices of other individuals. To describe a situation of strategic interaction, we need to define four things:

- **Players**: Who is involved in a game?
- **Rules**: Who moves when? What do they know when they move? What can they do (*i.e.,* what actions can they select)?
- **Outcomes**: What is the outcome of the game, for each possible actions chosen by players?
- **Payoffs**: What are the players' preferences (*i.e.,* utility) over the possible outcomes?

We assume each player is *rational* or *self-interested*, whose goal is to choose the actions that produce his most preferred outcomes.[1] When facing potential uncertainty over multiple outcomes, a rational player chooses actions that maximize his expected utilities. Under such a situation, a central problem is to identify the potential outcomes of the game, which we call *equilibrium*.

5.1.2 STRATEGIC FORM GAME

We first introduce *strategic form games* (also referred to as the normal form games). A strategic form game is a model for a game in which all players act simultaneously, without knowledge of other players actions. Thus, we only need to define the player set, the action set for each player, and the payoff (utility) function for each player. Formally,

Definition 5.1 **Strategic Form Game.** A strategic form game is a triplet $\langle \mathcal{I}$, $(\mathcal{S}_i)_{i\in\mathcal{I}}$, $(u_i)_{i\in\mathcal{I}}\rangle$ where

1. $\mathcal{I} = \{1, 2, ..., I\}$ is a finite set of game players.
2. $\mathcal{S}_i$ is a set of available actions (pure strategies) for player i. We further denote by $s_i \in \mathcal{S}_i$ an action for player i, and by $\boldsymbol{s}_{-i} = (s_j)_{j\neq i}$ a vector of actions for all players except i. The tuple $\boldsymbol{s} = (s_i, \boldsymbol{s}_{-i}) \in \mathbb{S}$ an action profile, where $\mathbb{S} \triangleq \Pi_i \mathcal{S}_i$ is the set of all possible action profiles. We further denote by $\mathbb{S}_{-i} \triangleq \Pi_{j\neq i}\mathcal{S}_j$ is the set of action profiles for all players except i.
3. $u_i : \mathbb{S} \to \mathbb{R}$ is the payoff (utility) function of player i, which maps every possible action profile in $\mathbb{S}$ to a real number, the utility.

Two important concepts in game theory are *dominated strategy* and *best response strategy* (or best response correspondence). We say a strategy (action) is *strictly dominated* for a player, if there exists some other strategy that always gives a better outcome for the player, no matter what strategies the other players choose.

Definition 5.2 **Strictly Dominated Strategy.** A strategy $s_i \in \mathcal{S}_i$ is strictly dominated for player i, if there exists some $s_i' \in \mathcal{S}_i$ such that

$$u_i(s_i, \boldsymbol{s}_{-i}) < u_i(s_i', \boldsymbol{s}_{-i}), \quad \forall \boldsymbol{s}_{-i} \in \mathbb{S}_{-i}.$$

We can also define a weaker version of dominated strategies.

[1] We assume that players' preference orderings are complete and transitive, like in most of the game theory literature (49; 50).

Definition 5.3 Weakly Dominated Strategy. A strategy $s_i \in \mathcal{S}_i$ is weakly dominated for player i, if there exists some $s_i' \in \mathcal{S}_i$ such that

$$u_i(s_i, \boldsymbol{s}_{-i}) \leq u_i(s_i', \boldsymbol{s}_{-i}), \quad \forall \boldsymbol{s}_{-i} \in \mathbb{S}_{-i},$$

with at least one inequality holds strictly.

We say a strategy is a *best response correspondence* (for a player i) to a particular strategy profile $\boldsymbol{s}_{-i}$ of other players, if it yields the best outcome for player i under $\boldsymbol{s}_{-i}$.

Definition 5.4 Best Response Correspondence. For each player i, the best response correspondence $B_i(\boldsymbol{s}_{-i}) : \mathbb{S}_{-i} \to \mathcal{S}_i$ is a mapping from the set $\mathbb{S}_{-i}$ into $\mathcal{S}_i$ such that

$$B_i(\boldsymbol{s}_{-i}) = \{s_i \in \mathcal{S}_i \mid u_i(s_i, \boldsymbol{s}_{-i}) \geq u_i(s_i', \boldsymbol{s}_{-i}), \forall s_i' \in \mathcal{S}_i\}.$$

When the strategy space is finite, and when the number of players and actions is small, it is convenient to represent a strategic form game in matrix form. To show this, consider the following "Matching Pennies" game and "Prisoner's Dilemma" game.

Matching Pennies Game

Two players turn the penny to "HEADS" or "TAILS" secretly and simultaneously. If the pennies match (both heads or both tails), Player 1 keeps both pennies, so wins one from Player 2 ($u_1 = 1$ for player 1, $u_2 = -1$ for player 2). If the pennies do not match (one heads and one tails), Player 2 keeps both pennies, so receives one from Player 1 ($u_1 = -1$ for player 1, $u_2 = 1$ for player 2). This is an example of a *zero-sum game*, where one player's gain is exactly the other player's loss. This game can be represented by the following matrix, where each row denotes the action of player 1, each column denotes the action of player 2, and the cell indexed by row x and column y contains a utility pair (a, b) with $a = u_1(x, y)$ and $b = u_2(x, y)$ for both players.

	HEADS	TAILS
HEADS	$(1, -1)$	$(-1, 1)$
TAILS	$(-1, 1)$	$(1, -1)$

In this game, player 1's best response correspondence is "HEADS" if player 2 selects the strategy "HEADS", and "TAILS" otherwise. We can also see that no strategy is dominated. To see this, we consider the strategy "HEADS" for player 1. It yields a better outcome for player 1 if player 2 selects the strategy "HEADS", whereas a worse outcome for player 1 if player 2 selects the strategy "TAILS". Similar result holds for the strategy "TAILS" as the game is symmetric.

Prisoner's Dilemma Game

Two players are arrested for a crime and placed in separate rooms. The authorities try to extract a confession from them. If they both remain silent, then the authorities will not be able to prove charges against them and they will both serve a short prison term, say 2 years ($u_i = -2$ for both players $i = 1, 2$), for minor offenses. If only one of them (say, player 1) confesses, his term will be reduced to 1 year ($u_1 = -1$ for player 1) and he will be used as a witness against the other person, who will get a sentence of 5 years ($u_2 = -5$ for player 2). If they both confess, they both get a smaller charge of 4 years ($u_i = -4$ for both players $i = 1, 2$) comparing with the worst case of 5 years. This game can be represented in matrix form as follows.

	SILENT	CONFESS
SILENT	$(-2, -2)$	$(-5, -1)$
CONFESS	$(-1, -5)$	$(-\mathbf{4}, -\mathbf{4})$

In this game, the player 1's best response correspondence is always "CONFESS", no matter what strategy player 2 chooses. Therefore the strategy "SILENT" is dominated by "CONFESS". To see this, we consider the strategy "CONFESS" for player 1. If player 2 selects the strategy "SILENT", it yields a better outcome ($u_1 = -1$) than "SILENT" ($u_1 = -2$) for player 1. If player 2 selects the strategy "CONFESS", it still yields a better outcome ($u_1 = -4$) than "SILENT" ($u_1 = -5$) for player 1.

It is also possible for the strategy space of a player (or even the number of players) to be infinite. For example, see the Cournot mode in Section 5.2.1.

5.1.3 NASH EQUILIBRIUM

Now we consider what outcome would result from the players' strategic interactions, given a particular game including the player set, acting rules, and payoff functions. This actually leads to the most important concept in game theory—*Nash Equilibrium.* At a high-level, a Nash Equilibrium is a profile of strategies, which has the property that no *single* player can improve his utility by deviating from the action profile, assuming that all other players act according to it (49; 50). More formally,

Definition 5.5 Nash Equilibrium. A pure strategy Nash Equilibrium of a strategic form game $\langle \mathcal{I}, (\mathcal{S}_i)_{i\in\mathcal{I}}, (u_i)_{i\in\mathcal{I}}\rangle$ is a strategy profile $\boldsymbol{s}^* \in \mathbb{S}$ such that for all $i \in \mathcal{I}$ the following condition holds

$$u_i(s_i^*, \boldsymbol{s}_{-i}^*) \geq u_i(s_i', \boldsymbol{s}_{-i}^*), \quad \forall s_i' \in \mathcal{S}_i.$$

The above definition can be restated in terms of a best-response correspondence:

Definition 5.6 Nash Equilibrium-Restated. A strategy profile $\boldsymbol{s}^* \in \mathbb{S}$ is a Nash Equilibrium of a strategic form game $\langle \mathcal{I}, (\mathcal{S}_i)_{i \in \mathcal{I}}, (u_i)_{i \in \mathcal{I}} \rangle$ if and only if

$$s_i^* = B_i(\boldsymbol{s}_{-i}^*), \quad \forall i \in \mathcal{I},$$

where $B_i(\cdot)$ is the best response correspondence defined in Definition 5.4.

It is worth noting that *not* every game produces a pure strategy Nash Equilibrium. To see this, recall the Matching Pennies Game mentioned above. There is actually no pure strategy Nash Equilibrium for the Matching Pennies Game, since for any strategy profile there is always a player who can increase his utility by unilaterally changing his strategy. For example, for strategy profile (HEADS, HEADS), player 2 can increase his utility by changing to "TAILS", for strategy profile (HEADS, TAILS), player 1 can increase his utility by changing to "HEADS", and so on.

It is further notable that Nash Equilibrium may not be the Pareto optimal solution. That is, there may be other strategy profiles under which *all* players achieve higher utilities than under a Nash Equilibrium. To see this, we consider the Prisoner's Dilemma Game mentioned above. It is easy to verify that (CONFESS, CONFESS) is the only Nash equilibrium for the Prisoner's Dilemma Game. However, both players can achieve higher utilities under the strategy profile (SILENT, SILENT).

In the situation that a game does not have a pure strategy Nash Equilibrium (*e.g.,* the Matching Pennies Game), what kind of outcome would emerge, and whether such a outcome is desirable? We will show that if we allow the players to *randomize* over their choice of actions, then we can find an equilibrium, which we call the *mixed strategy Nash Equilibrium.*

In order to formalize this notion, we will require some new notation. Let σ_i denote a mixed strategy for player i, which is a probability distribution function (or probability mass function for finite set $\mathcal{S}_i$) over all pure strategies $s_i \in \mathcal{S}_i$. For example, in the Matching Pennies Game, $\sigma_1 = (0.4, 0.6)$ is a mixed strategy for player 1, which states that player 1 picks "HEADS" with probability 0.4 and "TAILS" with probability 0.6. Let Σ_i denote the set of all mixed strategies of player i, *i.e.,* all probability distributions over $\mathcal{S}_i$. Let $\boldsymbol{\sigma} = (\sigma_i)_{i \in \mathcal{I}} \in \Sigma$ denote a mixed strategy profile for all players, where $\Sigma = \Pi_i \Sigma_i$ is the set of all mixed strategy profiles. Furthermore, let $\boldsymbol{\sigma}_{-i} = (\sigma_j)_{j \neq i}$ denote a mixed strategy profile for all players except i, and $\Sigma_{-i} = \Pi_{j \neq i} \Sigma_j$ denote the set of mixed strategy profile for all players except i.

Each player i's payoff under a mixed strategy profile $\boldsymbol{\sigma}$ is given by the expected value of pure strategy payoffs under the distribution σ. More precisely, we have

$$u_i(\boldsymbol{\sigma}) = \sum_{\boldsymbol{s} \in \mathbf{S}} \left(\Pi_{j=1}^{I} \sigma_j(s_j) \right) \cdot u_i(\boldsymbol{s}), \tag{5.1}$$

where $\boldsymbol{s} = (s_j)_{j\in\mathcal{I}}$ is a pure strategy profile, and $\Pi_{j=1}^{I}\sigma_j(s_j)$ is the probability of choosing a particular pure strategy profile $\boldsymbol{s}$.

Based on above, the mixed strategy Nash Equilibrium is defined as follows (49; 50).

Definition 5.7 **Mixed Strategy Nash Equilibrium.** A mixed strategy profile $\boldsymbol{\sigma}^*$ is a mixed strategy Nash Equilibrium if for every player i,

$$u_i(\sigma_i^*, \boldsymbol{\sigma}_{-i}^*) \geq u_i(\sigma_i', \boldsymbol{\sigma}_{-i}^*), \quad \forall \sigma_i' \in \Sigma_i.$$

Let $\mathtt{supp}(\sigma_i)$ denote the support of σ_i, defined as the set $\mathtt{supp}(\sigma_i) \triangleq \{s_i \in \mathcal{S}_i \mid \sigma_i(s_i) > 0\}$, that is, the support of σ_i is the set of pure strategies which are assigned positive probability. We have the following useful lemma for mixed strategy Nash Equilibrium.

Lemma 5.8 *A mixed strategy profile $\boldsymbol{\sigma}^*$ is a mixed strategy Nash Equilibrium if and only if for every player $i \in \mathcal{I}$, the following two conditions hold:*

1. *Every chosen action is equally good, that is, the expected payoff given $\boldsymbol{\sigma}_{-i}^*$ to every $s_i \in \mathtt{supp}(\sigma_i)$ is the same.*
2. *Every non-chosen action is not good enough, that is, the expected payoff given $\boldsymbol{\sigma}_{-i}^*$ to every $s_i \notin \mathtt{supp}(\sigma_i)$ must be no larger than the expected payoff to $s_i \in \mathtt{supp}(\sigma_i)$.*

Intuitively, the lemma states that for a player i, every action in the support of a mixed strategy Nash Equilibrium is a best response to $\boldsymbol{\sigma}_{-i}^*$. This lemma follows from the fact that if the strategies in the support have different payoffs, then it would be better to just take the pure strategy with the highest expected payoff. This would contradict the assumption that $\boldsymbol{\sigma}^*$ is a Nash Equilibrium. Using the same argument, it follows that the pure strategies which are not in the support must have lower (or equal) expected payoffs.

Recall the Matching Pennies Game mentioned above, it is easy to find that $\boldsymbol{\sigma}^* = (\sigma_1^*, \sigma_2^*)$ with $\sigma_i^* = (0.5, 0.5)$, $i = 1, 2$, is the mixed strategy Nash Equilibrium. It is easy to check that given player 2's equilibrium strategy $\sigma_2^* = (0.5, 0.5)$, player 1 achieves the same expected payoff in both pure strategies.

A following important problem is that under which conditions a strategic form game is guaranteed to produce a (mixed or pure) Nash equilibrium. The most important results regarding this are listed in the following theorems.

Theorem 5.9 **Existence (Nash 1950).** *Any finite strategic game has at least one mixed strategy Nash Equilibrium.*

Theorem 5.10 **Existence (Debreu-Fan-Glicksburg 1952).** *The strategic form game* $\langle \mathcal{I}, (\mathcal{S}_i)_{i\in\mathcal{I}}, (u_i)_{i\in\mathcal{I}} \rangle$ *has a pure Nash equilibrium, if for each* $i \in \mathcal{I}$ *the following condition hold:*

1. $\mathcal{S}_i$ *is non-empty, convex, and compact subset of a finite-dimensional Euclidean space.*

2. $u_i(\boldsymbol{s})$ *is continuous in* $\boldsymbol{s}$*, and quasi-concave in* s_i*.*

Both theorems can be proved by the Kakutani fixed point theorem (see (53)). Theorem 5.9 presents the existence of Nash equilibrium in a strategic form game with finite pure strategy sets (*e.g.,* the Matching Pennies Game and Prisoner's Dilemma Game). The Theorem 5.10 presents the existence of a pure Nash equilibrium in a strategic form game with infinite pure strategy sets (*e.g.,* the Cournot Competition Game). The existence of a mixed Nash equilibrium in this case is a special case of Theorem 5.10, since the expected payoff of each player in a mixed strategy is a convex combination of expected payoffs under pure strategies, and thus satisfies both assumptions in the theorem. In fact, a pure Nash equilibrium is a special case of a mixed Nash equilibrium.

Now we give a brief summary for Nash equilibrium. Given that Nash equilibrium is a very widely used notion, a natural question is why one should expect the Nash equilibrium to be the outcome in a strategic form game. One justification is that, since it represents a steady state situation, rational players somehow should reason their way to Nash equilibrium strategies; that is Nash equilibrium might arise through introspection. This justification requires that players are rational and know the payoff functions of all players, that they know their opponents are rational and know the payoff functions, that they know the opponents know and so on. A second justification is that Nash equilibria are self-enforcing. If players agree on a strategy profile before independently choosing their actions, then no player has an incentive to deviate if the agreed strategy profile is a Nash equilibrium.

5.1.4 EXTENSIVE FORM GAME

We have studied the strategic form games which are used to model one-shot games, in which each player chooses his action once and all players act simultaneously. In this section, we will study *extensive form games*, where players engage in sequential decision making (49; 50). Our focus will be on multi-stage games with observed actions where:

1. All previous actions (called history) are observed, *i.e.,* each player is perfectly informed of all previous events.

2. Some players may move simultaneously at some stage k.

Extensive form games can be conveniently represented by *tree* diagrams. To show this, we provide the "Market Entry Game" as an intuitive example. There are two players.

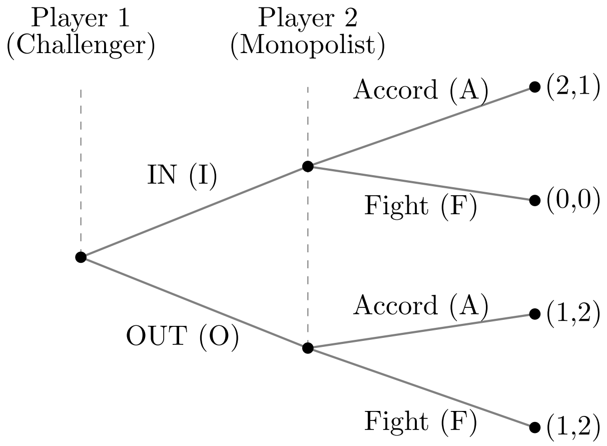

Figure 5.1: Market Entry Game.

Player 1, the challenger, can choose to enter the market (I) or stay out (O). Player 2, the monopolist, after observing the action of the challenger, chooses to accommodate him (A) or fight with him (F). The detailed process is shown in Figure 5.1. Note that when player 1 chooses "Out", there will be no difference for the player 2 to choose "Fight" or "Accord".

An extensive form game can be formally defined as follows (49; 50).

Definition 5.11 Extensive Form Game. An extensive form game consists of four main elements:

1. A set of players, $\mathcal{I} = \{1, 2, ..., I\}$.

2. Histories: A set $\mathcal{H}$ of sequences which can be finite or infinite, defined by

$$\begin{cases} \boldsymbol{h}^0 = \emptyset & \text{initial history} \\ \boldsymbol{h}^1 = \{\boldsymbol{s}^0\} & \text{history at stage 1} \\ \boldsymbol{h}^2 = \{\boldsymbol{s}^0, \boldsymbol{s}^1\} & \text{history at stage 2} \\ \quad ... & \quad ... \\ \boldsymbol{h}^k = \{\boldsymbol{s}^0, ..., \boldsymbol{s}^{k-1}\} & \text{history at stage k} \end{cases}$$

   where $\boldsymbol{s}^t = (s_1^t, s_2^t, ..., s_I^t)$ is the action profile at stage t.

   If the game has a finite number $(K+1)$ of stages, then it is a finite horizon game. Let $\mathcal{H}^k = \{\boldsymbol{h}^k\}$ be the set of all possible histories at stage k. Then $\mathcal{H}^{K+1} = \{\boldsymbol{h}^{K+1}\}$ is the set of all possible terminal histories (after stage K), and $\mathcal{H} = \bigcup_{k=0}^{K+1} \mathcal{H}^{K+1}$ is the set of all possible histories. Consider the "Market Entry Game" in Figure 5.1, we have: $\mathcal{H}^1 = \{\text{I, O}\}$ and $\mathcal{H}^2 = \{\text{(I, A), (I, F), (O, A), (O, F)}\}$.

3. Pure strategies for player i is defined as a contingency plan for every possible history. Let $\mathcal{S}_i(\boldsymbol{h}^k)$ be the set of actions available to player i under history $\boldsymbol{h}^k$, and $\mathcal{S}_i(\mathcal{H}^k) = \bigcup_{\boldsymbol{h}^k \in \mathcal{H}^k} \mathcal{S}_i(\boldsymbol{h}^k)$ be the set of actions available to player i under all possible histories at stage k. Let $a_i^k : \mathcal{H}^k \to \mathcal{S}_i(\mathcal{H}^k)$ be a mapping from $\mathcal{H}^k$ to $\mathcal{S}_i(\mathcal{H}^k)$ such that $a_i^k(\boldsymbol{h}^k) \in \mathcal{S}_i(\boldsymbol{h}^k)$. Then the pure strategy of player i is the set of all sequences $s_i = \{a_i^k\}_{k=0}^K$. The path of strategy profile $\boldsymbol{s}$ is $\boldsymbol{s}^0 = \boldsymbol{a}^0(\boldsymbol{h}^0)$, $\boldsymbol{s}^1 = \boldsymbol{a}^1(\boldsymbol{s}^0)$, $\boldsymbol{s}^2 = \boldsymbol{a}^2(\boldsymbol{s}^0, \boldsymbol{s}^1)$, and so on, where $\boldsymbol{a}^k(\cdot) = \big(a_1^k(\cdot), ..., a_I^k(\cdot)\big)$.

4. Preferences are defined on the outcome of the game $\mathcal{H}^{K+1}$ (after stage K). We can represent the preferences of player i by a utility function $u_i : \mathcal{H}^{K+1} \to \mathbb{R}$. As the strategy profile $\boldsymbol{s}$ determines the path $\boldsymbol{s}^0, ..., \boldsymbol{s}^k$, and hence $\boldsymbol{h}^{K+1}$, we will denote $u_i(\boldsymbol{s})$ as the payoff to player i under strategy profile $\boldsymbol{s}$.

It is notable that in an extensive form game, a strategy specifies the action the player chooses for *every* possible history. Consider the “Market Entry Game” in Figure 5.1. Player 1 moves in the first stage and player 2 moves in the second stage. The strategy of player 1 is the function $a_1^0 : \mathcal{H}^0 = \emptyset \to \mathcal{S}_1 = \{\text{I, O}\}$. The strategy of player 2 is the function $a_2^1 : \mathcal{H}^1 = \{\text{I, O}\} \to \mathcal{S}_2(\mathcal{H}^1)$. There are four possible strategies for player 2, which we can represent as AA, AF, FA, and FF, each corresponding to a contingency plan of player 2 for every possible history in $\mathcal{H}^1 = \{\text{I, O}\}$. That is, the strategy AA means player 2 will select “Accord” under both histories in $\mathcal{H}^1 = \{\text{I, O}\}$, the strategy FA means player 2 will select “Fight” under history $\boldsymbol{h}^1 = \{\text{I}\}$ and “Accord” under history $\boldsymbol{h}^1 = \{\text{O}\}$, and so on. If the strategy profile is (I, AF) or (I, AA), then the outcome will be $\{\text{I, A}\}$. On the other hand, if the strategy profile is (O, FA) or (O, AA), then the outcome will be $\{\text{O, A}\}$.

Based on above discussions, we can usually represent the extensive form game by an equivalent strategic form. The following matrix shows the strategic form of the “Market Entry Game” in Figure 5.1. Each row denotes the action of player 1, and each column denotes the action of player 2.

	AA	AF	FA	FF
I	$(2,1)$	$(2,1)$	$(0,0)$	$(0,0)$
O	$(1,2)$	$(1,2)$	$(1,2)$	$(1,2)$

5.1.5 SUBGAME PERFECT EQUILIBRIUM

Now we consider what outcome would occur in extensive form games. For strategic form games, we have studied Nash equilibrium as one of the solution concepts. We will similarly define Nash equilibrium for extensive form games. Formally,

Definition 5.12 Nash Equilibrium. A strategy profile $\boldsymbol{s}^*$ is a pure Nash Equilibrium for an extensive form game if for all $i \in \mathcal{I}$,

$$u_i(s_i^*, \boldsymbol{s}_{-i}^*) \geq u_i(s_i', \boldsymbol{s}_{-i}^*), \quad \forall s_i' \in \mathcal{S}_i,$$

where $\mathcal{S}_i = \bigcup_{k=0}^{K} \mathcal{S}_i(\mathcal{H}^k)$ is the strategy space of player i.

