How Markets Work

Lecture Notes on
Microeconomics and Industrial Organization

Johan Stennek

January 20, 2017

Contents

I. Why markets? 7

1. Exchange 8

1.1. Love of variety & returns to specialization 8

1.2. Absolute advantage 11

1.3. Comparative advantage 14

2. Prices 18

2.1. Problem 1: Coordination 18

2.2. Prices 22

3. Money 30

3.1. Problem 2: Double coincidence of wants 30

3.2. Commodity Money 30

3.3. Fiat money 34

4. Aims of this book 38

II. Basic Market Forms 43

5. Monopoly 44

5.1. Why study monopoly? 44

5.2. What is monopoly? 47

5.3. The monopoly model 49

5.4. Monopolist's choice 52

5.5. The determinants of price 63

5.6. Welfare 71
5.7. Third-degree price discrimination 76
5.8. Pros and cons of price regulation 80
5.9. Case Study: Value-Based pricing of pharmaceuticals 84
5.10. Exercises 88

6. Oligopoly **91**
6.1. Why study oligopoly? 91
6.2. Bertrand Oligopoly 93
6.3. What is price competition 100
6.4. Cournot oligopoly 102
6.5. The role of market concentration 111
6.6. Exercises 114

7. Bargaining **116**
7.1. Why study bargaining? 116
7.2. An extensive form bargaining game 123
7.3. Nash Bargaining Solution 135
7.4. Bilateral market power 139
7.5. Summary 149

8. Auctions **151**
8.1. Why study auctions? 151
8.2. Auction Designs and Types of Uncertainty 153
8.3. How to predict bidding behavior 155
8.4. Second-Price Sealed-Bid Auction 159
8.5. First-Price Sealed-Bid Auction 161
8.6. Comparison of Auction Designs 165

9. Quality Uncertainty and the Market Mechanism **168**
9.1. Why study quality uncertainty? 168
9.2. Bayes's Rule 172
9.3. Market for Lemons 178
9.4. Applications 187

III. Limits to Price Competition **189**

10. Sources of market power **190**

11. Product differentiation **193**
11.1. Horizontal Differentiation 196
11.2. Vertical Differentiation 209

12. Capacity Constraints **219**
12.1. Monopoly with capacity constraints 220
12.2. Oligopoly with capacity constraints 223

IV. Measuring competition **226**

13. Structure-Performance-Conduct **228**
13.1. Methodology 229
13.2. Results 230
13.3. Critique 231

14. New Empirical Industrial Organization **234**
14.1. Basic Ideas of NEIO 234
14.2. Inferring conduct from data on prices and quantities 241
14.3. Inferring conduct from data on concentration and market size 250

V. Non-Price Competition **265**

15. Product choice **266**
15.1. Monopoly 266
15.2. Oligopoly & Horizontal differentiation 274
15.3. Oligopoly & Vertical differentiation 288

16. Entry and market concentration **290**
16.1. Exogenous entry costs 291

16.2. Entry and Welfare 299
16.3. Additional