We have shown that the Nash equilibrium is a reasonable prediction of the outcome in strategic form games. However, this is *not* always the case in extensive form games. Recall the "Market Entry Game" in Figure 5.1. From its equivalent strategic form representation, we can see that this game has four pure strategy Nash equilibria: (I, AA), (I, AF), (O, FA) and (O, FF). However, the last two Nash equilibria (O, FA) and (O, FF) are not reasonable because they are not optimal for player 2 to choose the action "Fight" after history $\boldsymbol{h}^1 = \{\text{I}\}$. These two equilibria are sustained by the threat of the monopolist to play "Fight" under the history that player 1 chooses "In", but this threat is *non-credible.* In other words, if player 1 chooses "In", a rational player 2 will never select "Fight".

Hence, we will define a new equilibrium notion, called *subgame perfect equilibrium* (SPE), which requires the strategy of each player to be optimal not only at the start of the game but also after every history (54). Let $\boldsymbol{h}^k$ denote a history at stage k. We define $G(\boldsymbol{h}^k)$ as the game from $\boldsymbol{h}^k$ on with

- Histories: $\boldsymbol{h}^{K+1} = \{\boldsymbol{h}^k, \boldsymbol{a}^k, ..., \boldsymbol{a}^K\}$.
- Strategies: $s_{i|\boldsymbol{h}^k}$ is the restriction of s_i to histories in $G(\boldsymbol{h}^k)$.
- Payoffs: $u_i(s_i, \boldsymbol{s}_{-i}|\boldsymbol{h}^k)$ is the payoff of player i after histories in $G(\boldsymbol{h}^k)$.

Such a game is referred to as the *subgame* from history $\boldsymbol{h}^k$. Then a subgame perfect equilibrium is defined as follows:

Definition 5.13 Subgame Perfect Equilibrium. A strategy profile $\boldsymbol{s}^*$ is a subgame perfect equilibrium for an extensive form game if for every history $\boldsymbol{h}^k$, the restriction $s_{i|\boldsymbol{h}^k}^*$ is an Nash equilibrium of the subgame $G(\boldsymbol{h}^k)$.

Recall the "Market Entry Game" in Figure 5.1. The strategy profiles (O, FA) and (O, FF) are not subgame perfect equilibrium, since player 2's strategy "Fight" is not a Nash equilibrium strategy of the subgame from history $\boldsymbol{h}^1 = \{\text{I}\}$.

The definition of subgame perfect equilibrium provides a reasonable prediction for the outcome of an extensive form game. For finite horizon games, the subgame perfect equilibria can be constructed using *backward induction.* The following theorem provides a useful characterization for the subgame perfect equilibrium.

Theorem 5.14 **One-Stage Deviation Principle.** *For finite horizon games, $\boldsymbol{s}^*$ is a subgame perfect equilibrium if and only if for all i, t and $\boldsymbol{h}^t$, we have*

$$u_i(s_i^*, \boldsymbol{s}_{-i}^* | \boldsymbol{h}^t) \geq u_i(s_i, \boldsymbol{s}_{-i}^* | \boldsymbol{h}^t)$$

for all s_i satisfying $s_i(\boldsymbol{h}^t) \neq s_i^*(\boldsymbol{h}^t)$, and $s_{i|\boldsymbol{h}^k}(\boldsymbol{h}^{t+k}) = s_{i|\boldsymbol{h}^k}^*(\boldsymbol{h}^{t+k})$, $\forall k > 0, \boldsymbol{h}^{t+k} \in G(\boldsymbol{h}^k)$.

5.2 THEORY: OLIGOPOLY

Now we consider three classical game formulations for competitions among multiple entities (also called Oligopoly) (55): Cournot model, Bertrand model, and Hotelling model. We use these models to illustrate: (a) the translation of an informal statement of a problem into a strategic form representation of a game; and (b) the computations involved in solving for the game's Nash equilibrium.

5.2.1 COURNOT MODEL

Cournot model is an economic model used to describe interactions among firms (companies) that compete on the amount of output they will produce, which they decide independently of each other and at the same time (56). It is named after Antoine Augustin Cournot (1801-1877). A Cournot model usually has the following key features:

- There are at least two firms producing homogeneous (undifferentiated) products;
- Firms do not cooperate, *i.e.,* there is no collusion;
- Firms compete by setting quantities simultaneously. The total output quantity affects the market price;
- The firms are economically rational and act strategically, usually seeking to maximize profit given their competitors' decisions.

For simplicity, we consider a Cournot model between two firms, $\mathcal{I} = \{1, 2\}$. Each firm i decides the quantity q_i of output he will produce, under a fixed unit producing cost c_i. The market-clearing price is a decreasing function of the total quantity $Q = q_1 + q_2$, denoted by $P(Q)$. In such a competition model, what is the best quantity choice of each firm?

We first translate the problem into a strategic form game–*Coutnot Game*. Recall from Definition 5.1, we have the following strategic form representation of Cournot game:

- The set of players is $\mathcal{I} = \{1, 2\}$,
- The strategies available to each player $i \in \mathcal{I}$ is all nonnegative real number, *i.e.,* $q_i \in [0, \infty)$,

- The payoff received by each player i is a function of both players' strategies, defined by $\Pi_i(q_i, q_{-i}) = q_i \cdot P(Q) - c_i \cdot q_i$.

Next we identify the Nash equilibrium of the Coutnot game. Given player 2's strategy q_2, player 1's payoff (profit) is a function of his quantity q_1,

$$\Pi_1(q_1, q_2) = q_1 \cdot P(q_1 + q_2) - c_1 \cdot q_1.$$

It is easy to check the concavity of $\Pi_1(q_1)$. Thus, the optimal strategy for player 1 (or the best response of player 1) is given by the first-order condition,

$$q_1 \cdot P'(q_1 + q_2) + P(q_1 + q_2) - c_1 = 0.$$

Without loss of generality, we take $P(q_1 + q_2) = a - q_1 - q_2$ as an illustration. Then the best response of player 1 is

$$q_1^* = B_1(q_2) = \frac{a - q_2 - c_1}{2},$$

which is obviously a function of player 2's strategy q_2.

Similarly, given player 1's strategy q_1, the optimal strategy for player 2 or the best response of player 2 is

$$q_2^* = B_2(q_1) = \frac{a - q_1 - c_2}{2},$$

which is a function of player 1's strategy q_1 as well.

Recall from Definition 5.5, a strategy profile (q_1^*, q_2^*) is an Nash equilibrium if every player's strategy is the best response to others' strategies, that is, $q_1^* = B_1(q_2^*)$ and $q_2^* = B_2(q_1^*)$. This directly leads to the following pure strategy Nash equilibrium:

$$q_1^* = \frac{a + c_1 + c_2}{3} - c_1, \quad q_2^* = \frac{a + c_1 + c_2}{3} - c_2.$$

Figure 5.2 illustrates both players' best response functions and the Nash equilibrium. Geometrically, the Nash equilibrium is the intersection of both players' best response curves. Note that we consider a very simple version of Cournot game proposed by Cournot in 1883, and will not go into the details of numerous variations of Cournot games due to space limit. For more information about variations of Cournot games, please refer to (49; 50; 56).

5.2.2 BERTRAND MODEL

Bertrand model is an economic model used to describe interactions among firms (sellers) that set prices and their customers (buyers) that choose quantities at that price (56). It is named after Joseph Louis Francois Bertrand (1822-1900). A Bertrand model has the following key features:

- There are at least two firms producing homogeneous (undifferentiated) products;

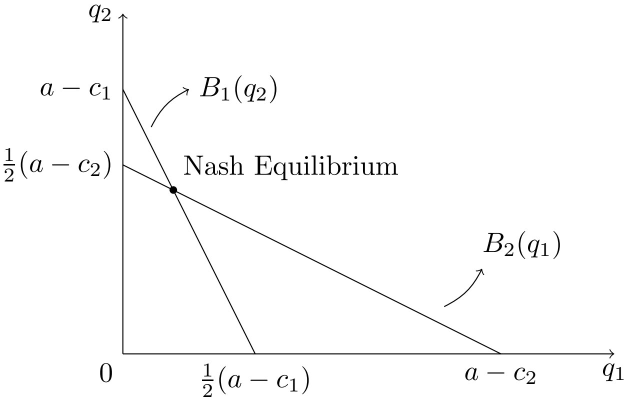

Figure 5.2: Cournot Game.

- Firms do not cooperate, *i.e.,* there is no collusion;
- Firms compete by setting prices simultaneously;
- Consumers buy everything from a firm with a lower price. If all firms charge the same price, consumers randomly select among them.
- The firms are economically rational and act strategically, usually seeking to maximize profit given their competitors' decisions.

Similarly, we consider a Bertrand model between two firms, $\mathcal{I} = \{1, 2\}$. Each firm i chooses the price p_i, rather than quantity as in the Cournot model. Consumers buy from the firm with a lower price, and the total consumer demand is a decreasing function of the market price, denoted by $D(P)$, where $P = \min\{p_1, p_2\}$. In such a competition model, what is the best price choice of each firm? It is important to note that Bertrand model is a different game than Cournot model: the strategy spaces are different, the payoff functions are different, and (as will be shown later) the market outcomes in the Nash equilibria of the two models are different.

The strategic form representation for Bertrand model, called *Bertrand Game*, is shown as follows:

- The set of players is $\mathcal{I} = \{1, 2\}$,
- The strategies available to each player $i \in \mathcal{I}$ is all nonnegative real number, *i.e.,* $p_i \in [0, \infty)$,
- The payoff (profit) received by each player i is a function of both players' strategies, defined by $\Pi_i(p_i, p_{-i}) = (p_i - c_i) \cdot D_i(p_i)$, where c_i is the unit producing cost and $D_i(p_i)$ is the consumers' demand to player i.

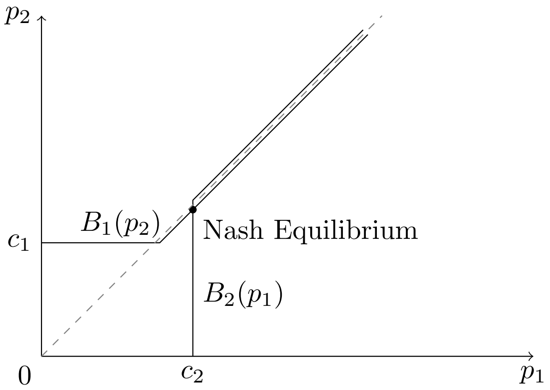

Figure 5.3: Bertrand Game.

Obviously, if player i's price is lower (higher) than the other player $-i$'s, then he gets the total (zero) consumer demand $D(P)$; and if two players' prices are the same, each player gets half of the total consumer demand $D(P)$. That is, $D_i(p_i) = D(p_i)$ if $p_i < p_{-i}$; $D_i(p_i) = 0$ if $p_i > p_{-i}$; and $D_i(p_i) = D(p_i)/2$ if $p_i = p_{-i}$. That is,

Next we identify the Nash equilibrium of the Bertrand game. Given player 2's strategy p_2, player 1's payoff is a function of his price p_1,

$$\Pi_1(p_1, p_2) = \begin{cases} (p_1 - c_i) \cdot D(p_1) & \text{if } p_1 < p_2 \\ \quad 0 & \text{if } p_1 > p_2 \\ (p_1 - c_i) \cdot D(p_1)/2 & \text{if } p_1 = p_2 \end{cases}$$

Thus, given player 2's strategy p_2, the optimal strategy for player 1 or the best response of player 1 is to select a price p_1 slightly lower than p_2, under the constraint that $p_1 \geq c_1$.

Similarly, given player 1's strategy p_1, the optimal strategy for player 2 or the best response of player 2 is to select a price p_2 slightly lower than p_1, under the constraint that $p_2 \geq c_2$. Thus, for any strategy profile (p_1, p_2), both players will gradually decrease their prices, until one player gets to his lowest acceptable price, *i.e.*, his producing cost. Therefore, the Nash equilibrium is given by

$$\begin{cases} p_1^* = [c_2]^-, \quad p_2^* \in [c_2, \infty) & \text{if } c_1 < c_2 \\ p_1^* \in [c_1, \infty), \quad p_2^* = [c_1]^- & \text{if } c_1 > c_2 \\ p_1^* = \ \ p_2^* = c & \text{if } c_1 = c_2 = c \end{cases}$$

where $[x]^-$ denotes the value slightly lower than x. The above Nash equilibrium implies that the lower producing cost firm will extract all the consumer demand, by setting a price slightly than the other firm's producing cost. This is geometrically illustrated in Figure 5.3. Note that the above classic Bertrand model assumes firms compete purely on price, ignoring non-price competition. In a more general case, firms can differentiate their products and charge a higher price. For detailed information, please refer to (49; 50; 56).

5.2.3 HOTELLING MODEL

Hotelling model is an economic model used to study the effect of locations on the competition among two or more firms (56). It is named after Harold Hotelling (1895-1973). A Hotelling model has the following key features:

- There are two firms selling the same good. The firms are located at different points in an interval [0, 1].
- The customers are uniformly distributed along the interval. Customers incur a transportation cost as well as the purchasing cost.
- The firms are economically rational and act strategically, usually seeking to maximize profit given their competitors' decisions.

We take the following model as an example of Hotelling model. Consider a one mile long beach on a hot summer day. There are two identical ice-cream shops on both ends of the beach: store 1 at $x = 0$ and store 2 at $x = 1$. The customers are uniformly distributed with density 1 along this beach. Customers incur a transportation cost w per unit of length (e.g., the value of time spent in travel). Thus, a customer at location $x \in [0, 1]$ will incur a transportation cost of wx when going to store 1 and $w(1 - x)$ when going to store 2.

Each customer comes into one ice-cream and obtains a satisfaction level of $\overline{s}$, which is large enough such that all customers want to purchase one ice-cream from one of the stores. Each store $i \in \{1, 2\}$ chooses a unit price p_i. A customer will choose a store that has the less generalized cost, *i.e.,* price plus transportation cost. Each store wants to choose the price to maximize the profit, by taking the unit cost into consideration.

The strategic form representation for this Hotelling model, called *Hotelling Game*, is shown as follows:

- The set of players is $\mathcal{I} = \{1, 2\}$,
- The strategies available to each player $i \in \mathcal{I}$ is all nonnegative real number, *i.e.,* $p_i \in [0, \infty)$,
- The payoff received by each player i is a function of both players' strategies, defined by $\Pi_i(p_i, p_{-i}) = (p_i - c_i) \cdot D_i(p_i)$, where c_i is the unit producing cost and $D_i(p_i)$ is the ratio of consumers coming to player i (which will be analyzed later).

Next we derive the Nash equilibrium of this game. First, let us compute the location of the customer who is indifferent of choosing either store, $x = l(p_1, p_2)$, where x is given

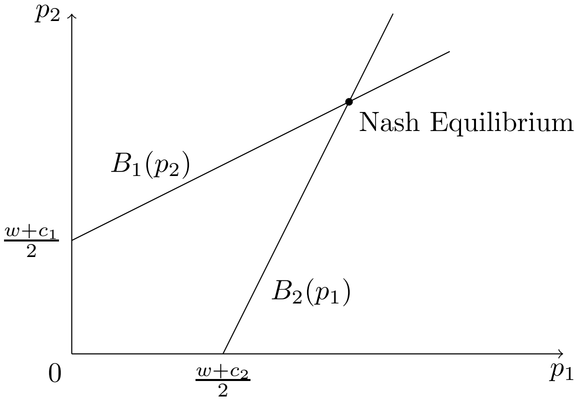

Figure 5.4: Hotelling Game.

by equating the generalized costs,

$$p_1 + w \cdot x = p_2 + w \cdot (1 - x).$$

Thus, the player' respective demand ratios are

$$D_1(p_1, p_2) = l(p_1, p_2) = \frac{p_2 - p_1 + w}{2w}$$

and

$$D_2(p_1, p_2) = 1 - l(p_1, p_2) = \frac{p_1 - p_2 + w}{2w}.$$

Given player 2's price p_2, the profit of player 1 is given by

$$\Pi_1(p_1, p_2) = (p_1 - c_1) \cdot \frac{p_2 - p_1 + w}{2w}.$$

Thus, the optimal strategy for player 1 or the best response of player 1 is given by the first order condition, *i.e.*,

$$p_1^* = B_1(p_2) = \frac{p_2 + w + c_1}{2}.$$

Similarly, given player 1's price p_1, the best response of player 2 is given by

$$p_2^* = B_2(p_1) = \frac{p_1 + w + c_2}{2}.$$

The Nash equilibrium of the Hotelling game is given by $p_1^* = B_1(p_2^*)$ and $p_2^* = B_2(p_1^*)$, *i.e.*,

$$p_1^* = \frac{3w + c_1 + c_2}{3} + \frac{c_1}{3}, \quad p_2^* = \frac{3w + c_1 + c_2}{3} + \frac{c_2}{3}.$$

Figure 5.4 illustrates both players' best response functions and the Nash equilibrium. Geometrically, the Nash equilibrium is the intersection of both players' best response curves. Note that the above classic Hotelling model assumes firms compete purely on price with fixed locations. In a more general case, firms can choose different locations so as to attract more consumers. For detailed information, please refer to (49; 50; 56).

5.3 APPLICATION I: WIRELESS SERVICE PROVIDER COMPETITION REVISITED

Here we revisit the wireless service provider competition model in Section 3.3. In Section 3.3, we solve the social welfare optimization problem, assuming that all providers are under the control of the same entity (*e.g.,* government). In this section, we look at the market competition case, where each provider determines its own price to maximize its own profit. This can be modeled as a multi-leader-follower provider competition game.

5.3.1 PROVIDER COMPETITION GAME

The provider competition game consists of two stages. In the first stage, providers announce prices $\boldsymbol{p} = [p_1, \cdots, p_J]$, where p_j is the unit resource price charged by provider j. In the second stage, each user $i \in \mathcal{I}$ chooses a demand vector $\boldsymbol{q_i} = [q_{i1}, \cdots, q_{iJ}]$, where q_{ij} is the demand to provider j. We denote by $\boldsymbol{q} = [\boldsymbol{q}_1, \cdots, \boldsymbol{q}_I]$ the demand vector of all users.

In the second stage where prices $\boldsymbol{p}$ are known, the goal of user i is to choose $\boldsymbol{q}_i$ to maximize its payoff, which is utility minus payment:

$$v_i(\boldsymbol{q_i}, \boldsymbol{p}) = u_i\left(\sum_{j=1}^{J} q_{ij}c_{ij}\right) - \sum_{j=1}^{J} p_j q_{ij}, \tag{5.2}$$

where c_{ij} is the *channel quality offset* for the channel between user i and the base station of provider j (see Example 3.11 and Assumption 2), and u_i is an increasing and concave utility function. In the first stage, a provider j chooses price p_j to maximize its revenue $p_j \sum_{i=1}^{I} q_{ij}$ subject to the resource constraint $\sum_{i=1}^{I} q_{ij} \le Q_j$, while taking into account the effect of the price on the demand of the users in the second stage. We consider linear pricing with no price discrimination across the users.

5.3.2 ANALYSIS OF THE TWO-STAGE GAME

In this section, we show that there exists a unique equilibrium (defined more precisely shortly) of the multi-leader-follower provider competition game. In particular, this equilibrium corresponds to the unique social optimal solution of SWO and the associated Lagrange multipliers discussed in Section 3.3.3. The idea is to show that the Lagrange multipliers as the prices announced by the providers at the equilibrium. Moreover, we show that there are at most $J - 1$ undecided users at this equilibrium.

First, we define the equilibrium concept (57):

Definition 5.15 (Subgame perfect equilibrium (SPE)). A price demand tuple $(\boldsymbol{p}^*, \boldsymbol{q}^*(\boldsymbol{p}^*))$ is a subgame perfect equilibrium for the provider competition game if no player has an incentive to deviate unilaterally at any stage of the game. In particular, each user $i \in \mathcal{I}$ maximizes its payoff given prices $\boldsymbol{p}^*$. Each provider $j \in \mathcal{J}$ maximizes its revenue given other providers' prices $p^*_{-j} = (p^*_1, \cdots, p^*_{j-1}, p^*_{j+1}, \cdots, p^*_J)$ and the users' demand $\boldsymbol{q}^*(\boldsymbol{p}^*)$.

We will compute the equilibrium concept using backward induction. In Stage II, we will compute the best response of the users $\boldsymbol{q}^*(\boldsymbol{p})$ as a function of any given price vector $\boldsymbol{p}$. Then in Stage I, we will compute the equilibrium prices $\boldsymbol{p}^*$. For equilibrium prices $\boldsymbol{p}^*$, the best response of the users $\boldsymbol{q}^*(\boldsymbol{p}^*)$ is uniquely determined via the BGR decoding.

Equilibrium strategy of the users in Stage II

Consider users facing prices $\boldsymbol{p}$ in the second stage. Each user solves a user payoff maximization (UPM) problem:

$$\textbf{UPM}: \max_{\boldsymbol{q}_i \geq \boldsymbol{0}} v_i = \max_{\boldsymbol{q}_i \geq \boldsymbol{0}} u_i \left(\sum_{j=1}^{J} q_{ij} c_{ij} \right) - \sum_{j=1}^{J} p_j q_{ij} \tag{5.3}$$

Lemma 5.16 *For each user $i \in \mathcal{I}$, there exists a unique nonnegative value x^*_i, such that $\sum_{j=1} c_{ij} q_{ij} = x^*_i$ for every maximizer $\boldsymbol{q}_i$ of the UPM problem. Furthermore, for any j such that $q_{ij} > 0$, $\frac{p_j}{c_{ij}} = \min_{k \in \mathcal{J}} \frac{p_k}{c_{ik}}$.*

Definition 5.17 (Preference set). For any price vector $\boldsymbol{p}$, user i's preference set $\mathcal{J}_i(\boldsymbol{p})$ includes each provider $j \in \mathcal{J}$ with $\frac{p_j}{c_{ij}} = \min_{k \in \mathcal{J}} \frac{p_k}{c_{ik}}$.

In light of Lemma 5.16, $\mathcal{J}_i$ is the set of providers from which user i might request a strictly positive amount of resource. Users can again be partitioned to decided and undecided based on the cardinality of their preference sets, analogous to the distinction made in Section 3.3.3. The preference set of a decided user i contains a singleton, and there is a unique vector $\boldsymbol{q}_i$ that maximizes his payoff. By contrast, the preference set of an undecided user i contains more than one provider, and any choice of $\boldsymbol{q}_i \geq \boldsymbol{0}$ such that $x^*_i = \sum_{j \in \mathcal{J}_i} q_{ij} c_{ij}$ maximizes his payoff.

There is a close relationship between the support sets from Section 3.3.3 and preference sets defined here. Facing prices $\boldsymbol{p}$, a user i *may* request positive resource only from providers who are in his preference set $\mathcal{J}_i$. By definition, he *actually* requests positive resource from providers who are in his support set $\hat{\mathcal{J}}_i$. So the support set of a user is a subset of his preference set. We can construct a BGR based on the preference sets similarly as in Section 3.3.3, and show that this BGR also has no loops with probability 1.

Suppose that the optimal Lagrange multipliers $\boldsymbol{p}^*$ from Section 3.3.3 are announced as prices. Since all users have access to complete network information, each of them can calculate all users' preference sets, and can construct the corresponding BGR. Undecided users can now uniquely determine their demand vector by independently running the same BGR decoding algorithm. The demand found through BGR decoding is unique as all demand vectors are considered at one time and equality of supply and demand is taken into account. We note that the demand found in this way is only one of an undecided user's infinitely many best responses under prices $\boldsymbol{p}^*$. However, only the demands given by the BGR decoding algorithm will balance the supply and demand for each provider at the optimal price $\boldsymbol{p}^*$. We will later show that this is the only subgame perfect equilibrium of the provider competition game.

Equilibrium strategy of the providers in Stage I

The optimal choice of prices for the providers depends on how the users' demand changes with respect to the price, which further depends on the users' utility functions. The quantity that indicates how a user's demand changes with respect to the price is the *coefficient of relative risk aversion* (10) of utility function u_i, i.e. $k_{RRA}^i = -xu_i''(x)/u_i'(x)$. We focus on a class of utility functions characterized in Assumption 4.

Assumption 4 *For each user $i \in \mathcal{I}$, the coefficient of relative risk aversion of its utility function is less than 1.*

Assumption 4 is satisfied by some commonly used utility functions, such as $\log(1+x)$ and the $\alpha-$fair utility functions $\frac{x^{1-\alpha}}{1-\alpha}$, for $\alpha \in (0,1)$. Under Assumption 4, a monopolistic provider will sell all of its resource Q_j to maximize its revenue. Intuitively, when a provider lowers its price, the demand of the users increases significantly enough that the change in revenue of the provider is positive. This encourages the provider to lower the price further such that eventually total demand equals total supply. In the case of multiple providers, Assumption 4 also ensures that all providers are willing to sell all their resources to maximize their revenues.

Theorem 5.18 *Under Assumptions 1, 2, and 3 in Section 3.3.2 and Assumption 4 above, the unique socially optimal demand vector $\mathbf{q}^*$ and the associated Lagrangian multiplier vector $\mathbf{p}^*$ of the SWO problem constitute the unique sub-game perfect equilibrium of the provider competition game.*

Detailed proof of Theorem 5.18 can be found in (34). It is interesting to see that the competition of providers does not reduce social efficiency. This is not a simple consequence of the strict concavity of the users' utility functions; it is also related to the elasticity of users' demands. Assumption 4 ensures that the demands are elastic enough such that a small decrease in price leads to a significant increase in demand and thus a net increase in revenue.

Under the optimal prices $\boldsymbol{p}^*$ announced by the providers in the first stage, the users in the second stage will determine the unique demand vector $\boldsymbol{q}^*$ using BGR decoding. On the other hand, if the providers charge prices other than $\boldsymbol{p}^*$, no best-response from the users will make the demand equals to the supply, which is a necessary condition for an equilibrium.

Since the presence of undecided users makes the analysis challenging, it is interesting to understand how many undecided users there can be in a given game. It turns out that such number is upperbounded by the number of providers J in the network.

Lemma 5.19 *Under any given price vector* $\mathbf{p}$ *in the first stage, the number of undecided users in the second stage is strictly less than* J.

The main idea is that if the number of undecided user nodes in a BGR is not smaller than the number of provider nodes, then there exists a loop in the BGR. This, however, occurs with zero probability, as shown in Section 3.3.3.

Figure 5.5 summarizes the three sets of concepts discussed in Sections 3.3.3, 3.3.4, and 5.3.1.

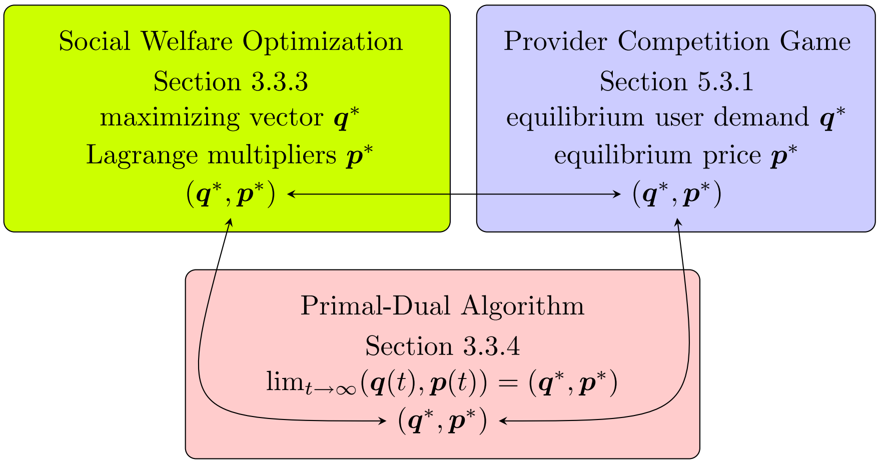
Social Welfare Optimization
Section 3.3.3
maximizing vector $\boldsymbol{q}^*$
Lagrange multipliers $\boldsymbol{p}^*$
$(\boldsymbol{q}^*, \boldsymbol{p}^*)$
Provider Competition Game
Section 5.3.1
equilibrium user demand $\boldsymbol{q}^*$
equilibrium price $\boldsymbol{p}^*$
$(\boldsymbol{q}^*, \boldsymbol{p}^*)$
Primal-Dual Algorithm
Section 3.3.4
$\lim_{t\to\infty}(\boldsymbol{q}(t), \boldsymbol{p}(t)) = (\boldsymbol{q}^*, \boldsymbol{p}^*)$
$(\boldsymbol{q}^*, \boldsymbol{p}^*)$

Figure 5.5: Relationship between different concepts for wireless service provide competition

5.4 APPLICATION II: COMPETITION WITH SPECTRUM LEASING

5.4.1 BACKGROUND

Wireless spectrum is often considered as a scarce resource, and thus has been tightly controlled by the governments through static license-based allocations. However, several recent field measurements show that many spectrum bands are often under-utilized even in densely populated urban areas (2). To achieve more efficient spectrum utilization, secondary users may be allowed to share the spectrum with the licensed primary users. Various dynamic spectrum access mechanisms have been proposed along this direction. One of the proposed mechanisms is dynamic spectrum leasing, where a spectrum owner dynamically transfers and trades the usage right of temporarily unused part of its licensed spectrum to secondary network operators or users in exchange for monetary compensation (58; 59; 60; 61). In this application, we study the competition of two secondary operators under the dynamic spectrum leasing mechanism.

Our study is motivated by the successful operations of mobile virtual network operators (MVNOs) in many countries today. An MVNO does not own wireless spectrum or even the physical infrastructure. It provides services to end-users by *long-term* spectrum leasing agreements with a spectrum owner. MVNOs are similar to the "switchless resellers" of the traditional landline telephone market. Switchless resellers buy minutes wholesale from the large long distance companies and resell them to their customers. It is shown by Dewenter and Haucap (62) that it can be more efficient for the spectrum owner to hire an MVNO as intermediary to retail its spectrum resource, as MVNO has a better understanding of local user population and local users' demand. However, an MVNO is often stuck in a long-term leasing contract with a spectrum owner and can not make flexible spectrum leasing and pricing decisions to match the dynamic demands of the users. The secondary operators considered in this section do not own wireless spectrum either. Compared with a traditional MVNO, the secondary operators can dynamically adjust their spectrum leasing and pricing decisions to match the users' demands that change with users' channel conditions.

Here we study the competition between secondary operators in spectrum acquisition and pricing. The secondary operators will dynamically lease spectrum from spectrum owners, and then compete to sell the resource to the secondary users to maximize their individual profits. We would like to understand how the operators make the equilibrium investment (leasing) and pricing (selling) decisions, considering operators' heterogeneity in leasing costs and wireless users' heterogeneity in transmission power and channel conditions.

5.4.2 NETWORK MODEL

We consider two operators ($i, j \in \{1, 2\}$ and $i \neq j$) and a set $\mathcal{K} = \{1, \ldots, K\}$ of users in an ad hoc network as shown in Fig. 5.6. The operators obtain wireless spectrum from different spectrum owners with different leasing costs, and compete to serve the same set $\mathcal{K}$ of users. Each user has a transmitter-receiver pair. We assume that users are equipped with software defined radios and can transmit in a wide range of frequencies as instructed by the operators,

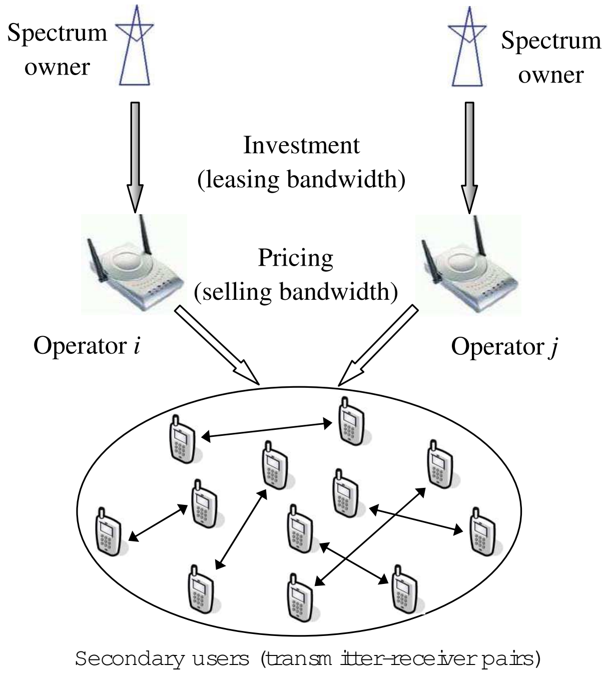

Figure 5.6: Network model for the secondary network operators.

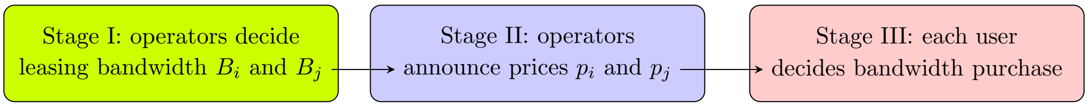
Stage I: operators decide leasing bandwidth B_i and B_j
Stage II: operators announce prices p_i and p_j
Stage III: each user decides bandwidth purchase

Figure 5.7: Three-stage dynamic game: the duopoly's leasing and pricing, and the users' resource allocation

but do not have the capability of spectrum sensing in cognitive radios. Such a network structure puts most of the implementation complexity for dynamic spectrum leasing and allocation on the operators, and thus is easier to implement than a "full" cognitive radio network especially for a large number of users. A user may switch among different operators' services (e.g. WiMAX, 3G) depending on operators' prices. It is important to study the competition among multiple operators as operators are normally not cooperative.

The interactions between the two operators and the users can be modeled as a *three-stage dynamic game*, as shown in Fig. 5.7. Operators i and j first simultaneously determine their leasing bandwidths in Stage I, and then simultaneously announce the prices to the users in Stage II. Finally, each user chooses to purchase bandwidth from *only one operator* to maximize its payoff in Stage III.

Here are several key notations for our problem:

- *Leasing decisions B_i and B_j*: leasing bandwidths of operators i and j in Stage I, respectively.

- *Costs C_i and C_j*: the fixed positive leasing costs per unit bandwidth for operators i and j, respectively. These costs are determined by the negotiation between the operators and their own spectrum suppliers.

- *Pricing decisions p_i and p_j*: prices per unit bandwidth charged by operators i and j to the users in Stage II, respectively.

- *The User k's demand w_{ki} or w_{kj}*: the bandwidth demand of a user $k \in \mathcal{K}$ from operator i or j. A user can only purchase bandwidth from one operator.

5.4.3 USERS' PAYOFFS AND OPERATORS' PROFITS

We assume that the users share the spectrum using OFDM to avoid mutual interferences. If a user $k \in \mathcal{K}$ obtains bandwidth w_{ki} from operator i, then it achieves a data rate (in nats) of

$$r_k(w_{ki}) = w_{ki} \ln\left(\frac{P_k^{\max} h_k}{n_0 w_{ki}}\right), \tag{5.4}$$

where $P_k^{\max}$ is user k's maximum transmission power, n_0 is the noise power density, h_k is the channel gain between user k's transmitter and receiver. The channel gain h_k is independent of the operator, as the operator only sells bandwidth and does not provide a physical infrastructure. We also assume that each user experiences a flat fading over the entire spectrum, such as in the current 802.11d/e standard where the channels are formed through proper interleaving. Here we assume that user k spreads its power $P_k^{\max}$ across the entire allocated bandwidth w_{ki}. Furthermore, we focus on the high SNR regime where $\mathtt{SNR} \gg 1$, such that Shannon capacity $\ln(1+\mathtt{SNR})$ can be approximated by $\ln(\mathtt{SNR})$. To simplify later discussions, we let

$$g_k = \frac{P_k^{\max} h_k}{n_0},$$

thus g_k/w_{ki} is the user k's SNR.

If a user k purchases bandwidth w_{ki} from operator i, it receives a *payoff* of

$$u_k(p_i, w_{ki}) = w_{ki} \ln\left(\frac{g_k}{w_{ki}}\right) - p_i w_{ki}, \tag{5.5}$$

which is the difference between the data rate and the payment. The payment is proportional to price p_i announced by operator i.

For an operator i, its profit is the difference between the revenue and the total cost, i.e.,

$$\pi_i(B_i, B_j, p_i, p_j) = p_i Q_i(B_i, B_j, p_i, p_j) - B_i C_i, \tag{5.6}$$

where $Q_i(B_i, B_j, p_i, p_j)$ and $Q_j(B_i, B_j, p_i, p_j)$ are realized demands of operators i and j. The concept of realized demand will be defined later in Definition 5.23.

5.4.4 ANALYSIS OF THE THREE-STAGE GAME

We will use backward induction to compute the subgame perfect equilibrium (SPE). We will start with Stage III and analyze the users' behaviors given the operators' investment and pricing decisions. Then we look at Stage II and analyze how operators make the pricing decisions taking the users' demands in Stage III into consideration. Finally, we look at the operators' leasing decisions in Stage I knowing the results in Stages II and III.

In the following analysis, we only focus on pure strategy SPE and rule out mixed SPE in the multi-stage game. We use *conditionally* SPE (63) to denote an SPE with pure strategies only, where the network's pure strategies constitute a Nash equilibrium in every subgame. In the following analysis, we derive the conditionally SPE, which is also referred to as equilibrium for simplicity.

Spectrum Allocation in Stage III

In Stage III, each user needs to decide purchase how much spectrum from which operator, based on the prices p_i and p_j announced by the operators in Stage II.

If a user $k \in \mathcal{K}$ obtains bandwidth w_{ki} from operator i, then its payoff $u_k(p_i, w_{ki})$ is given in (5.5). Since this payoff is concave in w_{ki}, the unique *demand* that maximizes the payoff is

$$w_{ki}^*(p_i) = \arg\max_{w_{ki} \geq 0} u_k(p_i, w_{ki}) = g_k \exp(-(1 + p_i)). \tag{5.7}$$

Demand $w_{ki}^*(p_i)$ is always positive, linear in g_k, and decreasing in price p_i. Since g_k is linear in channel gain h_k and transmission power $P_k^{\max}$, then a user with a better channel condition or a larger transmission power has a larger demand.

Next we explain how each user decides which operator to purchase from. The following definitions help the discussions.

Definition 5.20 (Preferred User Set). The Preferred User Set $\mathcal{K}_i^P$ includes the users who prefer to purchase from operator i.

Definition 5.21 (Preferred Demand). The Preferred Demand D_i is the total demand from users in the preferred user set $\mathcal{K}_i^P$, i.e.,

$$D_i(p_i, p_j) = \sum_{k \in \mathcal{K}_i^P(p_i, p_j)} g_k \exp(-(1 + p_i)). \tag{5.8}$$

The notations in (5.8) imply that both set $\mathcal{K}_i^P$ and demand D_i only depend on prices (p_i, p_j) in Stage II and are independent of operators' leasing decisions (B_i, B_j) in Stage I. Such dependance can be discussed in two cases:

1. *Different Prices* $(p_i < p_j)$: every user $k \in \mathcal{K}$ *prefers* to purchase from operator i since

$$u_k(p_i, w_{ki}^*(p_i)) > u_k(p_j, w_{kj}^*(p_j)).$$

We have $\mathcal{K}_i^P = \mathcal{K}$ and $\mathcal{K}_j^P = \emptyset$, and

$$D_i(p_i, p_j) = G\exp(-(1 + p_i)) \text{ and } D_j(p_i, p_j) = 0,$$

where $G = \sum_{k \in \mathcal{K}} g_k$ represents the aggregate wireless characteristics of the users. This notation will be used heavily later in the section.

2. *Same Prices* $(p_i = p_j = p)$: every user $k \in \mathcal{K}$ is indifferent between the operators and randomly chooses one with equal probability. In this case,

$$D_i(p,p) = D_j(p,p) = G\exp(-(1+p))/2.$$

Now let us discuss how much demand an operator can actually satisfy, which depends on the bandwidth investment decisions (B_i, B_j) in Stage I.

It is useful to define the following terms.

Definition 5.22 (Realized User Set). The Realized User Set $\mathcal{K}_i^R$ includes the users whose demands are satisfied by operator i.

Definition 5.23 (Realized Demand). The Realized Demand Q_i is the total demand of users in the Realized User Set $\mathcal{K}_i^R$, i.e.,

$$Q_i\left(B_i, B_j, p_i, p_j\right) = \sum_{k \in \mathcal{K}_i^R(B_i, B_j, p_i, p_j)} g_k \exp(-(1 + p_i)).$$

Notice that both $\mathcal{K}_i^R$ and Q_i depend on prices (p_i, p_j) in Stage II and leasing decisions (B_i, B_j) in Stage I. Calculating the Realized Demands also requires considering two different pricing cases.

1. *Different prices* $(p_i < p_j)$: The Preferred Demands are $D_i(p_i, p_j) = G\exp(-(1 + p_i))$ and $D_j(p_i, p_j) = 0$.

- *If Operator i has enough resource* (*i.e.*, $B_i \geq D_i\,(p_i, p_j)$): all Preferred Demand will be satisfied by operator i. The Realized Demands are

$$\begin{aligned} Q_i &= \min(B_i, D_i(p_i, p_j)) = G\exp(-(1+p_i)), \\ Q_j &= 0. \end{aligned}$$

- *If Operator i has limited resource* (*i.e.*, $B_i < D_i\,(p_i, p_j)$): since operator i cannot satisfy the Preferred Demand, some demand will be satisfied by operator j if it has enough resource. Since the realized demand $Q_i(B_i, B_j, p_i, p_j) = B_i = \sum_{k\in\mathcal{K}_i^R} g_k \exp(-(1+p_i))$, then $\sum_{k\in\mathcal{K}_i^R} g_k = B_i \exp(1+p_i)^2$. The remaining users want to purchase bandwidth from operator j with a total demand of $(G - B_i \exp(1+p_i))\exp(-\,(1+p_j))$. Thus the Realized Demands are

$$\begin{aligned} Q_i &= \min(B_i, D_i(p_i, p_j)) = B_i, \\ Q_j &= \min\left(B_j, \frac{G - B_i\exp(1+p_i)}{\exp\,(1+p_j)}\right). \end{aligned}$$

2. *Same prices* ($p_i = p_j = p$): both operators will attract the same Preferred Demand $G\exp(-(1+p))/2$. The Realized Demands are

$$\begin{aligned} Q_i &= \min\left(B_i, \frac{G}{2\exp(1+p)} + \max\left(\frac{G}{2\exp(1+p)} - B_j, 0\right)\right), \\ Q_j &= \min\left(B_j, \frac{G}{2\exp(1+p)} + \max\left(\frac{G}{2\exp(1+p)} - B_i, 0\right)\right). \end{aligned}$$

Operators' Pricing Competition in Stage II

In Stage II, the two operators simultaneously determine their prices (p_i, p_j) considering the users' preferred demands in Stage III, given the investment decisions (B_i, B_j) in Stage I.

An operator i's profit is defined earlier in (5.6). Since the payment B_iC_i is fixed at this stage, operator i's profit maximization problem is equivalent of maximization of its revenue p_iQ_i. Note that users' total demand Q_i to operator i depends on the received power of each user (product of its transmission power and channel gain). We assume that an operator i knows users' transmission powers and channel conditions. This can be achieved similarly as in today's cellular networks, where users need to register with the operator when they enter the network and frequently feedback the channel conditions. Thus we assume that an operator knows the user population and user demand.

Game 1 (Pricing Game) *The competition between the two operators in Stage II can be modeled as the following game:*

- *Players: two operators i and j.*

[2] Here we consider a large number of users and each user is non-atomic (infinitesimal). Thus an individual user's demand is infinitesimal to an operator's supply and we can claim equality holds for $Q_i = B_i$.

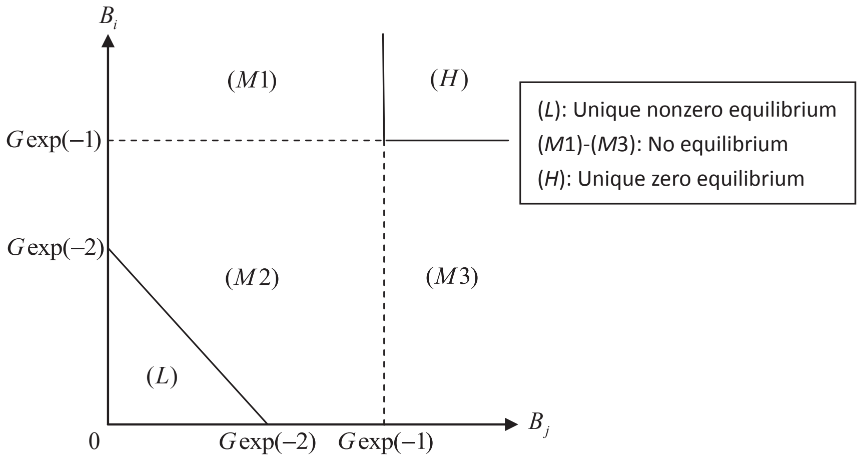
B_i
$(M1)$
(H)
$G\exp(-1)$
$G\exp(-2)$
$(M2)$
$(M3)$
(L)
0
$G\exp(-2)$
$G\exp(-1)$
B_j
(L): Unique nonzero equilibrium
(M1)-(M3): No equilibrium
(H): Unique zero equilibrium

Figure 5.8: Pricing equilibrium types in different (B_i, B_j)

- *Strategy space: operator i can choose price p_i from the feasible set $\mathcal{P}_i = [0, \infty)$. Similarly for operator j.*
- *Payoff function: operator i wants to maximize the revenue $p_i Q_i(B_i, B_j, p_i, p_j)$. Similarly for operator j.*

At an equilibrium of the pricing game, (p_i^*, p_j^*), each operator maximizes its payoff assuming that the other operator chooses the equilibrium price, i.e.,

$$p_i^* = \arg\max_{p_i \in \mathcal{P}_i} p_i Q_i(B_i, B_j, p_i, p_j^*), \quad i = 1, 2, i \neq j.$$

In other words, no operator wants to unilaterally change its pricing decision at an equilibrium.

Next we will investigate the existence and uniqueness of the pricing equilibrium. First, we show that it is sufficient to only consider symmetric pricing equilibrium for Game 1.

Proposition 5.24 *Assume both operators lease positive bandwidth in Stage I, i.e., $\min(B_i, B_j) > 0$. If pricing equilibrium exists, it must be symmetric $p_i^* = p_j^*$.*

The intuition is that no operator will announce a price higher than its competitor to avoid losing its Preferred Demand. This property significantly simplifies the search for all possible equilibria.

Next we show that the symmetric pricing equilibrium is a function of (B_i, B_j) as shown in Fig. 5.8.

Theorem 5.25 *The equilibria of the pricing game are as follows.*

- *Low Investment Regime ($B_i + B_j \leq G\exp(-2)$ as in region (L) of Fig. 5.8): there exists a unique nonzero pricing equilibrium*

$$p_i^*(B_i, B_j) = p_j^*(B_i, B_j) = \ln\left(\frac{G}{B_i + B_j}\right) - 1. \tag{5.9}$$

  The operators' profits in Stage II are

$$\pi_{II,i}(B_i, B_j) = B_i\left(\ln\left(\frac{G}{B_i + B_j}\right) - 1 - C_i\right), \tag{5.10}$$

$$\pi_{II,j}(B_i, B_j) = B_j\left(\ln\left(\frac{G}{B_i + B_j}\right) - 1 - C_j\right). \tag{5.11}$$

- *Medium Investment Regime ($B_i + B_j > G\exp(-2)$ and $\min(B_i, B_j) < G\exp(-1)$ as in regions (M1)-(M3) of Fig. 5.8): there is no pricing equilibrium.*

- *High Investment Regime ($\min(B_i, B_j) \geq G\exp(-1)$ as in region (H) of Fig. 5.8): there exists a unique zero pricing equilibrium*

$$p_i^*(B_i, B_j) = p_j^*(B_i, B_j) = 0,$$

  and the operators' profits are negative for any positive values of B_i and B_j.

Intuitively, higher investments in Stage I will lead to lower equilibrium prices in Stage II. Theorem 5.25 shows that the only interesting case is the low investment regime where both operators' total investment is no larger than $G\exp(-2)$, in which case there exists a unique positive symmetric pricing equilibrium. Notice that same prices at equilibrium do not imply same profits, as the operators can have different costs (C_i and C_j) and thus can make different investment decisions (B_i and B_j) as shown next.

Operators' Leasing Strategies in Stage I

In Stage I, the operators need to decide the leasing amounts (B_i, B_j) to maximize their profits. Based on Theorem 5.25, we only need to consider the case where the total bandwidth of both the operators is no larger than $G\exp(-2)$. We emphasize that the analysis of Stage I is not limited to the case of low investment regime; we actually also consider the medium investment regime and the high investment regime. The key observation is that an SPE will not include any investment decisions (B_i, B_j) in the medium investment regime, as it will not lead to a pricing equilibrium in Stage II. Moreover, any investment decisions in the high investment regime lead to zero operator revenues and are strictly dominated by any decisions in low investment regime. After the above analysis, the operators only need to consider possible equlibria in the low investment regime in Stage I.

Game 2 (Investment Game) *The competition between the two operators in Stage I can be modeled as the following game:*

- *Players: two operators i and j.*
- *Strategy space: the operators will choose (B_i, B_j) from the set $\mathcal{B} = \{(B_i, B_j) : B_i + B_j \leq G\exp(-2)\}$. Notice that the strategy space is coupled across the operators, but the operators do not cooperate with each other.*
- *Payoff function: the operators want to maximize their profits in (5.10) and (5.11), respectively.*

At an equilibrium of the investment game, (B_i^*, B_j^*), each operator has maximized its payoff assuming that the other operator chooses the equilibrium investment, i.e.,

$$B_i^* = \arg \max_{0 \leq B_i \leq \frac{G}{\exp(2)} - B_j^*} \pi_{II,i}(B_i, B_j^*), \quad i = 1, 2, i \neq j.$$

To calculate the investment equilibria of Game 2, we can first calculate operator i's best response given operator j's (not necessarily equilibrium) investment decision, i.e.,

$$B_i^*(B_j) = \arg \max_{0 \leq B_i \leq \frac{G}{\exp(2)} - B_j} \pi_{II,i}(B_i, B_j), \quad i = 1, 2, i \neq j.$$

By looking at operator i's profit in (5.10), we can see that a larger investment decision B_i will lead to a smaller price. The best choice of B_i will achieve the best tradeoff between a large bandwidth and a small price.

After obtaining best investment responses of duopoly, we can then calculate the investment equilibria, given different costs C_i and C_j.

Theorem 5.26 *The duopoly investment (leasing) equilibria in Stage I are summarized as follows.*

- *Low Costs Regime ($0 < C_i + C_j \leq 1$, as region (L) in Fig. 5.9): there exists infinitely many investment equilibria characterized by*

  $$B_i^* = \rho G\exp(-2), \;\; B_j^* = (1-\rho)G\exp(-2), \tag{5.12}$$

  where ρ can be any value that satisfies

  $$C_j \leq \rho \leq 1 - C_i. \tag{5.13}$$

  The operators' profits are

  $$\pi_{I,i} = B_i^*(1 - C_i), \;\; \pi_{I,j} = B_j^*(1 - C_j).$$

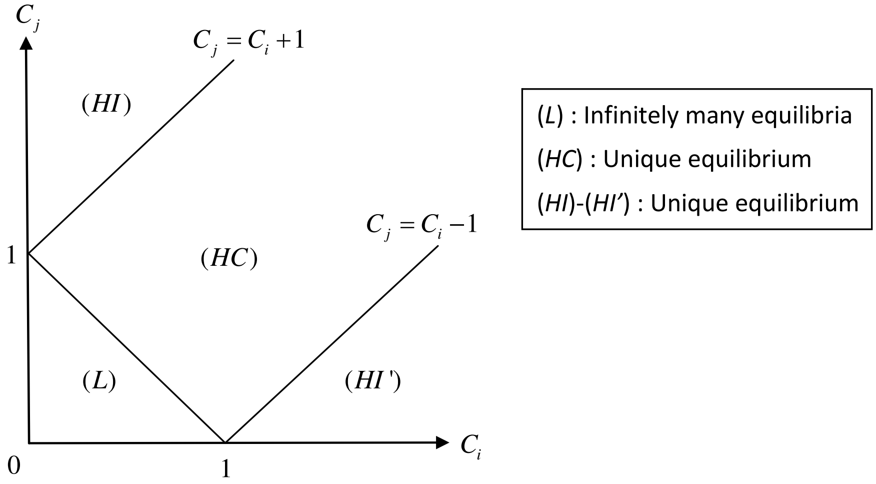
C_j
$C_j = C_i + 1$
(HI)
(HC)
$C_j = C_i - 1$
1
(L)
(HI')
0
1
C_i
(*L*) : Infinitely many equilibria
(*HC*) : Unique equilibrium
(*HI*)-(*HI'*) : Unique equilibrium

Figure 5.9: Leasing equilibrium types in different (C_i, C_j)

- *High Comparable Costs Regime ($C_i + C_j > 1$ and $|C_j - C_i| \leq 1$, as region (HC) in Fig. 5.9): there exists a unique investment equilibrium*
$$B_i^* = \frac{(1 + C_j - C_i)G}{2} \exp(-\frac{C_i + C_j + 3}{2}), \tag{5.14}$$
$$B_j^* = \frac{(1 + C_i - C_j)G}{2} \exp(-\frac{C_i + C_j + 3}{2}). \tag{5.15}$$
The operators' profits are
$$\pi_{I,i} = \left(\frac{1 + C_j - C_i}{2}\right)^2 G \exp(-\left(\frac{C_i + C_j + 3}{2}\right)),$$
$$\pi_{I,j} = \left(\frac{1 + C_i - C_j}{2}\right)^2 G \exp(-\left(\frac{C_i + C_j + 3}{2}\right)).$$

- *High Incomparable Costs Regime ($C_j > 1 + C_i$ or $C_i > 1 + C_j$, as regions (HI) and (HI′) in Fig. 5.9): For the case of $C_j > 1 + C_i$, there exists a unique investment equilibrium with*
$$B_i^* = G \exp(-(2 + C_i)), \; B_j^* = 0,$$
i.e., operator i acts as the monopolist and operator j is out of the market. The operators' profits are
$$\pi_{I,i} = G \exp(-(2 + C_i)), \; \pi_{I,j} = 0.$$
The case of $C_i > 1 + C_j$ can be analyzed similarly.

5.5 CHAPTER SUMMARY

This chapter discusses how multiple players compete with each other in a market, where pricing is one of the major decisions to make.

To understand the competition, we first introduce the basis of game theory. Game theory describes how multiple strategic players make their decisions to maximize their own payoffs, by taking the other players' decisions into consideration. This chapter introduces the basics of noncooperative static and dynamic games with complete information. We first introduce the strategic form game, which is often used to model the simultaneous decisions of all players (also called static game). We define several important concepts including strategy, best response, and finally the Nash equilibrium. We further differentiate between pure strategy Nash equilibrium and mixed strategy Nash equilibrium, and show several classifical existence results that date back to the 50's. Then we move on to introduce the extensive form game, where players make sequential decisions (also called dynamic game). In such a game, the game history becomes very important, and the strategy is no long a single action but a contingency plan based on the game history. We further introduce the concept of subgame perfect equilibrium, which is a generalization of the Nash equilibrium in a dynamic game.

With the knowledge of game theory, we are able to understand the oligopoly models, which characterize the competition between multiple firms in the same market. We introduce three types of oligopoly models: the Cournot model where firms compete based on quantity, the Bertrand model where firms compete based on pricing, and the Hotelling model that captures the impact of locations on the competition. Although we have used two firms (duopoly) as examples when introducing these three models, the results can be easily generalized to the case of more than two firms.

We illustrate the application of theory using two applications. The first one is a revisit of the wireless service provider competition in Section 3.3. Instead of looking at the social optimal pricing as in Section 3.3, here we study how multiple providers will set their prices to attract users, by considering the users' locations and other providers' prices into consideration. We model the interactions by a multi-leader-follower game. Perhaps the most surprising result is that under proper conditions of the utility function, we can show that the unique subgame perfect equilibrium of the game is exactly the same as the unique global optimal solution of the social welfare optimization problem. This result holds regardless of the number of providers in the network, and thus is quite general and encouraging in practice. The second application considers a duopoly between two secondary operators, who will decide their capacity investments through spectrum leasing and market competition through spectrum pricing. We model the interactions between the operators and users as a three-stage multi-leader-follower game, and derive the conditionally SPE with pure strategies in each stage of the game. It turns out that when the leasing costs are low for both providers, then they engage in severe market competition and there are infinitely many equilibria. When the leasing costs are higher, the market will have a unique equilibrium, either two operators sharing the market or only the lower cost operator dominates the market. For more details especially mathematically proofs related to the two applications, please see (34; 64).

CHAPTER 6

Network Externalities

In previous chapters, we assumed that a decision made by a consumer or producer has no external effects on other consumers or producers who are not directly involved. That is, we have considered the interaction and effect between the directly involved players only.

In practical, however, there are many situations where external or third-party effects are important. In these situations, third parties' actions lead to either benefits or costs to players who are not involved directly. In economics, these external effects are termed as *externalities* (65; 66).

In this chapter, we will first introduce the basic theory of externality following (65; 66). Then we present two wireless examples. In the first one, users generate negative externalities to each other due to interferences. The key idea to resolve this is to internalize the externality through Pigovian tax. In the second example, two wireless networks interconnect with other to increase the customer base. We examine how wireless networks determine the access pricing to maximize either the social welfare or their individual profits.

6.1 THEORY: NETWORK EXTERNALITIES

In this section, we cover the basic concepts of externalities, and study some classic externalities. We will see how externalities can be a source of market or network inefficiency, and study some approaches to combate such inefficiencies.

6.1.1 WHAT IS EXTERNALITY?

Simply speaking, externalities are benefits or costs that are imposed by the actions of one player on a third-party not directly involved (65; 66). The neighbors who breathe the smoke, a wireless user who experience the interferences from nearby transmitters, the shoppers who enjoy the department store Christmas displays – these are all good examples of consumers experiencing the costs or benefits imposed by other consumers. Such costs and benefits are said to be *external* and are thus called *externalities*. External costs (like the smoke and interference) are called *negative* externalities, and external benefits (like the pleasure from enjoying the Christmas decorations) are called *positive* externalities.

In the context of a free market, a classic definition of externality suggests that "any indirect effect that either a production or a consumption activity has on the utility function or the consumption set of a consumer, or the production set of a producer, is an *externality* (66)." By "indirect", it is meant that the effect concerns a third-party agent other than the one who exerts this activity or who is involved directly, and the effect is not transferred through prices (non-pecuniary). More simply, an externality is an economic side effect. A general definition of externality is given below.

Definition 6.1 Externality. An externality is any side effect (benefit or cost) that is imposed by the actions of a player on a third-party not directly involved.

As mentioned previously, externalities may be positive or negative. A negative externality is the cost of an externality, also called external cost; and a positive externality is the benefit of an externality, also called external benefit. Classic examples of negative externalities include *pollution* (such as air pollution, water pollution, and noise pollution) and *interference* in wireless networks. In the examples of pollution, the producer or consumer finances the goods produced, but the third-party society must bear the cost of pollution that is introduced into the environment as a by-product. In the example of wireless interference, the mobile user transmits his own data for benefits, but other users will suffer performance degradation in transmission due to the interference caused by this user. One of the most common examples of positive externalities is *network effect*. With network effect, more users consuming goods or service makes a good or service more valuable. Network effect is an important theme in telecommunication networks and online social networks. The more people own telephones or access to a social network, the more valuable the telephone or the social network is to each user, since he can connect to more people by his own telephone or through the social network.[1]

Externalities can cause market failure if the price mechanism does not take into account the external costs and external benefits of production and consumption. The reason is that the producers or consumers are interested in maximizing their profits only. Therefore, they will only take into account the private costs and private benefits arising from their supply or demand of the product, but not the social costs and social benefits. As a consequence, the producers (or consumers) profit maximizing level of supply (or demand) will deviate from the social optimum level. We show this in Figure 6.1.

The first subfigure (left) in Figure 6.1 illustrates the production deviation induced by the negative externality (or external cost) of production. The social cost includes not only the producers' private costs, but also the external costs. Thus, the social marginal cost is larger than the private marginal cost, as shown in the figure. Accordingly, the social optimum level of supply is Q^*, which is smaller than the producers' private profit maximizing level of supply Q_1. Similarly, The second subfigure (right) illustrates the consumption deviation induced by the positive externality (or external benefit) of consumption. The social benefit (or revenue) includes not only the consumers' private benefits, but also the external

[1] The expression "network effect (67; 68)" is applied most commonly to positive network externalities as in telecommunication networks and online social networks. Note that negative network externalities can also occur, where more users consuming a good make the good less valuable. Network congestion is one of the most common examples of negative network externalities.

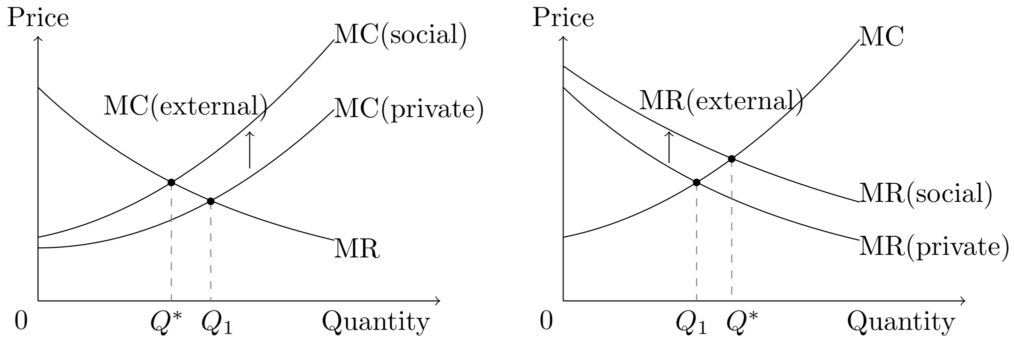

Figure 6.1: Market failures arising from (left) negative production externalities and (right) positive consumption externalities.

benefits. Thus, the social marginal revenue is larger than the private marginal revenue, as shown in the figure. Accordingly, the social optimum level of demand is Q^*, which is larger than the consumers' private profit maximizing level of demand Q_1.

6.1.2 NEGATIVE EXTERNALITY

A negative externality, also called *external cost* or *external diseconomy*, is an action of a player that imposes a negative side effect on a third party (65; 66). As shown previously, negative externalities may cause market failures such as over-production and over-consumption of the product. Many negative externalities are related to the environmental consequences of production and consumption. In wireless networks, a typical negative externality is related to the interference to mobile users caused by other users' transmissions.

Example: Pollution

Pollution is one of the common examples of negative externalities. In a pollution model, the producer benefits from his production activity, but a third party such as the society must bear the cost of pollution. Consumer's consumption activities can also cause pollution (like the smoker).

Consider a simple example of pollution. There are two firms: a chemical plant (CH) and a water company (WT). The chemical plant produces chemical products and discharges wasterwater into a river, which causes the water pollution in the river. Each chemical product is sold at a market equilibrium price of \$10 each unit. That is, the marginal revenue of each unit of product is \$10 for the chemical plant. The water company produces bottled water by drawing water from the river. The chemical plant's wasterwater lowers the quality of the water, and therefore the water company must incur additional costs to purify the water before bottling it. Such an additional cost is the negative externality of

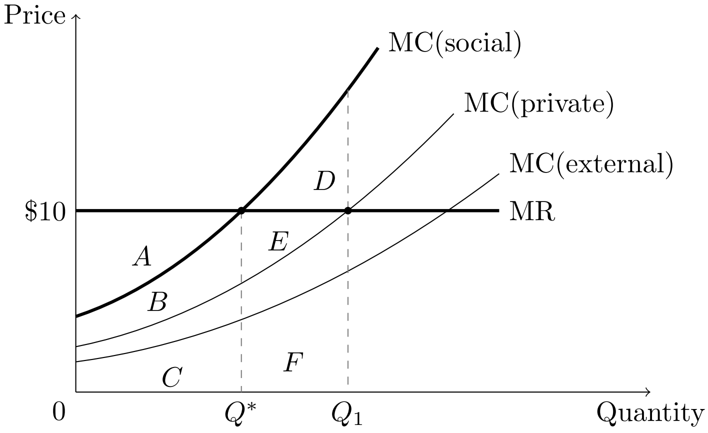

Figure 6.2: Marginal costs in the pollution model.

the production activity of the chemical plant. Figure 6.2 shows the private marginal cost of the chemical plant, and the external marginal cost incurred by the water company. The social cost is the summation of both costs.

Since there is no incentive for the chemical plant to cover the external cost, it will choose the quantity of productions that maximizes its own profit, i.e., that equalizes its private marginal cost and marginal revenue (shown by the point Q_1 in the figure). However, this is not optimal from a social perspective. It is easy to derive that the social optimum quantity is Q^*, which equalizes the social marginal cost and the marginal revenue. Obviously, Q^* is smaller than Q_1.

Now let's consider the chemical plant's and the water company's revenue (or cost) at different level of quantity. As the chemical plant's profit maximizing quantity level Q_1, the chemical plant's total surplus is the sum of areas A, B, and E (denoted by $A + B + E$). The water company's total surplus is $-(C + F)$ due to the external cost of water purifying. Notice that $B = C$ and $F = D + E$, since the social marginal cost is the sum of the private marginal cost of the chemical plant and the external cost incurred by the water company. Thus, the social surplus at quantity Q_1 is $A - D$. At the social optimal quantity level Q^*, the chemical plant's total surplus is $A + B$, and the water company's total surplus is $-C$. Thus, the social surplus at quantity Q^* is A. This shows that with negative externalities (and positive externalities as well), the individual profit maximizing decision may hurt more or less the social surplus.

This example leaves us a key question: *whether it is possible to motivate the chemical plant to produce the social optimal quantity level in the absence of a centralized planner, and how, if so?* Most earlier economists argued that a decentralized competitive price system could reach the social optimum by either costlessly internalizing the externality by government or assessing taxes on the firm creating the negative externality. Due to space limits, we will discuss the second approach briefly as follows.

Solution: Pigovian Tax

Pigovian tax, named after Arthur C. Pigou (1877-1959), is proposed to against the market failure caused by negative externalities. A Pigovian tax is a tax levied on a market activity that generates negative externalities (69). In the presence of negative externalities, the social cost of a market activity is not covered by the private cost of the activity. In such a case, the market outcome is not efficient and may lead to over-production or over-consumption of the product. A Pigovian tax equal to the negative externality is thought to correct the market outcome back to efficiency.

To deal with over-production (or over-consumption similarly), Pigou recommended a tax placed on the offending producer (or consumer) for each unit of production (or consumption). If the government can accurately measure the external cost or social cost, the tax could equalize the marginal private cost and the marginal social cost, and therefore eliminate the market failure. In more specific terms, the producer would have to pay for the externality cost that it created. This would effectively reduce the quantity of the product produced, moving the economy back to an efficient equilibrium.

Figure 6.3 illustrates the effect of a Pigovian tax on the chemical plant's output. The tax shifts the private marginal cost curve up by the amount of the tax, and the shaded area $A_2 + B$ is the tax revenue. As shown in the figure, with the tax, the chemical plant has the incentive to reduce its output to the socially optimum quantity level Q^*. The chemical plant's surplus at quantity level Q^* is $(A + B) - (A_2 + B) = A_1$, and the water company's total surplus is still $-C$.

Although this Pigovian tax works perfectly in theory, the practical implementation is very difficult due to a lack of complete information on the marginal social cost. Most of the criticism of the Pigovian tax relate to the determination of the marginal social cost and therefore the tax. In fact, Pigou himself also pointed out in (69) that the assumption that the government can determine the marginal social cost of a negative externality and convert that amount into a monetary value is a weakness of the Pigovian tax. Later, we will see that a per unit tax is not always the ideal solution to the externality problem and, in some cases, can lead to an inferior solution.

Further Discussions: The Coase Theorem

In the 1960s, Ronald Coase, a Nobel Prize-winner, argued that the traditional analysis on externality was incomplete (70). In terms of our example, Coase would say that the fundamental difficulty is not that the chemical plant creates an externality but that no one owns the quality of water. Coase would argue that the chemical plant will produce the socially optimal output, if (i) transaction costs are negligible, and (ii) one or the other

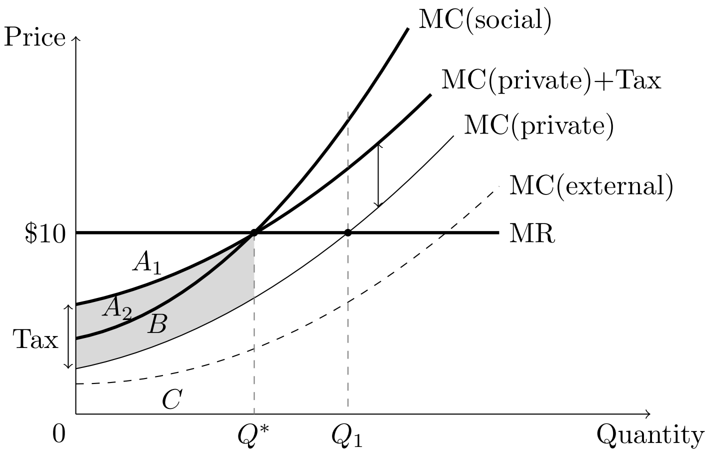

Figure 6.3: Illustration of the Pigovian tax.

party has clearly defined property rights in water quality. Transaction costs refer to the cost of negotiating, verifying, and enforcing contracts. What was truly remarkable at the time was Coase's claim that the output produced by the chemical plant did not depend on which party processes the property right. In other words, even if the chemical plant itself owns the property right of the water quality, it will produce the social optimal quantity level voluntarily!

We show the Coase's claim by the pollution example mentioned above. Suppose the water company owns the right of the water quality. Then it can costlessly collect external costs from the chemical plant, if the chemical plant degrades the water quality. In this case, the marginal cost for the chemical plant would exactly be the sum of its private marginal cost and the external cost. Thus, the chemical plant will select the social optimal output. Things are more surprising in the case that the chemical plant itself owns the right of the water quality. It seems that the chemical plant would over-produce to maximize its profit, but this will not happen. The reason is that with the right of the water quality, the chemical plant essentially has the ability to charge the water company for keeping the water high quality.[2] In other words, the chemical plant can charge the water company a certain money for the decrease of its output from Q_1 to Q^*. Obviously, as long as the money is well designed, e.g., between $[C, C+D]$, both the chemical plant and water company has the incentive to accept such a deal.

Now let us go back to Coase's results. In the previous claims, Coase actually introduced an alternative approach to solve negative externalities, using the *property rights*

[2] Such an ability is achieved by the chemical plant's threat of producing a large number of products (and accordingly degrading the water quality).

theory (70). Based on the property rights theory, Coase pointed in his famous article "The Problem of Social Cost (70)" that if trade in an externality is possible and there are no transaction costs, bargaining will lead to an efficient outcome regardless of the initial allocation of property rights. This is called the *Coase theorem.* Formally,

Theorem 6.2 **Coase Theorem.** *As long as private property rights are well defined and transaction costs are negligible, exchange will eliminate divergence and lead to efficient use of resources or highest valued use of resources.*

Despite of the perfection in theory, there are some criticisms about the practical application of the theorem, among which a key criticism is that the theorem is almost always inapplicable in economic reality, because real-world transaction costs are rarely low enough to allow for efficient bargaining. In fact, Coase also emphasized that the Coase theorem is applied in such a problem that: transactions are "often extremely costly, sufficiently costly at any rate to prevent many transactions that would be carried out in a world in which the pricing system worked without cost (70)."

6.1.3 POSITIVE EXTERNALITY

A positive externality, also called *external benefit* or *external economy*, is an action of a product on consumers that imposes a positive side effect on a third party. As shown previously, positive externalities may cause market failures such as under-production and under-consumption of the product. In computer networking, a typical positive externality is the network effect, where the higher usage of certain products makes them more valuable (67; 68). Recall the examples of telecommunication networks and online social networks, more people using the network increases the network's value to other people. The analysis for positive externalities is very similar to that for negative externalities.

Example: Network Effect

A classic example of positive externalities is that of *network effect* (67; 68). In these situations, a product displays positive network effects when more usage of the product by any user increases the product's value for other users (and sometimes all users).

Network effect is one of the most important underlying economic concepts in industrial organization of IT industries, especially in telecommunication networks and online social networks. Network effects were first studied in the context of long-distance telephony in the early 1970's (one of the earliest papers on the topic is (71)). Today, they are widely recognized as a critical aspect of the industrial organization of IT industries, and are prevalent in a wide variety of sectors, including software, microprocessors, telecommunications, e-commerce and electronic marketplaces. Empirical evidence of network effects has been found in product categories as diverse as spreadsheets, databases, networking equipment, and DVD players.

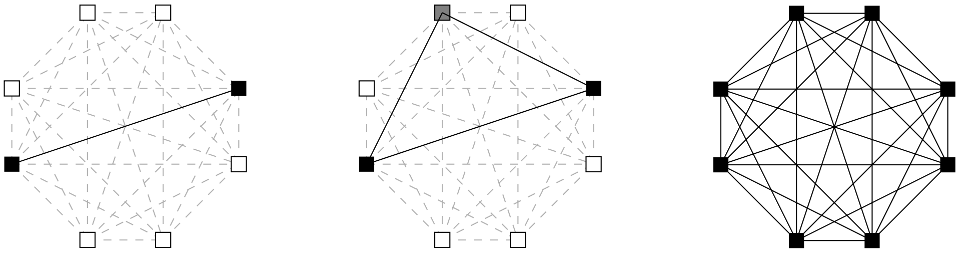

Figure 6.4: Illustration of Network Effect. (I) Each user gets one unit of benefit; (II) With the joining of a new user, each user gets two units of benefit; and (III) When all eight users join the network, each user gets seven units of benefit.

Consider a very simple example of network effect in telecommunication networks, where each user's benefit is simply defined as the number of users connected. The more people own telephones, the more valuable the telephone is to each owner. This creates a positive externality because a user may purchase a telephone without the intention of creating more value for other users, but does so in any case. The reason is that the purchasing activity of one user essentially increases the range that other users can connect to, and accordingly increase the value of other users' telephones. We show this in Figure 6.4, where the solid square denotes the user with a telephone, and the hollow square denotes the user without telephone. When there are only two users in the network, each user can only connect to one person (as shown in subfigure I). If a third person joins the network, each of the two earlier users benefit from the joining of the third user, since now each of them can connect to two persons (as shown in subfigure II). We can similarly see the increase of network value to every existing user as more users join the network (as shown in subfigure III).

Network effects become significant after a certain subscription percentage has been achieved, called *critical mass*. At the critical mass point, the value obtained from the good or service is greater than or equal to the price paid for the good or service. As the value of the good is determined by the user base, this implies that after a certain number of people have subscribed to the service or purchased the good, additional people will subscribe to the service or purchase the good due to the value exceeding the price.

Thus, a key business concern will be: *how to attract users prior to reaching critical mass*. One way is to rely on extrinsic motivation, such as a payment, a fee waiver, or a request for friends to sign up. A more natural strategy is to build a system that has enough value without network effects, at least to early adopters. Then, as the number of users increases, the system becomes even more valuable and is able to attract a wider user base.[3]

[3] Beyond critical mass, the increasing number of subscribers generally cannot continue indefinitely. After a certain point, most networks become either congested or saturated, stopping future uptake. Congestion occurs due to overuse. The applicable analogy is that of a telephone network. While the number of users is below the

This issue is particularly important for online social networks, whose value greatly relies on the number of subscribed users in the network. This is also the reason why the QQ (an online social communicating tool) of Tencent company is so popular in China. Even though Tencent company sometimes provides poor service and charges higher price (for some applications), many new people still prefer the QQ network over other alternatives, because the large number of subscribed users make it more valuable than other social networks. Another example is the Facebook company which just went public. The high evaluation of Facebook is largely because of the enormous subscribed users around the world.

Different Types of Network Effect

There are many ways to classify networks effects. One popular segmentation views network effects as being of four kinds as shown below (67; 68).

1. *Direct network effects.* The simplest network effects are direct: increases in usage lead to direct increases in value. The original example of telephone service is a good illustration of this kind. Another example is online social networks, where users directly benefit from the participation of other users.

2. *Indirect network effects.* Network effects may also be indirect, where increased usage of one product spawns the production of increasingly valuable complementary goods, and this in turn results in an increase in the value of the original product. Examples of complementary goods include software (such as an Office suite for operating systems) and DVDs (for DVD players). This is why Windows and Linux might compete not just for users, but also for software developers. A more precise name for this would be the cross-side network effect, which is different from network benefits that cross distinct markets.

3. *Two-sided network effects.* Network effects can also be two-sided, where the usage increase by one set of users increases the value of a complementary product to another distinct set of users, and vice versa. Hardware/software platforms, reader/writer software pairs, marketplaces and matching services all display this kind of network effects. In many cases, one may think of indirect network effects as an one-directional version of two-sided network effects.

4. *Local network effects.* The structure of an underlying social network affects who benefits from whom. For example, a good displays local network effects when each consumer is influenced directly by the decisions of only a typically small subset of other

congestion point, each additional user adds additional value to every other customer. However, at some point the addition of an extra user exceeds the capacity of the existing system. After this point, each additional user decreases the value obtained by every other user.

consumers, instead of being influenced by the increase of the total number of consumers. Instant messaging is an example of a product that displays local network effects.

6.2 APPLICATION I: DISTRIBUTED WIRELESS INTERFERENCE COMPENSATION

6.2.1 BACKGROUND

Mitigating interference is a fundamental problem in wireless networks. A basic technique for this is to control the nodes' transmit powers. In an ad hoc wireless network power control is complicated by the lack of centralized infrastructure, which necessitates the use of distributed approaches. This application addresses distributed power control for rate adaptive users in a wireless network. We consider a spread spectrum (SS) network, where all users spread their power over a single frequency band. The transmission rate for each user depends on the received signal-to-interference plus noise ratio (SINR). Our objective is to coordinate user power levels to optimize overall performance, measured in terms of total network utility. To achieve this, we propose a protocol in which the users exchange price signals that indicate the negative externality of received interference.

Because we assume that the users cooperate, we ignore incentive issues, which may occur in networks with non-cooperative users. For example, in that scenario a user may attempt to manipulate its announced interference prices to increase its own utility at the expense of the overall network utility. That can, of course, compromise the performance of the distributed algorithms presented here. Although we do not explore this issue further in this work, we note that it may be possible to "hard wire" the power control algorithm into handsets, making such a manipulation of price information difficult.

6.2.2 NETWORK MODEL

We consider a snap-shot of an ad hoc network with a set $\mathcal{M} = \{1, ..., M\}$ of distinct node pairs. As shown in Fig. 6.5, each pair consists of one dedicated transmitter and one dedicated receiver.[4] We use the terms "pair" and "user" interchangeably in the following. In this section, we assume that each user i transmits an SS signal spread over the total bandwidth of B Hz. Over the time-period of interest, the channel gains of each pair are fixed. The channel gain between user i's transmitter and user j's receiver is denoted by h_{ij}. Note that in general $h_{ij} \neq h_{ji}$, since the latter represents the gain between user j's transmitter and user i's receiver.

[4] For example, this could represent a particular schedule of transmissions determined by an underlying routing and MAC protocol.

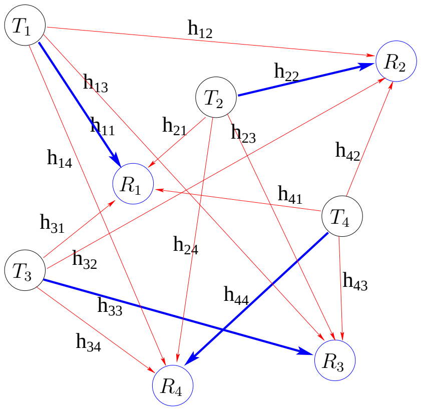

Figure 6.5: An example wireless network with four users (pairs of nodes) (T_i and R_i denote the transmitter and receiver of "user" i, respectively).

Each user i's quality of service is characterized by a utility function $u_i\left(\gamma_i\right)$, which is an increasing and strictly concave function of the received SINR,

$$\gamma_i\left(\boldsymbol{p}\right) = \frac{p_i h_{ii}}{n_0 + \frac{1}{B}\sum_{j \neq i} p_j h_{ji}}, \tag{6.1}$$

where $\boldsymbol{p} = (p_1, \cdots, p_M)$ is a vector of the users' transmission powers and n_0 is the background noise power. The users' utility functions are coupled due to mutual interference. An example utility function is a *logarithmic utility function* $u_i\left(\gamma_i\right) = \theta_i \log\left(\gamma_i\right)$, where θ_i is a user dependent priority parameter.[5]

The problem we consider is to specify $\boldsymbol{p}$ to maximize the utility summed over all users, where each user i must also satisfy a transmission power constraint, $p_i \in \mathcal{P}_i = \left[P_i^{\min}, P_i^{\max}\right]$, i.e.,

$$\max_{\{\boldsymbol{p}: p_i \in \mathcal{P}_i \ \forall i\}} \sum_{i=1}^{M} u_i\left(\gamma_i(\boldsymbol{p})\right). \tag{P1}$$

Note that a special case is $P_i^{\min} = 0$; i.e., the user may choose not to transmit.[6]

Although $u_i(\cdot)$ is concave, the objective in Problem P1 may not be concave in $\boldsymbol{p}$. However, it is easy to verify that any local optimum, $\boldsymbol{p}^* = (p_1^*, ..., p_M^*)$, of this problem will be regular (see p. 309 of (35)), and so must satisfy the Karush-Kuhn-Tucker (KKT) necessary conditions:

[5] In the high SINR regime, logarithmic utility approximates the Shannon capacity $\log\left(1+\gamma_i\right)$ weighted by θ_i. For low SINR, a user's rate is approximately linear in SINR, and so this utility is proportional to the logarithm of the rate.

[6] Occasionally, for technical reasons, we require $P_i^{\min} > 0$; in these cases, $P_i^{\min}$ can be chosen arbitrarily small so that this restriction has little effect. Note that for certain utilities, e.g., $\theta_i \log\left(\gamma_i\right)$, all assigned powers must be strictly positive, since as $P_i \to 0$, the utility approaches $-\infty$.

Lemma 6.3 KKT conditions: *For any local maximum* $\boldsymbol{p}^*$ *of Problem P1, there exist unique Lagrange multipliers* $\lambda_{1,u}^*, ..., \lambda_{M,u}^*$ *and* $\lambda_{1,l}^*, ..., \lambda_{M,l}^*$ *such that for all* $i \in \mathcal{M}$*,*

$$\frac{\partial u_i\left(\gamma_i\left(\boldsymbol{p}^*\right)\right)}{\partial p_i} + \sum_{j \neq i} \frac{\partial u_j\left(\gamma_j\left(\boldsymbol{p}^*\right)\right)}{\partial p_i} = \lambda_{i,u}^* - \lambda_{i,l}^*, \tag{6.2}$$

$$\lambda_{i,u}^*(p_i^* - P_i^{\max}) = 0,\ \ \lambda_{i,l}^*(P_i^{\min} - p_i^*) = 0,\ \ \lambda_{i,u}^*, \lambda_{i,l}^* \geq 0. \tag{6.3}$$

Let

$$\pi_j\left(p_j, p_{-j}\right) = -\frac{\partial u_j\left(\gamma_j\left(p_j, p_{-j}\right)\right)}{\partial I_j\left(p_{-j}\right)}, \tag{6.4}$$

where $I_j\left(p_{-j}\right) = \sum_{k \neq j} p_k h_{kj}$ is the total interference received by user j (before bandwidth scaling). Here, $\pi_j\left(p_j, p_{-j}\right)$ is always nonnegative and represents user j's marginal increase in utility per unit decrease in total interference. Using (6.4), condition (6.2) can be written as

$$\frac{\partial u_i\left(\gamma_i\left(\boldsymbol{p}^*\right)\right)}{\partial p_i} - \sum_{j \neq i} \pi_j\left(p_j^*, p_{-j}^*\right) h_{ij} = \lambda_{i,u}^* - \lambda_{i,l}^*. \tag{6.5}$$

Viewing $\pi_j \left(= \pi_j\left(p_j, p_{-j}\right)\right)$ as a *price* charged to other users for generating interference to user i, condition (6.5) is a necessary and sufficient optimality condition for the problem in which each user i specifies a power level $p_i \in \mathcal{P}_i$ to maximize the following surplus function

$$s_i\left(p_i; p_{-i}, \pi_{-i}\right) = u_i\left(\gamma_i\left(p_i, p_{-i}\right)\right) - p_i \sum_{j \neq i} \pi_j h_{ij}, \tag{6.6}$$

assuming fixed p_{-i} and π_{-i} (i.e., each user is a price taker and ignores any influence he may have on these prices). User i therefore maximizes the difference between its utility minus its payment to the other users in the network due to the interference it generates. The payment is its transmit power times a weighted sum of other users' prices, with weights equal to the channel gains between user i's transmitter and the other users' receivers. This pricing interpretation of the KKT conditions motivates the following asynchronous distributed pricing (ADP) algorithm.

6.2.3 ASYNCHRONOUS DISTRIBUTED PRICING (ADP) ALGORITHM

In the ADP algorithm, each user announces a single price and all users set their transmission powers based on the received prices. Prices and powers are asynchronously updated. For $i \in \mathcal{M}$, let $T_{i,p}$ and $T_{i,\pi}$, be two unbounded sets of positive time instances at which user i updates its power and price, respectively. User i updates its power according to

$$\mathcal{W}_i(p_{-i}, \pi_{-i}) = \arg\max_{\hat{p}_i \in \mathcal{P}_i} \ s_i\left(\hat{p}_i; p_{-i}, \pi_{-i}\right),$$

```
Algorithm 6 The ADP Algorithm
(1) INITIALIZATION: For each user $i \in \mathcal{M}$ choose some power $p_i(0) \in \mathcal{P}_i$ and price $\pi_i(0) \geq 0$.
(2) POWER UPDATE: At each $t \in T_{i,p}$, user $i$ updates its power according to
        $p_i(t) = \mathcal{W}_i\left(p_{-i}(t^-), \pi_{-i}(t^-)\right).$
(3) PRICE UPDATE: At each $t \in T_{i,\pi}$, user $i$ updates its price according to
        $\pi_i(t) = \mathcal{C}_i\left(\boldsymbol{p}(t^-)\right).$
```

which corresponds to maximizing the surplus in (6.6). Each user updates its price according to

$$\mathcal{C}_i(\boldsymbol{p}) = -\frac{\partial u_i\left(\gamma_i\left(\boldsymbol{p}\right)\right)}{\partial I_i\left(p_{-i}\right)},$$

which corresponds to (6.4). Using these update rules, the ADP algorithm is given in Algorithm 6. Note that in addition to being asynchronous across users, each user also need not update its power and price at the same time.[7]

In the ADP algorithm not only are the powers and prices generated in a distributed fashion, but also each user only needs to acquire limited information. To see this note that the power update function can be written as[8]

$$\mathcal{W}_i(p_{-i}, \pi_{-i}) = \left[\frac{p_i}{\gamma_i\left(\boldsymbol{p}\right)} g_i\left(\frac{p_i}{\gamma_i(\boldsymbol{p})}\left(\sum_{j\neq i}\pi_j h_{ij}\right)\right)\right]_{P_i^{\min}}^{P_i^{\max}},$$

where $\frac{p_i}{\gamma_i(\boldsymbol{p})}$ is independent of p_i, and

$$g_i\left(x\right) = \begin{cases} \infty, & 0 \leq x \leq u_i'\left(\infty\right), \\ \left(u_i'\right)^{-1}\left(x\right), & u_i'\left(\infty\right) < x < u_i'\left(0\right), \\ 0, & u_i'\left(0\right) \leq x. \end{cases}$$

Likewise, the price update can be written as

$$\mathcal{C}_i\left(\boldsymbol{p}\right) = \frac{\partial u_i(\gamma_i(\boldsymbol{p}))}{\partial \gamma_i(\boldsymbol{p})} \frac{\left(\gamma_i(\boldsymbol{p})\right)^2}{B p_i h_{ii}}.$$

[7] Of course, simultaneous updates of powers and prices per user and synchronous updating across all users are just special cases of Algorithm 6 .

[8] Notation $[x]_a^b$ means $\max\left\{\min\left\{x, b\right\}, a\right\}$.

From these expressions, it can be seen that to implement the updates, each user i only needs to know: (i) its own utility u_i, the current SINR γ_i and channel gain h_{ii}, (ii) the "adjacent" channel gains h_{ij} for $j \in \mathcal{M}$ and $j \neq i$, and (iii) the price profile $\boldsymbol{\pi}$. By assumption each user knows its own utility. The SINR γ_i and channel gain h_{ii} can be measured at the receiver and fed back to the transmitter. Measuring the adjacent channel gains h_{ij} can be accomplished by having each receiver periodically broadcast a beacon; assuming reciprocity, the transmitters can then measure these channel gains. The adjacent channel gains account for only $1/M$ of the total channel gains in the network; each user does not need to know the other gains. The price information could also be periodically broadcast through this beacon. Since each user announces only a single price, the number of prices scales linearly with the size of the network. Also, numerical results show that there is little effect on performance if users only convey their prices to "nearby" transmitters, i.e., those generating the strongest interference (72).

Denote the set of fixed points of the ADP algorithm by

$$\mathcal{F}^{ADP} \equiv \{(\boldsymbol{p}, \boldsymbol{\pi}) \,|\, (\boldsymbol{p}, \boldsymbol{\pi}) = (\boldsymbol{\mathcal{W}}(\boldsymbol{p}, \boldsymbol{\pi}), \boldsymbol{\mathcal{C}}(\boldsymbol{p}))\}, \tag{6.7}$$

where $\boldsymbol{\mathcal{W}}(\boldsymbol{p}, \boldsymbol{\pi}) = (\mathcal{W}_k(p_{-k}, \pi_{-k}))_{k=1}^{M}$ and $\boldsymbol{\mathcal{C}}(\boldsymbol{p}) = (\mathcal{C}_k(\boldsymbol{p}))_{k=1}^{M}$. Using the strict concavity of $u_i(\gamma_i)$ in γ_i, the following result can be easily shown.

Lemma 6.4 *A power profile* $\mathbf{p}^*$ *satisfies the KKT conditions of Problem P1 (for some choice of Lagrange multipliers) if and only if* $(\boldsymbol{p}^*, \mathcal{C}(\boldsymbol{p}^*)) \in \mathcal{F}^{ADP}$.

If there is only one solution to the KKT conditions, then it must be the global maximum and the ADP algorithm would reach that point if it converges. In general, $\mathcal{F}^{ADP}$ may contain multiple points including local optima or saddle points.

6.2.4 CONVERGENCE ANALYSIS OF ADP ALGORITHM

We next characterize the convergence of the ADP algorithm by viewing it in a game theoretic context. A natural generalization of the NCP game is to consider a game where each player i's strategy includes specifying both a power p_i and a price π_i to maximize a payoff equal to the surplus in (6.6). However, since there is no penalty for user i announcing a high price, it can be shown that each user's best response is to choose a large enough price to force all other users transmit at $P_i^{\min}$. This is certainly not a desirable outcome and suggests that the prices should be determined externally by another procedure.[9] Instead, we consider the following *Fictitious Power-Price (FPP) control game*,

$$G_{FPP} = [\mathcal{FW} \cup \mathcal{FC}, \left\{\mathcal{P}_i^{\mathcal{FW}}, \mathcal{P}_i^{\mathcal{FC}}\right\}, \left\{s_i^{\mathcal{FW}}, s_i^{\mathcal{FC}}\right\}],$$

[9] A similar situation arises in (73), where users in a multi-hop network announce prices charging other users for packets they forward. In that case, the prices also cannot be determined by individual surplus optimizations.

where the players are from the union of the sets $\mathcal{FW}$ and $\mathcal{FC}$, which are both copies of $\mathcal{M}$. $\mathcal{FW}$ is a *fictitious power player set*; each player $i \in \mathcal{FW}$ chooses a power p_i from the strategy set $\mathcal{P}_i^{\mathcal{FW}} = \mathcal{P}_i$ and receives payoff

$$s_i^{\mathcal{FW}}\left(p_i; p_{-i}, \pi_{-i}\right) = u_i\left(\gamma_i\left(\boldsymbol{p}\right)\right) - \sum_{j \neq i} \pi_j h_{ij} p_i. \tag{6.8}$$

$\mathcal{FC}$ is a *fictitious price player set*; each player $i \in \mathcal{FC}$ chooses a price π_i from the strategy set $\mathcal{P}_i^{\mathcal{FC}} = [0, \bar{\pi}_i]$ and receives payoff

$$s_i^{\mathcal{FC}}\left(\pi_i; \boldsymbol{p}\right) = -\left(\pi_i - \mathcal{C}_i\left(\boldsymbol{p}\right)\right)^2. \tag{6.9}$$

Here $\bar{\pi}_i = \sup_{\boldsymbol{p}} \mathcal{C}_i\left(\boldsymbol{p}\right)$, which could be infinite for some utility functions.

In G_{FPP}, each user in the ad hoc network is split into two fictitious players, one in $\mathcal{FW}$ who controls power p_i and the other one in $\mathcal{FC}$ who controls price π_i. Although users in the real network cooperate with each other by exchanging interference information (instead of choosing prices to maximize their surplus), each fictitious player in G_{FPP} is selfish and maximizes its own payoff function. In the rest of this section, a "user" refers to one of the M transmitter-receiver pairs in set $\mathcal{M}$, and a "player" refers to one of the $2M$ fictitious players in the set $\mathcal{FW} \cup \mathcal{FC}$.

In G_{FPP} the players' best responses are given by

$$\mathcal{B}_i^{\mathcal{FW}}\left(p_{-i}, \pi_{-i}\right) = \mathcal{W}_i\left(p_{-i}, \pi_{-i}\right), \forall i \in \mathcal{FW}$$

and

$$\mathcal{B}_i^{\mathcal{FC}}\left(\boldsymbol{p}\right) = \mathcal{C}_i\left(\boldsymbol{p}\right), \forall i \in \mathcal{FC},$$

where $\mathcal{W}_i$ and $\mathcal{C}_i$ are the update rules for the ADP algorithm. In other words, the ADP algorithm can be interpreted as if the players in G_{FPP} employ asynchronous *myopic best response (MBS)* updates, i.e. the players update their strategies according their best responses assuming the other player's strategies are fixed. It is known that the set of fixed points of MBS updates are the same as the set of NEs of a game (74, Lemma 4.2.1). Therefore, we have:

Lemma 6.5 *$(\boldsymbol{p}^*, \boldsymbol{\pi}^*) \in \mathcal{F}^{ADP}$ if and only if $(\boldsymbol{p}^*, \boldsymbol{\pi}^*)$ is a NE of G_{FPP}.*

Together with Lemma 6.4, it follows that proving the convergence of asynchronous MBS updates of G_{FPP} is sufficient to prove the convergence of the ADP algorithm to a solution of KKT conditions. We next analyze this convergence using supermodular game theory (74).

We first introduce some definitions[10]. A real m-dimensional set $\mathcal{V}$ is a *sublattice* of $\mathbb{R}^m$ if for any two elements $a, b \in \mathcal{V}$, the component-wise minimum, $a \wedge b$, and the

[10] More general definitions related to supermodular games are given in (74).

component-wise maximum, $a \vee b$, are also in $\mathcal{V}$. In particular, a compact sublattice has a (component-wise) smallest and largest element. A twice differentiable function f has *increasing differences* in variables (x, t) if $\partial^2 f / \partial x \partial t \geq 0$ for any feasible x and t.[11] A function f is *supermodular* in $\boldsymbol{x} = (x_1, .., x_m)$ if it has increasing differences in (x_i, x_j) for all $i \neq j$.[12] Finally, a game $G = [\mathcal{M}, \{\mathcal{P}_i\}, \{s_i\}]$ is *supermodular* if for each player $i \in \mathcal{M}$, (a) the strategy space $\mathcal{P}_i$ is a nonempty and compact sublattice, and (b) the payoff function s_i is continuous in all players' strategies, is supermodular in player i's own strategy, and has increasing differences between any component of player i's strategy and any component of any other player's strategy. The following theorem summarizes several important properties of these games.

Theorem 6.6 *In a supermodular game $G = [\mathcal{M}, \{\mathcal{P}_i\}, \{s_i\}]$,*

(a) *The set of NEs is a nonempty and compact sublattice and so there is a component-wise smallest and largest NE.*

(b) *If the users' best responses are single-valued, and each user uses MBS updates starting from the smallest (largest) element of its strategy space, then the strategies monotonically converge to the smallest (largest) NE.*

(c) *If each user starts from any feasible strategy and uses MBS updates, the strategies will eventually lie in the set bounded component-wise by the smallest and largest NE. If the NE is unique, the MBS updates globally converge to that NE from any initial strategies.*

Properties (a) follows from Lemma 4.2.1 and 4.2.2 in (74); (b) follows from Theorem 1 of (75) and (c) can be shown by Theorem 8 in (76).

Next we show that by an appropriate strategy space transformation certain instances of G_{FPP} are equivalent to supermodular games, and so Theorem 6.6 applies. We first study a simple two-user network, then extend the results to a M-user network.

Two-user networks

Let G_{FPP}^2 be the FPP game corresponding to a two user network; this will be a game with four players, two in $\mathcal{FW}$ and two in $\mathcal{FC}$. First, we check whether G_{FPP}^2 is supermodular. Each user $i \in \mathcal{FW}$ clearly has a nonempty and compact sublattice (interval) strategy set, and so does each user $i \in \mathcal{FC}$ if $\bar{\pi}_i < \infty$.[13] Each player's payoff function is (trivially) supermodular in its own one-dimensional strategy space. The remaining increasing difference

[11] If we choose x to maximize a twice differentiable function $f(x,t)$, then the first order condition gives $\partial f(x,t) / \partial x|_{x=x^*} = 0$, and the optimal value x^* increases with t if $\partial^2 f / \partial x \partial t > 0$.

[12] A function f is always supermodular in a single variable x.

[13] When $P_i^{\min} = 0$, this bounded price restriction is not satisfied for utilities such as $u_i(\gamma_i) = \theta_i \gamma_i^{\alpha} / \alpha$ with $\alpha \in [-1, 0)$, since $\pi_i = \theta_i \gamma_i^{\alpha+1} / (p_i h_{ii} B)$ is not bounded as $p_i \to 0$. However, as noted above, we can set $P_i^{\min}$ to some arbitrarily small value without effecting the performance.

condition for the payoff functions does *not* hold with the original definition of strategies $(\boldsymbol{p}, \boldsymbol{\pi})$ in G^2_{FPP}. For example, from (6.8),

$$\frac{\partial s_i^{\mathcal{FW}}}{\partial p_i \partial \pi_j} = -h_{ij} < 0, \forall j \neq i,$$

e.g. a higher price leads the other users to decrease their powers. However, if we define $\pi'_j = -\pi_j$ and consider an equivalent game where each user $j \in \mathcal{FC}$ chooses π'_j from the strategy set $[-\bar{\pi}_j, 0]$, then

$$\frac{\partial s_i^{\mathcal{FW}}}{\partial p_i \partial \pi'_j} = h_{ij} > 0, \forall j \neq i,$$

i.e. $s_i^{\mathcal{FW}}$ has increasing differences in the strategy pair $\left(p_i, \pi'_j\right)$ (or equivalently $(p_j, -\pi_j)$). If all the users' strategies can be redefined so that each player's payoff satisfies the increasing differences property in the transformed strategies, then the *transformed FPP game* is supermodular.

Denote

$$CR_i\left(\gamma_i\right) = -\frac{\gamma_i u''_i\left(\gamma_i\right)}{u'_i\left(\gamma_i\right)},$$

and let $\gamma_i^{\min} = \min\{\gamma_i(\boldsymbol{p}) : p_i \in \mathcal{P}_i \, \forall i\}$ and $\gamma_i^{\max} = \max\{\gamma_i(\boldsymbol{p}) : p_i \in \mathcal{P}_i \, \forall i\}$. An increasing, twice continuously differentiable, and strictly concave utility function $u_i\left(\gamma_i\right)$ is defined to be

- Type I if $CR_i\left(\gamma_i\right) \in [1, 2]$ for all $\gamma_i \in \left[\gamma_i^{\min}, \gamma_i^{\max}\right]$;
- Type II if $CR_i\left(\gamma_i\right) \in (0, 1]$ for all $\gamma_i \in \left(\gamma_i^{\min}, \gamma_i^{\max}\right]$.

The term $CR_i\left(\gamma_i\right)$ is called the *coefficient of relative risk aversion* in economics (10) and measures the relative concaveness of $u_i\left(\gamma_i\right)$. Many common utility functions are either Type I or Type II, as shown in Table 6.1.

Table 6.1: Examples of Type I and II utility functions

Type I	Type II
$\theta_i \log(\gamma_i)$	$\theta_i \log(\gamma_i)$
$\theta_i \gamma_i^\alpha / \alpha$ (with $\alpha \in [-1, 0)$)	$\theta_i \gamma_i^\alpha / \alpha$ (with $\alpha \in (0, 1)$)
$1 - e^{-\theta_i \gamma_i}$ (with $\frac{1}{\gamma_i^{\min}} \leq \theta_i \leq \frac{2}{\gamma_i^{\max}}$)	$1 - e^{-\theta_i \gamma_i}$ (with $\theta_i \leq \frac{1}{\gamma_i^{\max}}$)
$a\left(\gamma_i\right)^2 + b\gamma_i + c$ (with $0 \leq -3a\gamma_i^{\max} \leq b \leq -4a\gamma_i^{\min}$)	$a\left(\gamma_i\right)^2 + b\gamma_i + c$ (with $b \geq -4a\gamma_i^{\max} > 0$)
	$\theta_i \log\left(1 + \gamma_i\right)$

The logarithmic utility function is both Type I and II. A Type I utility function is "more concave" than a Type II one. Namely, an increase in one user's transmission power would induce the other users to increase their powers, i.e.,

$$\frac{\partial^2 u_i\left(\gamma_i\left(\boldsymbol{p}\right)\right)}{\partial p_i \partial p_j} \geq 0, \forall j \neq i.$$

A Type II utility would have the opposite effect, i.e.,

$$\frac{\partial^2 u_i\left(\gamma_i\left(\boldsymbol{p}\right)\right)}{\partial p_i \partial p_j} \leq 0, \forall j \neq i.$$

The strategy spaces must be redefined in different ways for these two types of utility functions to satisfy the requirements of a supermodular game.

Proposition 6.7 *G^2_{FPP} is supermodular in the transformed strategies $(p_1, p_2, -\pi_1, -\pi_2)$ if both users have Type I utility functions.*

Proposition 6.8 *G^2_{FPP} is supermodular in the transformed strategies $(p_1, -p_2, \pi_1, -\pi_2)$ if both users have Type II utility functions.*

The proofs of both propositions consist of checking the increasing differences conditions for each player's payoff function. These results along with Theorem 6.6 enable us to characterize the convergence of the ADP algorithm. For example, if the two users have Type I utility functions (and $\bar{\pi}_1, \bar{\pi}_2 < \infty$), then $\mathcal{F}^{ADP}$ is nonempty. In case of multiple fixed points, there exist two extreme ones $\left(\boldsymbol{p}^L, \boldsymbol{\pi}^L\right)$ and $\left(\boldsymbol{p}^R, \boldsymbol{\pi}^R\right)$, which are the smallest and largest fixed points in terms of strategies $(p_1, p_2, -\pi_1, -\pi_2)$. If users initialize with $\left(\boldsymbol{p}\left(0\right), \boldsymbol{\pi}\left(0\right)\right) = \left(P_1^{\min}, P_2^{\min}, \bar{\pi}_1, \bar{\pi}_2\right)$ or $(P_1^{\max}, P_2^{\max}, 0, 0)$, the power and prices converge monotonically to $\left(\boldsymbol{p}^L, \boldsymbol{\pi}^L\right)$ or $\left(\boldsymbol{p}^R, \boldsymbol{\pi}^R\right)$, respectively. If users start from arbitrary initial power and prices, then the strategies will eventually lie in the space bounded by $\left(\boldsymbol{p}^L, \boldsymbol{\pi}^L\right)$ and $\left(\boldsymbol{p}^R, \boldsymbol{\pi}^R\right)$. Similar arguments can be made with Type II utility functions with a different strategy transformation. Convergence of the powers for both types of utilities is illustrated in Fig. 6.6.

M-user Networks

Proposition 6.7 can be easily generalized to a network with $M > 2$:

Corollary 6.9 *For an M-user network if all users have Type I utilities, G_{FPP} is a supermodular in the transformed strategies $\left(\boldsymbol{p}, -\boldsymbol{\pi}\right)$.*

In this case, Theorem 6.6 can again be used to characterize the structure of $\mathcal{F}^{ADP}$ as well as the convergence of the ADP algorithm. On the other hand, it can be seen that

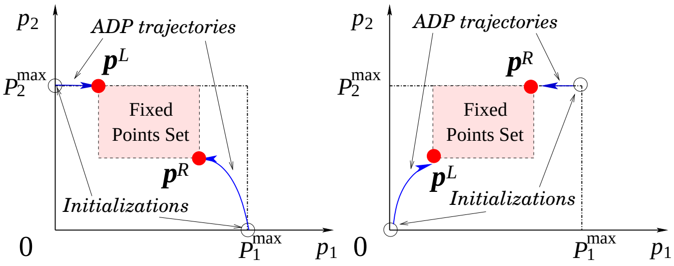

Figure 6.6: Examples of the trajectories of the power profiles under the ADP algorithm for a two-user network with Type I (left) or Type II (right) utility functions. In both cases, from the indicated initializations the power profiles will monotonically converge to the indicated "corner" fixed points.

the strategy redefinition used in Proposition 6.8, can not be applied with $M > 2$ users so that the increasing differences property holds for every pair of users.

With logarithmic utility functions, it is shown in (77) that Problem P1 is a strictly concave maximization problem over the transformed variables $y_i = \log p_i$. In this case Problem P1 has a unique optimal solution, which is the only point satisfying the KKT conditions. It follows from Lemma 6.4 and Lemma 6.5 that G_{FPP} will have a unique NE corresponding to this optimal solution and the ADP algorithm will converge to this point from any initial choice of powers and prices.[14] With some minor additional conditions, the next proposition states that these properties generalize to other Type I utility functions.

Proposition 6.10 *In an M-user network, if for all $i \in \mathcal{M}$:*

a) $P_i^{\min} > 0$, and

b) $CR_i(\gamma_i) \in [a, b]$ for all $\gamma_i \in [\gamma_i^{\min}, \gamma_i^{\max}]$, where $[a, b]$ is a strict subset of $[1, 2]$;

then Problem P1 has a unique optimal solution, to which the ADP algorithm globally converges.

6.2.5 SIMULATION RESULTS

We simulate a network contained in a 10m×10m square area. Transmitters are randomly placed in this area according to a uniform distribution, and the corresponding receiver is randomly placed within 6m×6m square centered around the transmitter. There are $M = 10$ users, each with utility $u_i = \log(\gamma_i)$. The channel gains $h_{ij} = d_{ij}^{-4}$, $P_i^{\max}/n_0$=40dB, and B=128. Figure 6.7 shows the convergence of the powers and prices for each user under the ADP algorithm for a typical realization, starting from random initializations. Also, for comparison we show the convergence of these quantities using a gradient-based algorithm as in (77) with a step-size of 0.01.[15]. Both algorithms converge to the optimal power allocation, but the ADP algorithm converges much faster; in all the cases we have simulated, the ADP algorithm converges about 10 times faster than the gradient-based algorithm (if the latter converges). The ADP algorithm, by adapting power according to the best response updates, is essentially using an "adaptive step-size" algorithm: users adapt the power in "larger" step-sizes when they are far away from the optimal solution, and use finer steps when close to the optimal.

[14] Moreover, if each user $i \in \mathcal{M}$ starts from profile $(p_i(0), \pi_i(0)) = \left(P_i^{\min}, \theta_i/(n_0 B)\right)$ or $\left(P_i^{\max}, 0\right)$, then their strategies will monotonically converge to this fixed point

[15] In our experiments, a larger step-size than 0.01 would often not converge

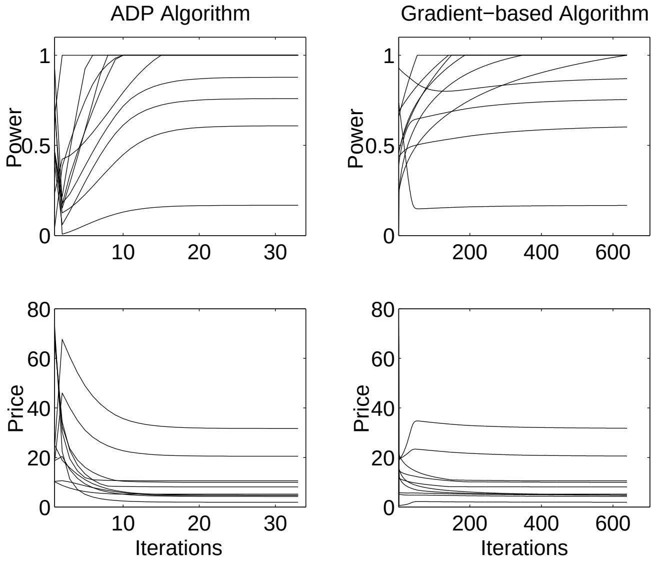

Figure 6.7: Convergence of the prices and powers for the ADP algorithm (left) and a gradient algorithm (right) in a network with 10 users and logarithmic utility functions. Each curve corresponds to the power or price for one user with a random initialization.

6.3 APPLICATION II: WIRELESS ACCESS PRICING

6.3.1 BACKGROUND

Today's wireless networks are interconnected. A subscriber of one network can communicate with subscribers in many other networks. Inter-network communications bring benefits to all networks involved, as no single network needs to build an infrastructure covering the entire market. In other words, the existence of one network brings positive externality to the customers of other networks. However, serving users from other networks brings negative impacts on a network's own available resource. This can be compensated by the *access pricing* charged between network providers. A user who initiates a communication session needs to pay the *user pricing* to its own network provider.

This application focuses on the interaction between user pricing and access pricing in multiple interconnected wireless communication networks. Users communicate with each other through providers' base stations. Users experience different channel conditions to their base stations, and hence require different amount of network resources to achieve the same data rate. The uplink and downlink transmissions of the same communication session may also consume different amount of resources depending on the locations of users and base stations. These unique features make the analysis more challenging compared with the existing access pricing study for wireline networks.

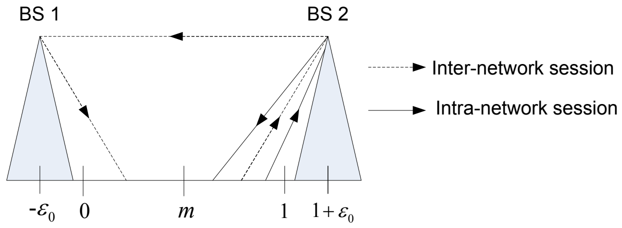

Figure 6.8: An illustration of the wireless market.

6.3.2 SYSTEM MODEL

Wireless Market Model

Following the tradition of network access pricing literature (e.g., (78)), we adopt a one-dimensional market model as in Fig. 6.8. There are two networks, each represented by a base station at one end of the market. Users are uniformly located along the segment [0,1]. The average channel gain between a user and a base station follows the large-scale distance based attenuation. To avoid having an infinitely large channel gain, the base-stations are ϵ_0 away from their closest users. In the rest of the application, we use the terms "network", "provider", and "base station" interchangeably.

We assume that the user-network associations are fixed, such that users in $[0, m]$ subscribe to network 1 and users in $[m, 1]$ subscribe to network 2. This means that network 1 has a market share of m, and network 2 has a market share of $1 - m$. For example, some part of California is covered by AT&T but not Verizon (and vice versa). The analysis based on this restrictive assumption of fixed market share will help us understand the more general scenario of flexible market share.

User's Utility, Payment, Payoff, and Demand

A one-way data communication session involves a source user and destination user. The session is *intra-network* when both users belong to the same network, or *inter-network* when the two users belong to different networks. We adopt the α-fair utility function (79) to represent the quality of service (QoS) of a session in terms of its data rate y,

$$u(y) = \frac{y^{1-\alpha}}{1-\alpha},$$

where the utility parameter $\alpha \in (0, 1)$ (78). If a source user i belonging to network j starts a session with the data rate y_{ij}, then it pays network j the *user pricing* $\pi_j y_{ij}$.[16] The source

[16] We assume that the destination node does not need to pay the user pricing. This is similar as the pricing of multimedia messaging services in may countries today.

user i's payoff (which is also the payoff of the communication session) is

$$r_i(y_{ij}, \pi_j) = \frac{y_{ij}^{1-\alpha}}{1-\alpha} - \pi_j y_{ij}.$$

The optimal *demand* of data rate that maximizes utility is

$$y_{ij}^*(\pi_j) = \pi_j^{-1/\alpha},$$

which has a constant elasticity of $1/\alpha$. A small α denotes an elastic application and a large α denotes an inelastic application. Source user i's optimal payoff is

$$r_i(y_{ij}^*(\pi_j), \pi_j) = \frac{\pi_j^{1-1/\alpha}}{1-\alpha} - \pi_j^{1-\frac{1}{\alpha}},$$

which only depends on the network price π_j and is independent of the user's location. This is desirable in practice, as the user only needs to keep track of the total data usage instead of where he conducts the communications.

Network Costs

A communication session involves both a uplink transmission (from the source user to its network's base station) and a downlink transmission (from the destination user's base station to the destination user). Each part of the transmissions involves a cost proportional to the bandwidth consumed. For a source user i belonging to network j, the relationship between the transmission rate y_{ij} and the consumed bandwidth B_{ij} depends on the distance between the user and the base station d_{ij}, the uplink transmission power per unit bandwidth P^u, and the background noise density n_0. We assume an Orthogonal Frequency Division Multiple Access scheme with equal power allocation, and no two users (either in a same or different networks) interfere with each other. Thus

$$y_{ij} = B_{ij} \log\left(1 + \frac{P^u h_{ij}\left(d_{ij}\right)}{n_0}\right),$$

where $h_{ij}(d_{ij})$ is the channel gain depending on the distance d_{ij}. One possible choice is $h_{ij}(d_{ij}) = {d_{ij}}^{-\beta}$, where β is the channel attenuation factor (usually between 2 to 4). The analysis in this application can be generalized to other distance based channel models as we keep the function $h_{ij}(d_{ij})$ abstract most of the time. The total bandwidth cost for supporting this uplink transmission is $c_j B_{ij}$. The cost for the downlink transmission can be computed similarly, except that P^u will be replaced by P^d and c_j will be replaced by the cost of the corresponding network. For notation simplicities, we denote

$$g^d(d_{ij}) \equiv \log\left(1 + \frac{P^d h_{ij}\left(d_{ij}\right)}{n_0}\right),$$
$$g^u(d_{ij}) \equiv \log\left(1 + \frac{P^u h_{ij}\left(d_{ij}\right)}{n_0}\right).$$

Let us compute the total cost of serving one communication session. If source user i at location d_i is with network 1, then network 1's cost of the uplink transmission is:

$$c_1 B_{i1} = c_1 \frac{y_{i1}}{\log\left(1+\frac{P^u h_{ij}(d_i)}{n_0}\right)} = c_1 \frac{y_{i1}}{g^u(d_i)}.$$

Serving a user close-by (with a small value of d_i and thus a larger $g^u(d_i)$) costs less compared with serving a user far away. For the downlink traffic, since the destination may be with network 1 or network 2, we need to compute the *expected* downlink cost. In our model, users are uniformly distributed along the segment [0,1], and each of them will have equal probability of receiving data. The expected cost is $c_1 \int_0^m \frac{y_{i1}}{g^d(r)} dr$ to network 1 and $c_2 \int_0^{1-m} \frac{y_{i1}}{g^d(r)} dr$ to network 2. Hence, the total expected cost to support a session initiated by a user i located at d_i in network 1 with data rate y_{i1} is:

$$c_1 \frac{y_{i1}}{g^u(d_i)} + c_1 \int\limits_0^m \frac{y_{i1}}{g^d(r)} dr + c_2 \int\limits_0^{1-m} \frac{y_{i1}}{g^d(r)} dr.$$

The first two terms are related to network 1 and the third term is related to network 2.

In the rest of the discussions, we denote

$$B^u(m) \equiv \int_0^m \frac{1}{g^u(r)} dr \text{ and } B^d(m) \equiv \int_0^m \frac{1}{g^d(r)} dr,$$

which represent the average bandwidth needed to serve one unit of uplink and a downlink transmission of a network with a market share m, respectively.

Access Price

For an inter-network session, the network with the destination user cannot charge the source user, but bears the cost of the downlink transmission. To compensate this additional cost, the destination network charges the source network an *access pricing.* As illustrated in Fig. 6.9, to complete an inter-network session initiated from network 1, network 1 pays network 2 a_2 per unit of bandwidth consumed in network 2. The access price a_1 is defined similarly.

6.3.3 SOCIAL OPTIMAL USER PRICING

A regulator cares about the social welfare, which is the total payoffs of all entities in the market, i.e., the total user utility minus the total network cost. The payments (either from users to networks or between networks) are internal transfers and do not affect the social welfare. However, a regulator typically cannot directly control how much resources that users consume. In this section, we will look at the case where the regulator maximizes the social welfare by controlling the user pricing π_1 and π_2.[17]

[17] The access prices a_1 and a_2 cancel out and do not affect social welfare.

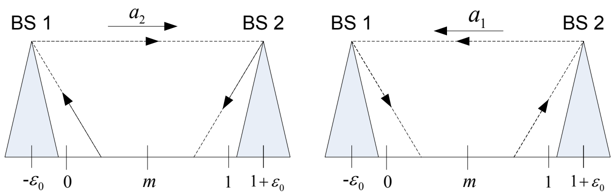

Figure 6.9: Illustration of the access pricing. Left: To complete an inter-network traffic initiated from network 1, network 2 charges network 1 a_2 per unit of bandwidth. Right: the access price a_1 is defined similarly.

The social welfare $SW(\pi_1, \pi_2, m, c_1, c_2)$ is:

$$\begin{aligned} SW(\pi_1, \pi_2, m, c_1, c_2) = m\frac{{\pi_1}^{\alpha-\frac{1}{\alpha}}}{1-\alpha} + (1-m)\frac{{\pi_2}^{\alpha-\frac{1}{\alpha}}}{1-\alpha} \\ - \pi_1^{-\frac{1}{\alpha}} f(m, c_1, c_2) - \pi_2^{-\frac{1}{\alpha}} f(1-m, c_2, c_1), \quad (6.10) \end{aligned}$$

where $f(m, c_j, c_{-j})$ represents the total cost in serving sessions originated from network j of a market share m,

$$f(m, c_j, c_{-j}) = c_j B^u(m) + mc_j B^d(m) + mc_{-j} B^d(1-m).$$

Here c_{-j} denotes the per unit bandwidth cost of the network other than network j. For example, if $j = 1$, then $c_{-j} = c_2$.

The regulator's objective is to choose π_1 and π_2 to maximize the social welfare. From (6.10), it is clear that the social welfare is decoupled across π_1 and π_2. We can also show that the social welfare is quasi-concave in both π_1 and π_2, and thus the optimal user prices can be obtained through solving the first order conditions.

Proposition 6.11 *Socially optimal user prices are* $\pi_1^S = f(m, c_1, c_2)/m$ *and* $\pi_2^S = f(1-m, c_2, c_1)/(1-m)$.

Proposition 6.11 implies that the social optimal user prices do not depend on the utility parameter α. Proposition 6.11 works for any channel function. Next, we study the impacts of market share and bandwidth costs with a particular choice of channel function of $h_{ij}(d_{ij}) = d_{ij}^{-3}$. We have two observations from the results in Fig. 6.10.

Observation 6.12 A network's social optimal price first decreases and then increases with its market share.

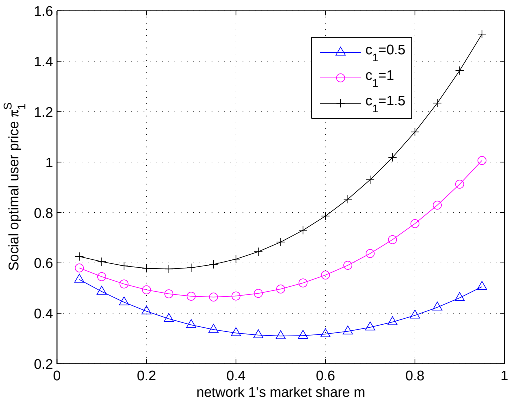

Figure 6.10: Social optimal user price π_1^S of network 1 with different market share m and cost c_1. Here $h_{ij}(d_{ij}) = d_{ij}^{-3}$, $P^d = P^u = 1$, and $c_2 = 1$.

Figure 6.10 illustrates this observation for network 1, which has a market share m. When m is small, only users close to BS1 are associated with network 1 (see Fig. 6.8). Since the average channel condition of these users to BS1 is high, the average cost of serving one unit of data rate is low. When the market share m increases, it is beneficial for the regulator to decrease the price π_1^S to induce more demands of the users, which increases users' utility with a relatively small increase in the total cost.

When the market share m is large, many users associated with network 1 are located far away from BS1, and thus the average cost of serving users is high. Any additional user joining the network due to the increase of market share requires the highest cost to serve. As a result, it is beneficial for the regulator to increase the price so as to decrease the operational cost, with a small penalty of users' utility decrease.

Comparing three curves in Figure 6.10, we have the following.

Observation 6.13 A network's social optimal price increases in its bandwidth cost.

6.3.4 SOCIAL OPTIMAL ACCESS PRICING

Very often the regulator cannot even control the user pricing in practice. This may be due to the complexity of the regulation, or due to the fact that the government just does not want to micro-manage the telecommunication industry. However, the regulator can still achieve the social optimality by setting the access pricing.

Mathematically, we can model the system as a two-stage decision process. In stage one, the regulator determines the access pricing a_1 and a_2 between two network operators.

In stage two, each network j chooses the user pricing π_j to maximize its profit, given a_1 and a_2. We will look at the stage 2 problem first.

Networks' Profit-Maximizing User Pricing Given Fixed Access Prices: $\pi_1^*(a_1, a_2)$ and $\pi_2^*(a_1, a_2)$

We begin by deriving the network profits. Consider a session originated from a user i located at d_i in network 1. Network 1's profit from this session equals the payment received from user i minus the total expected cost for either an intra-network session and an inter-network session, i.e.,

$$\pi_1^{1-\frac{1}{\alpha}} - c_1 \frac{\pi_1^{-\frac{1}{\alpha}}}{g^u(d_i)} - c_1 \int_0^m \frac{\pi_1^{-\frac{1}{\alpha}}}{g^d(r)} dr - a_2 \int_0^{1-m} \frac{\pi_1^{-\frac{1}{\alpha}}}{g^d(r)} dr.$$

For an inter-network session originating from a user i located at d_i in network 2, network 1's profit equals the access price payment received from network 2 minus the cost in supporting the downlink communication, i.e.,

$$(a_1 - c_1) \int_0^m \frac{\pi_2^{\frac{-1}{\alpha}}}{g^d(r)} dr.$$

Combining the above analysis, the profits of network 1 (R_1) and of network 2 (R_2) are

$$\begin{aligned} R_1(\pi_1, \pi_2) = m\pi_1^{1-\frac{1}{\alpha}} &- c_1\pi_1^{-\frac{1}{\alpha}} B^u(m) - mc_1\pi_1^{-\frac{1}{\alpha}} B^d(m) \\ &- ma_2\pi_1^{-\frac{1}{\alpha}} B^d(1-m) + (1-m)(a_1 - c_1)\pi_2^{-\frac{1}{\alpha}} B^d(m) \end{aligned}$$

and

$$\begin{aligned} R_2(\pi_1, \pi_2) = (1-m)\pi_2^{1-\frac{1}{\alpha}} &- c_2\pi_2^{-\frac{1}{\alpha}} B^u(1-m) - (1-m)c_2\pi_2^{-\frac{1}{\alpha}} B^d(1-m) \\ &- (1-m)a_1\pi_2^{-\frac{1}{\alpha}} B^d(m) + m(a_2 - c_2)\pi_1^{-\frac{1}{\alpha}} B^d(1-m). \end{aligned}$$

By optimizing R_1 over π_1 and optimizing R_2 over π_2, we have the following results.

Proposition 6.14 *For any given access prices a_1 and a_2 set by the regulator, networks 1 and 2 set the following user prices to maximize their individual profits,*

$$\pi_1^*(a_2) = \frac{c_1 B^u(m) + mc_1 B^d(m) + a_2 B^d(1-m)}{m(1-\alpha)}, \tag{6.11}$$

and

$$\pi_2^*(a_1) = \frac{c_2 B^u(1-m) + (1-m)c_2 B^d(1-m) + a_1 B^d(m)}{(1-m)(1-\alpha)}. \tag{6.12}$$

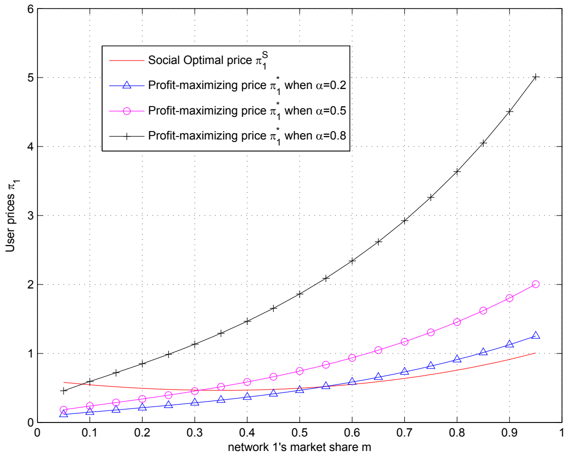

Figure 6.11: Social optimal user price π_1^S and profit-maximizing user price π_1^* with different values of market share m and utility parameter α. Here $h_{ij}(d_{ij}) = d_{ij}^{-3}$, $P^d = P^u = 1$, and $c_1 = c_2 = 1$.

Observation 6.15 A network j's profit-maximizing user pricing π_j^* depends on the rival network's access pricing a_{-j} and is independent of its own access pricing a_j.

Before studying the regulator's optimal choice of access pricing, let us consider the case where no access pricing is set, i.e., $a_1 = a_2 = 0$. We can compare networks' profit maximizing user pricing (π_1^* and π_2^*) with the social optimal user pricing (π_1^S and π_2^S) computed in Section 6.3.3.

Observation 6.16 With $a_1 = a_2 = 0$, the profit-maximizing user prices π_1^* and π_2^* always lead to a smaller social welfare comparing with the one achieved under π_1^S and π_2^S, except for one value of the market share.

Figure 6.11 shows the social optimal user price π_1^S and the profit-maximizing user price π_1^* for three different values of α under different market shares. For each choice of α, π_1^* only intersects with π_1^S once, and the prices are different for all other values of m. For example, when $\alpha = 0.2$ and market share $m < 0.55$, we have $\pi_1^* < \pi_1^S$ and users' demand is larger in the profit-maximizing case than in the social optimal case. It is the other way around when $m > 0.55$. Neither case is desirable from the regulator's point of view.

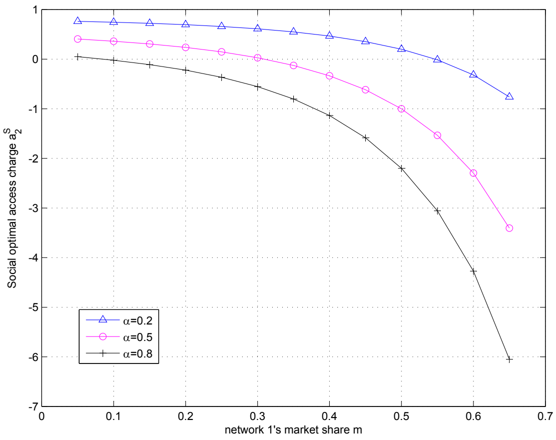

Figure 6.12: Social optimal access price a_2^S with different market share m and utility parameter α. Here $h_{ij}(d_{ij}) = d_{ij}^{-3}$, $P^d = P^u = 1$, and $c_1 = c_2 = 1$.

Social Optimal Access Pricing: a_1^S and a_2^S

Now consider the stage 1 problem. The regulator can set the proper access pricing a_1^S and a_2^S so that the networks' profit-maximizing behavior is aligned with the social optimality objective, i.e., $\pi_1^*(a_2^S) = \pi_1^S$ and $\pi_2^*(a_1^S) = \pi_2^S$.

By comparing the values of π_1^S and π_2^S in Proposition 6.11 and the values of π_1^* and π_2^* in (6.11) and (6.12), we have the following.

Proposition 6.17 *The social optimal access prices are*

$$a_1^S = \frac{(1-m)(1-\alpha)\pi_2^S - c_2 B^d(1-m)}{(1-m)B^d(1-m)} - \frac{c_2 B^d(1-m)}{B^d(m)},$$
$$a_2^S = \frac{m(1-\alpha)\pi_1^S - c_1 B^d(m)}{mB^d(m)} - \frac{c_1 B^d(m)}{B^d(1-m)},$$

where π_1^S and π_2^S are defined in Proposition 6.11.

Figure 6.12 shows the social optimal access price a_2^S with different market share of network 1 m and utility parameter α. We have the following observation.

Observation 6.18 The social optimal access price of a network decreases in its rival network's market share.

Observation 6.18 can be explained as follows. When a network's market share (e.g., m of network 1) increases, its average bandwidth cost of serving users also increases. Due to the profit-maximzing nature of the network, it tends to increase the user pricing (π_1^*) significantly to compensate such cost increase (see Fig. 6.11). Thus the access price charged by the other network (network 2) should decrease to provide incentive for the network to maintain the social optimal user pricing.

In fact, in the extreme case where the average cost has increased so much due to a very large market share, the access price from its rival network needs to be negative (i.e., network 2 pays to network 1 for using network 2's resource, as the case of $\alpha = 0.8$ and $m = 0.6$ in Fig. 6.12) to reach the social optimality. For example, assume that network 1 is a cellular network that has a large coverage area, a large market share, and an average of not very high channel condition to the users. Network 2 is a commercial Wi-Fi service provider that has a small coverage area, a small market share, and an average of excellent channel condition to the users. Then in order to maintain the social optimal user pricing, i.e., keeping the user price of the cellular network π_1^* low enough, the Wi-Fi service provider needs to pay the cellular provider for a file transfer from a cellular user to a Wi-Fi user, as the cellular network is bearing most of the network costs in supporting the communication session. In other words, the cellular network provides positive externality to the wifi network. Notice that here we consider the case where the regulator determines the access pricing; it will be a quite different story if the networks themselves optimize the access pricing to maximize their profits.

Observation 6.19 The social optimal access price increases in the elasticity of the users' utilities.

We also note from Fig. 6.12 that the utility parameter α has a significant impact on the social optimal access pricing. A smaller α (e.g., a higher elasticity) means a higher optimal access price (under the same market share). Since today's wireless networks are becoming more data-centric, we can expect that the overall users' utility functions (determined by the applications) will become more elastic and thus the optimal access price will become higher.

6.4 CHAPTER SUMMARY

This chapter discusses how we should price for negative and positive externalities in networks. We first explain the concept of externality, which reflects the side effect that is imposed by the actions of a player on a third-party not directly involved. We then illustrate that both positive and negative externalities can lead to a deviation from the social optimal solution. As an example of the negative externality, we illustrate how pollution from a chemical plan will affect the business of a water company, and how such negative externality can be internalized by the method of Pigovian tax. As an example of the positive externality, we explain the different kinds of network effects.

We then illustrate the application of theory using two applications. In the first application, we consider the problem of optimal distributed power control in wireless ad hoc networks. The mutual interferences pose negative externality among wireless users. To mitigate such negative externality, the wireless users will charge each other the interference prices, which are basically distributively computed Pigovian tax. Under proper technic a conditions, the ADP algorithm designed based on asynchronous power and price updates will converge to a global or local optimal solution, much faster than the usual gradient based method with step size. In the second application, we study the choice of wireless user pricing and access pricing from a regulator's perspective. In the case where the regulator can directly control the user pricing, we show that a network's social optimal user pricing is always increasing in the cost per unit bandwidth. When the regulator can only control the access price, we show that it is still possible to achieve the social optimality after taking the networks' profit-maximizing behaviors into consideration. The social optimal access pricing decreases with the rival network's market share and increases in the elasticity of users' utility functions. For more details especially mathematically proofs related to the two applications, please see (80; 81).

CHAPTER 7

Outlook

In this book, we often assume that the information of all commodities are observable to all market players.[1] We also assume that one player (or one type of players) in the market has the market power to determine the solution (such as price), and others can only accept or reject the solution. In reality, however, the violation of these two assumptions is often the norm instead the exception. In particular, market information is often incomplete to most market players, and the market power is often distributed among different market players.

A number of new issues arise under information asymmetry, among which the most important one is the *truthfulness* (also called *incentive compatibility*). That is, how to design a truthful (or incentive compatible) mechanism that credibly elicits the private information held by some market players. The significance of the truthful mechanism is suggested by the revelation principle (82), which states that "for any outcome resulting from any mechanism, there always exists a *payoff-equivalent* revelation mechanism where the players truthfully report their private information." The principle is extremely useful. It allows one to solve for an outcome or equilibrium by assuming all players truthfully report their private information (subject to the incentive compatibility constraint). In other words, it helps to eliminate the need to consider either strategic behavior or lying. This means that to find an optimal mechanism (to ace hive certain objective, e.g., profit maximization or social welfare maximization), we do not need to search from the enormously large set of mechanisms where players act arbitrarily, but only need to consider those truthful mechanisms where players act truthfully. Typical examples of truthful mechanisms include auction and contract.[2]

Problems become significantly different, when the market power is distributed among multiple market players. Specifically, when one player or one type of players (e.g., firms or consumers) has the total market power, the player(s) tries to extract the social surplus as much as possible. For example, in a monopoly market where the monopolist has the total market power, the monopolist can set a monopoly price or perform price discrimination to maximize his profit. In a oligopoly market where the firms have the total market power, they can set strategic prices or quantities to maximize their own profits against others' strategies. This type of self-interested interaction among players is essentially referred to as

[1] The only exception is Chapter 4, where we discussed the price differentiation under incomplete market information.

[2] Note this is not to say that all auction designs or contract designs are truthful. But rather, a considerable part (and possibly, the most important part) of auctions and contracts are truthful mechanisms.

the non-cooperative game theory. When the market power is distributed among different types of market players (e.g., firms and consumers), the self-interested interaction may not work due to the conflict interests among the players. Consider a simple example where a firm sells a single product to a consumer. How to determine the price of the product if both the firm and the consumer have certain market power? This problem cannot be solved by a self-interested interaction, since the increase of one player's surplus must lead to the decrease of the other's surplus. A well studied discipline to this type of problem is *bargaining*. Simply speaking, bargaining solution is such an outcome that both players feel *acceptable*, rather than strictly prefer in terms of certain criterion. This type of bargaining interaction is essentially referred to as the cooperative game theory.

In the rest of the chapter, we will briefly discuss the connections and differences between pricing, auction, contract, and bargaining models, and provide pointers to some key related wireless literature. We hope to provide more detailed discussions regarding the theory and applications of these different economic mechanisms in a future book.

7.1 AUCTION

An auction is a process of buying and selling goods or services by offering them up for bid, taking bids, and then selling the item to the highest bidder(s). Typical issues studied by auction theorists include the efficiency of a given auction design, optimal and equilibrium bidding strategies, and revenue comparison. There are many possible designs (or sets of rules) for an auction, among which the most important two are the *allocation rule* (who is/are the winner(s) of an auction) and the *charge rule* (what is the payment(s) of winner(s)). In this sense, an auction can be viewed as a special kind of the pricing model.

One key difference between the pricing model and the auction model is their application scenarios. Specifically, the network pricing in this book is often used in the symmetric and complete information scenario, where the decision-makers knows complete information about the commodities or the market; whereas the auction model is often used in the asymmetric information scenario where the market players (called bidders) hold certain private information that the decision-makers (called auctioneers) do not know. In a pricing model, the decision-makers determine the market price based on their known information; while in an auction model, the auctioneers let the market (i.e., the community of bidders) set the price due to the uncertainty about the bidders' valuations. With a careful design, the bidders have the incentive to bid for the commodity in a truthful manner, and the auctioneers can efficiently allocate the commodity without knowing the bidders' private valuations in advance.

Auction has been widely used in wireless networks for network resource allocation and performance optimization. In (83), Huang et al. proposed two divisible auction mechanisms for power allocation to achieve efficiency and fairness, respectively. In (84), Huang et al. extended the auction mechanisms to a cooperative communication system with multiple relays. In (85), Li et al. proposed several truthful (strategy-proof) spectrum auction mechanisms to achieve the efficiency closed to social optimal. In (86), Gandhi et al. proposed a real-time spectrum auction framework under interference constraints. In (87; 88), Zheng et al. proposed truthful single side spectrum auction and double spectrum auction, respectively, both considering spectrum reuse. In (89), Wang et al. proposed a general framework for truthful double auction for spectrum sharing. In (90), Gao et al. proposed a multi-shot spectrum auction mechanism to achieve social optimal efficiency in dynamic spectrum sharing.

7.2 CONTRACT

A contract is an agreement entered into voluntarily by two or more parties with the intention of creating a legal obligation. In economics, contract theory studies how the economic agents construct *contractual* arrangements, generally in the presence of asymmetric information. Thus, it is closely connected to the truthful (or incentive compatible) mechanism design. A standard practice in the contract theory is to represent the behavior of a decision-maker under certain numerical utility structures, and then apply an optimization algorithm to identify optimal decisions. Such a procedure has been used in the contract theory framework to several typical situations, such as *moral hazard*, *adverse selection*, *signalling* and *screening*. The common spirit of these models is to motivate one party (the agent) to act on behalf of another (the principal), e.g., to truthfully reveal information.

In moral hazard models, the information asymmetry is the principal's inability to observe and/or verify the agent's action (termed as hidden action). Contracts that depend on observable and verifiable output can often be employed to create incentives for the agent to act in the principal's interest. In adverse selection models, the principal is not informed about a certain characteristic of the agent (termed as hidden information). Two commonly used methods to model adverse selection are signaling games and screening games. The idea of signalling is that one party (usually the agent) credibly conveys some information about itself to another party (usually the principal). The idea of screening is that one party (usually the principle) offers multiple contract options, which are incentive compatible for another party (usually the agent) such that every agent selects the option indented for his type. The main difference of signalling and screening is that who moves first. In signalling games, the informed agent moves (signaling) first, and the process is essentially a Stackelberg game with the agent as the leader. In screening games, however, the uninformed principle moves (offering the options) first, and the process is essentially a Stackelberg game with the principle as the leader. In the context of monopoly market, the second degree price discrimination (in Section 4) is essentially a screening model.

Contract has also been widely used in wireless networks. In (91), Gao et al. proposed a quality-price screening contract for secondary spectrum trading, where the seller offers a menu of prices for different qualities to attract different types of buyers. The authors show that with the contract, the seller can extract more surplus from the buyers with private information. In (92), Kalathil et al. proposed a contract-based spectrum sharing mechanism to avoid possible manipulating in a spectrum auction. The authors show that it is possible to achieve socially optimal rate allocations with contracts in licensed bands. In (93), Duan et al. proposed a time-power screening contract for cooperative spectrum sharing between PUs and SUs, where the SUs relay traffic for PUs in exchange of the guaranteed access time on the PUs' licensed spectrums. In such a market, the price is essentially the amount of power the SUs offer. In (94; 95), Kasbekar and Sarkar et al. considered the secondary spectrum trading with two types of contracts: the guaranteed-bandwidth contract, which provides guaranteed access to a certain amount of bandwidth for a specified duration of time, and the opportunistic-access contract, which offers restricted (uncertain) access rights on a certain amount of bandwidth at the current time slot. In (96), Gao et al. studies the secondary spectrum trading in a hybrid market with both contract users and spot purchasing users.

7.3 BARGAINING

Bargaining is a type of negotiation in which the buyer and seller of a good or service discuss the price which will be paid and the exact nature of the transaction that will take place, and eventually come to an agreement. In this sense, bargaining is an alternative pricing strategy to fixed prices. Bargaining arises when the market power is distributed among different market players (and thus no participant has the total market power to determine the solution solely). Solutions to bargaining come in two flavors: an *axiomatic approach* where desired properties of a solution are satisfied, and a *strategic approach* where the bargaining procedure is modeled in detail as a sequential game. Typical solutions of bargaining include Nash bargaining solution, Shapely value, Harsanyi value, and so on.

The study of bargaining was initiated by J. Nash in 1950 (97), who provided an axiomatic solution for the outcome of the negotiation among 2 players. In 1982, A. Rubinstein proposed a sequential non-cooperative game (named Rubinstein bargaining game) between 2 players (98), where the player alternating offers through an infinite time horizon. As one player offers a proposal, the other player decides to accept or reject. If a proposal is rejected (by the responder), the proposer and responder change their roles, that is, the previous responder becomes a proposer offering a proposal, and the previous proposer become a responder deciding to accept or reject the proposal. Rubinstein characterized the subgame perfect equilibrium of this game, and concluded that the subgame perfect equilibrium of this non-cooperative Rubinstein bargaining game is equivalent to the Nash bargaining solution given by the axiomatic approach. This to some extent connects the axiomatic approach and the non-cooperative strategic approach for bargaining models.

Although many studies consider that players bargain independently and in an uncoordinated fashion, a survey of recent economic journals reveals that most applied bargaining papers actually analyze group bargaining problems (99). That is, more often than not, players form groups and bargain jointly in order to improve their anticipated payoff. Examples include labor disputes between the management which represents the stockholders of a factory, and a union which represents the workers (100; 101). In order to predict the bargaining result in such settings, it is necessary to analyze both the inter-group bargaining as well as the intra-group bargaining. Usually, bargaining first takes place among the different groups and accordingly the members of each group bargain with each other in order to distribute the acquired welfare. In most cases, the grouping improves the payoff of the group members (102; 103; 104), since it leverages their bargaining power. Moreover, often the bargaining outcome depends on the bargaining protocol, i.e. bargaining concurrently or sequentially (and the sequence the players bargain). This aspect was studied in (105; 106), where one dominant player optimally selects in each stage a weak player to bargain.

Bargaining has also been widely used in wireless networks. In (107), Zhang et al. proposed a cooperation bandwidth allocation strategy based on the Nash bargaining solution in a wireless cooperative relaying network. In (108), Cao et al. proposed a local bargaining approach to fair spectrum allocation in mobile ad-hoc networks. In (109), Han et al. proposed a fair scheme to allocate subcarrier, rate, and power for multiuser OFDMA systems based on Nash bargaining solutions and coalitions. The above works studied the bargaining problem using axiomatic approaches. In (110), Yan et al. studied the bargaining problem using strategic approaches. Specifically, they considered dynamic bargaining between one primary user and several secondary users in a cognitive cooperative network with incomplete network information. Moreover, in (111), Boche et al. studied the necessary requirements for the existence and uniqueness of Nash bargaining solution and proportional fairness solution, and showed that the classical requirement of compact comprehensive convex utility set can be generalized to certain strictly log-convex and noncompact sets.

Bibliography

[1] G. Staple and K. Werbach, “The end of spectrum scarcity,” *IEEE Spectrum*, pp. 48–52, March 2004.

[2] M. McHenry, “NSF spectrum occupancy measurements project summary,” *Shared Spectrum Company*, 2005.

[3] P. Bahl, R. Chandra, T. Moscibroda, R. Murty, and M. Welsh, “White space networking with wi-fi like connectivity,” in *ACM SIGCOMM Computer Communication Review*, vol. 39, no. 4. ACM, 2009, pp. 27–38.

[4] A. Technica, “Broadcasters sue FCC over white space broadband decision.” [Online]. Available: http://arstechnica.com/tech-policy/news/2009/03/broadcasters-sue-fcc-over-white-space-broadband-decision.ars

[5] T. Nguyen, H. Zhou, R. Berry, M. Honig, and R. Vohra, “The impact of additional unlicensed spectrum on wireless services competition,” in *New Frontiers in Dynamic Spectrum Access Networks (DySPAN), 2011 IEEE Symposium on*. IEEE, 2011, pp. 146–155.

[6] Cisco, “Cisco visual networking index: Global mobile data traffic forecast update, 2011–2016,” White Paper, 2012.

[7] L. Duan, J. Huang, and J. Walrand, “Economic analysis of 4g network upgrade,” submitted to *IEEE Journal on Selected Areas in Communications*, 2011.

[8] B. P. Pashigian, *Price Theory and Applications*. McGraw-Hill Inc., 1995.

[9] S. Landsburg, *Price Theory and Applications*. South-Western Pub, 2010.

[10] A. Mas-Colell, M. Whinston, and J. Green, *Microeconomic theory*. Oxford university press New York, 1995.

[11] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge Univ Pr, 2004.

[12] D. Bertsekas, *Nonlinear programming*. Athena Scientific, 1999.

[13] S. Wright, *Primal-dual interior-point methods*. Society for Industrial Mathematics, 1997, vol. 54.

[14] R. Duffin, E. Peterson, and C. Zener, *Geometric programming: theory and application.* Wiley New York, 1967.

[15] M. Chiang, *Geometric programming for communication systems.* Now Publishers Inc, 2005.

[16] K. Arrow, L. Hurwicz, and H. Uzawa, "Constraint qualifications in maximization problems," *Naval Research Logistics Quarterly*, vol. 8, no. 2, pp. 175–191, 1961.

[17] H. Kuhn and A. Tucker, "Nonlinear programming," in *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, vol. 1. California, 1951, pp. 481–492.

[18] "Mobile station-base station compatibility standard for dual-mode wideband spread spectrum cellular system, tia/eia interim standard 95(is-95-a), washington, dc: Telecommunications industry association," May 1995.

[19] "TIA/EIA IS-856 CDMA 2000: High Rate Packet Data Air Interface Specification," Nov 2000.

[20] J. Ohm, "Advances in Scalable Video Coding," *Proceedings of the IEEE*, vol. 93, no. 1, pp. 42–56, 2005.

[21] J. Xin, C. Lin, and M. Sun, "Digital video transcoding," *Proceedings of the IEEE*, vol. 93, no. 1, pp. 84–97, 2005.

[22] A. Vetro and C. Sun, "Video transcoding architectures and techniques: an overview," *Signal Processing Magazine, IEEE*, vol. 20, no. 2, pp. 18–29, 2003.

[23] Z. Li, F. Zhai, and A. Katsaggelos, "Video summarization for energy efficient wireless streaming," *Optical Fibers: Technology. Edited by Rayss, Jan; Culshaw, Brian; Mignani, Anna G. Proceedings of the SPIE,*, vol. 5960, pp. 763–774, 2005.

[24] T. M. Cover and J. Thomas, *Elements of Information Theory.* Wiley-Interscience, 1991.

[25] R. Srikant, *The Mathematics of Internet Congestion Control.* Birkhauser Boston, 2004.

[26] J. Huang, Z. Li, M. Chiang, and A. K. Katsaggelos, "Joint source adaptation and resource pricing for multi-user wireless video streaming," *IEEE Transations on Circuits and Systems for Video Technology*, vol. 18, no. 5, pp. 582–595, May 2008.

[27] A. Sampath, P. S. Kumar, and J. Holtzman, "Power control and resource management for multimedia CDMA wireless system," in *Proc. IEEE PIMRC'95*, vol. 1, 1995, pp. 91–95, centralized resource allocation.

[28] K. Kumaran and L. Qian, "Uplink Scheduling in CDMA Packet-Data Systems," *Wireless Networks*, vol. 12, no. 1, pp. 33–43, 2006.

[29] P. Marbach, "Analysis of a static pricing scheme for priority services," *IEEE/ACM Transactions on Networking (TON)*, vol. 12, no. 2, pp. 312–325, 2004.

[30] J. Huang, R. Berry, and M. L. Honig, "Auction-based spectrum sharing," *Mobile Networks and Applications*, vol. 11, no. 3, pp. 405–418, Jun 2006.

[31] F. Kelly, A. Maulloo, and D. Tan, "Rate control for communication networks: shadow prices, proportional fairness and stability," *Journal of the Operational Research society*, vol. 49, no. 3, pp. 237–252, 1998.

[32] X. Lin and N. B. Shroff, "Utility maximization for communication networks with multi-path routing," submitted to *IEEE Transactions on Automatic Control*, 2004.

[33] T. Voice, "Stability of congestion control algorithms with multi-path routing and linear stochastic modelling of congestion control," Ph.D. dissertation, Ph. D. dissertation, University of Cambridge, Cambridge, UK, 2006.

[34] V. Gajic, J. Huang, and B. Rimoldi, "Competition of wireless providers for atomic users," *arXiv:1007.1087v1*, 2010.

[35] D. Bertsekas, *Nonlinear Programming*, 2nd ed. Belmont, Massachusetts: Athena Scientific, 1999.

[36] M. Chen and J. Huang, "Optimal resource allocation for OFDM uplink communication: A primal-dual approach," in *Conference on Information Sciences and Systems*, Princeton University, NJ, USA, March 2008, pp. 926–931.

[37] H. Khalil and J. Grizzle, *Nonlinear systems.* Macmillan Publishing Company New York, 1992.

[38] S. Boyd and L. Vanderberghe, *Convex Optimization.* Cambridge University Press, 2004.

[39] B. Pashigian, *Price theory and applications.* McGraw-Hill, 1998.

[40] S. Shenker, "Fundamental design issues for the future internet," *IEEE Journal of Selected Areas in Communications*, vol. 13, pp. 1176–1188, 1995.

[41] M. Yuksel and S. Kalyanaraman, "Elasticity considerations for optimal pricing of networks," in *Proceedings of IEEE International Symposium on Computer Communications (ISCC)*, Antalya, Turkey, 2003.

[42] P. Marbach and R. Berry, “Downlink resource allocation and pricing for wireless networks,” in *IEEE INFOCOM*, New York, NY, 2002.

[43] H. R. Varian, *Microeconomic Analysis*, 3rd ed. New York: W. W. Norton & Company, 1992.

[44] T. Basar and R. Srikant, “Revenue-maximizing pricing and capacity expansion in a many-users regime,” in *IEEE INFOCOM*, vol. 1, 2002, pp. 294–301.

[45] V. Valancius, C. Lumezanu, N. Feamster, R. Johari, and V. Vazirani, “How many tiers? pricing in the internet transit market,” *SIGCOMM-Computer Communication Review*, vol. 41, no. 4, p. 194, 2011.

[46] J. van Lint and R. Wilson, *A course in combinatorics*. Cambridge Univ Pr, 2001.

[47] J. Huang and X. Huang, “Revenue management for cognitive spectrum underlay networks: An interference elasticity perspective,” in *Asia-Pacific Conference on Communications*, Shanghai, China, October 2009.

[48] S. Li and J. Huang, “Revenue maximization for communication networks with usage-based pricing,” *under revision of IEEE Transactions on Networking*, 2012.

[49] M. Osborne and A. Rubinstein, *A course in game theory*. The MIT press, 1994.

[50] R. Gibbons, *A primer in game theory*. FT Prentice Hall, 1992.

[51] J. Friedman, *Game theory with applications to economics*. Oxford University Press New York, 1986.

[52] L. Petrosjan and V. Mazalov, *Game theory and applications*. Nova Science Pub Inc, 2002, vol. 8.

[53] S. Kakutani, “A generalization of brouwer’s fixed point theorem,” *Duke Mathematical Journal*, vol. 8, no. 3, pp. 457–459, 1941.

[54] E. Maskin and J. Tirole, “Markov perfect equilibrium: I. observable actions,” *Journal of Economic Theory*, vol. 100, no. 2, pp. 191–219, 2001.

[55] J. Friedman, *Oligopoly theory*. Cambridge Univ Pr, 1983.

[56] Y. Narahari, D. Garg, R. Narayanam, and H. Prakash, *Game theoretic problems in network economics and mechanism design solutions*. Springer, 2009.

[57] D. Fudenberg and J. Tirole, *Game Theory*. MIT Press, 1991.

[58] S. Jayaweera and T. Li, "Dynamic spectrum leasing in cognitive radio networks via primary-secondary user power control games," *IEEE Trans. Wireless Commun*, vol. 8, no. 6, 2009.

[59] Q. Zhao and B. Sadler, "A survey of dynamic spectrum access," *IEEE Signal Processing Magazine*, vol. 24, no. 3, pp. 79 – 89, May 2007.

[60] O. Simeone, I. Stanojev, S. Savazzi, Y. Bar-Ness, U. Spagnolini, and R. Pickholtz, "Spectrum leasing to cooperating secondary ad hoc networks," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 1, pp. 203 – 213, Jan 2008.

[61] J. Jia and Q. Zhang, "Competitions and dynamics of duopoly wireless service providers in dynamic spectrum market," in *Proceedings of the 9th ACM international symposium on Mobile ad hoc networking and computing.* ACM, 2008, pp. 313–322.

[62] R. Dewenter and J. Haucap, "Incentives to license virtual mobile network operators (mvnos)," *Access pricing: Theory and practice*, pp. 305–325, 2006.

[63] R. Gibbens, R. Mason, and R. Steinberg, "Internet service classes under competition," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 12, pp. 2490–2498, 2000.

[64] L. Duan, J. Huang, and B. Shou, "Duopoly competition in dynamic spectrum leasing and pricing," *IEEE Transactions on Mobile Computing*, forthcoming.

[65] R. Cornes and T. Sandler, *The theory of externalities, public goods, and club goods.* Cambridge Univ Pr, 1996.

[66] L. Blume and D. Easley, *The new Palgrave dictionary of economics.* Palgram Macmillan, 2008.

[67] M. Katz and C. Shapiro, "Network externalities, competition, and compatibility," *The American economic review*, vol. 75, no. 3, pp. 424–440, 1985.

[68] S. Liebowitz and S. Margolis, "Network externality: An uncommon tragedy," *The Journal of Economic Perspectives*, vol. 8, no. 2, pp. 133–150, 1994.

[69] A. Pigou, *The economics of welfare.* Transaction Publishers, 1952.

[70] R. Coase, *The problem of social cost.* Wiley Online Library, 1960.

[71] J. Rohlfs, "A theory of interdependent demand for a communications service," *The Bell Journal of Economics and Management Science*, vol. , pp. 16–37, 1974.

[72] J. Huang, R. Berry, and M. L. Honig, "Performance of distributed utility-based power control for wireless ad hoc networks," in *IEEE Military Communications Conference*, Atlantic City, NJ, USA, October 2005.

[73] Y. Qiu and P. Marbach, "Bandwidth allocation in ad hoc networks: A price-based approach," in *IEEE INFOCOM*, 2003.

[74] D. M. Topkis, *Supermodularity and Complementarity.* Princeton University Press, 1998.

[75] E. Altman and Z. Altman, "S-modular games and power control in wireless networks," *IEEE Trans. on Automatic Control*, vol. 48, no. 5, pp. 839–842, May 2003.

[76] P. Milgrom and J. Roberts, "Rationalizability, learning and equilibrium in games with strategic complementarities," *Econometrica*, vol. 58, no. 6, pp. 1255–1277, 1990.

[77] M. Chiang, "Balancing transport and physical layers in wireless multihop networks: Jointly optimal congestion control and power control," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 1, pp. 104– 116, Jan 2005.

[78] J. Laffont, P. Rey, and J. Tirole, "Network competition: I. overview and nondiscriminatory pricing," *RAND Journal of Economics*, vol. 29, no. 1, pp. 1–37, 1998.

[79] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Transactions on Networking*, vol. 8, no. 5, pp. 556 – 567, October 2000.

[80] J. Huang, R. Berry, and M. L. Honig, "Distributed interference compensation for wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 5, pp. 1074–1084, May 2006.

[81] K. F. Leung and J. Huang, "Regulating wireless access pricing," in *IEEE ICC*, Kyoto, Japan, June 2011.

[82] R. Myerson, "Optimal auction design," *Mathematics of operations research*, pp. 58–73, 1981.

[83] J. Huang, R. Berry, and M. Honig, "Auction-based spectrum sharing," *Mobile Networks and Applications*, vol. 11, no. 3, pp. 405–418, 2006.

[84] J. Huang, Z. Han, M. Chiang, and H. Poor, "Auction-based resource allocation for cooperative communications," *Selected Areas in Communications, IEEE Journal on*, vol. 26, no. 7, pp. 1226–1237, 2008.

[85] X. Li, P. Xu, S. Tang, and X. Chu, "Spectrum bidding in wireless networks and related," *Computing and Combinatorics*, vol. , pp. 558–567, 2008.

[86] S. Gandhi, C. Buragohain, L. Cao, H. Zheng, and S. Suri, "A general framework for wireless spectrum auctions," in *New Frontiers in Dynamic Spectrum Access Networks, 2007. DySPAN 2007. 2nd IEEE International Symposium on.* IEEE, 2007, pp. 22–33.

[87] X. Zhou, S. Gandhi, S. Suri, and H. Zheng, "ebay in the sky: strategy-proof wireless spectrum auctions," in *Proceedings of the 14th ACM international conference on Mobile computing and networking.* ACM, 2008, pp. 2–13.

[88] X. Zhou and H. Zheng, "Trust: A general framework for truthful double spectrum auctions," in *INFOCOM 2009, IEEE.* IEEE, 2009, pp. 999–1007.

[89] S. Wang, P. Xu, X. Xu, S. Tang, X. Li, and X. Liu, "Toda: truthful online double auction for spectrum allocation in wireless networks," in *New Frontiers in Dynamic Spectrum, 2010 IEEE Symposium on.* IEEE, 2010, pp. 1–10.

[90] L. Gao, Y. Xu, and X. Wang, "Map: Multi-auctioneer progressive auction for dynamic spectrum access," *Mobile Computing, IEEE Transactions on*, vol. , no. 99, pp. 1–1, 2010.

[91] L. Gao, X. Wang, Y. Xu, and Q. Zhang, "Spectrum trading in cognitive radio networks: A contract-theoretic modeling approach," *Selected Areas in Communications, IEEE Journal on*, vol. 29, no. 4, pp. 843–855, 2011.

[92] D. Kalathil and R. Jain, "Spectrum sharing through contracts," in *IEEE Dynamic Spectrum Access Networks (DySPAN)*, Singapore, April 2010.

[93] L. Duan, L. Gao, and J. Huang, "Contract-based cooperative spectrum sharing," in *New Frontiers in Dynamic Spectrum Access Networks (DySPAN), 2011 IEEE Symposium on.* IEEE, 2011, pp. 399–407.

[94] G. Kasbekar, S. Sarkar, K. Kar, P. Muthusamy, and A. Gupta, "Dynamic contract trading in spectrum markets," in *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on.* IEEE, 2010, pp. 791–799.

[95] P. Muthuswamy, K. Kar, A. Gupta, S. Sarkar, and G. Kasbekar, "Portfolio optimization in secondary spectrum markets," in *Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), 2011 International Symposium on.* IEEE, 2011, pp. 249–256.

[96] L. Gao, J. Huang, Y. Chen, and B. Shou, "Contrauction: An integrated contract and auction design for dynamic spectrum sharing," in *46th Conference on Information Sciences and Systems (CISS),*, 2012.

[97] J. Nash Jr, "The bargaining problem," *Econometrica: Journal of the Econometric Society*, pp. 155–162, 1950.

[98] A. Rubinstein, “Perfect equilibrium in a bargaining model,” *Econometrica: Journal of the Econometric Society*, pp. 97–109, 1982.

[99] S. Chae and P. Heidhues, “A group bargaining solution,” *Mathematical Social Sciences*, vol. 48, no. 1, pp. 37–53, 2004.

[100] Y. Chica and M. P. Espinosa, “Efficient union formation and bargaining rules in the labor market,” *European Journal of Social Sciences*, vol. 8, 2009.

[101] R. I. L. S. Dobbelaere, “Collective bargaining under non-binding contracts,” Tinbergen Institute Discussion Paper, Tech. Rep., 2011.

[102] S. Chae and H. Moulin, “Bargaining among groups: an axiomatic viewpoint,” *Working Papers*, vol. , p. , 2004.

[103] J. Vidal-Puga, *Game Theory and Information.* EconWPA, 2005.

[104] S. Chae and P. Heidhues, “Buyers’ alliances for bargaining power,” *Journal of Economics & Management Strategy*, vol. 13, no. 4, pp. 731–754, 2004.

[105] S. Moresi, S. Salop, and Y. Sarafidis, “A model of ordered bargaining with applications,” Working paper, Tech. Rep., 2008.

[106] D. Li, “One-to-many bargaining with endogenous protocol,” Working paper, Tech. Rep., 2010.

[107] Z. Zhang, J. Shi, H. Chen, M. Guizani, and P. Qiu, “A cooperation strategy based on nash bargaining solution in cooperative relay networks,” *Vehicular Technology, IEEE Transactions on*, vol. 57, no. 4, pp. 2570–2577, 2008.

[108] L. Cao and H. Zheng, “Distributed spectrum allocation via local bargaining,” in *Proc. IEEE SECON*, vol. 5, 2005, pp. 119–127.

[109] Z. Han, Z. Ji, and K. Liu, “Fair multiuser channel allocation for ofdma networks using nash bargaining solutions and coalitions,” *Communications, IEEE Transactions on*, vol. 53, no. 8, pp. 1366–1376, 2005.

[110] Y. Yang, J. Huang, and J. Wang, “Dynamic bargaining for relay-based cooperative spectrum sharing,” *IEEE Journal on Selected Areas in Communications*, 2012.

[111] H. Boche and M. Schubert, “Nash bargaining and proportional fairness for wireless systems,” *Networking, IEEE/ACM Transactions on*, vol. 17, no. 5, pp. 1453–1466, 2009.

Author's Biography

JIANWEI HUANG

Jianwei Huang is an Assistant Professor in the Department of Information Engineering at the Chinese University of Hong Kong. He received B.S. in Electrical Engineering from Southeast University (Nanjing, Jiangsu, China) in 2000, M.S. and Ph.D. in Electrical and Computer Engineering from Northwestern University (Evanston, IL, USA) in 2003 and 2005, respectively. He worked as a Postdoc Research Associate in the Department of Electrical Engineering at Princeton University during 2005-2007. He was a visiting scholar in the School of Computer and Communication Sciences at Ecole Polytechnique Federale De Lausanne (EPFL) during the Summer Research Institute in June 2009, and a visiting scholar in the Department of Electrical Engineering and Computer Sciences at University of California-Berkeley in August 2010. He is a Guest Professor of Nanjing University of Posts and Telecommunications.

Dr. Huang currently leads the Network Communications and Economics Lab (ncel.ie.cuhk.edu.hk), with the main research focus on nonlinear optimization and game theoretical analysis of communication networks, especially on network economics, cognitive radio networks, and smart grid. He is the recipient of the IEEE Marconi Prize Paper Award in Wireless Communications in 2011, the International Conference on Wireless Internet Best Paper Award 2011, the IEEE GLOBECOM Best Paper Award in 2010, the IEEE ComSoc Asia-Pacific Outstanding Young Researcher Award in 2009, and Asia-Pacific Conference on Communications Best Paper Award in 2009.

Dr. Huang has served as Editor of *IEEE Journal on Selected Areas in Communications - Cognitive Radio Series*, Editor of *IEEE Transactions on Wireless Communications*, Guest Editor of *IEEE Journal on Selected Areas in Communications* special issue on "Economics of Communication Networks and Systems", Lead Guest Editor of *IEEE Journal of Selected Areas in Communications* special issue on "Game Theory in Communication Systems", Lead Guest Editor of *IEEE Communications Magazine* Feature Topic on "Communications Network Economics", and Guest Editor of several other journals including *(Wiley) Wireless Communications and Mobile Computing*, *Journal of Advances in Multimedia*, and *Journal of Communications*.

Dr. Huang has served as Chair of IEEE MMTC (Multimedia Communications Technical Committee), the TPC Co-Chair of IEEE GLOBEBOM Selected Areas in Communications Symposium (Game Theory for Communications Track) 2013, the TPC Co-Chair of IEEE WiOpt (International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks) 2012, the Publicity Co-Chair of IEEE Communications Theory Workshop 2012, the TPC Co-Chair of IEEE ICCC Communication Theory and Security Symposium 2012, the Student Activities Co-Chair of IEEE WiOpt 2011, the TPC Co-Chair of IEEE GlOBECOM Wireless Communications Symposium 2010, the TPC Co-Chair of IWCMC (the International Wireless Communications and Mobile Computing) Mobile Computing Symposium 2010, and the TPC Co-Chair of GameNets (the International Conference on Game Theory for Networks) 2009. He is also TPC member of leading conferences such as INFOCOM, MobiHoc, ICC, GLBOECOM, DySPAN, WiOpt, NetEcon, and WCNC. He is a senior member of IEEE.

LIN GAO

Lin Gao is a Postdoc research associate in the Department of Information Engineering at the Chinese University of Hong Kong. He received the B.S. degree in Information Engineering from Nanjing University of Posts and Telecommunications in 2002, and the M.S. and Ph.D. degrees in Electronic Engineering from Shanghai Jiao Tong University in 2006 and 2010, respectively. His research interests are in the area of wireless communications and communication theory, in particular, MIMO and OFDM techniques, cooperative communications, multi-hop relay networks, cognitive radio networks, wireless resource allocation, network economics and game theoretical models.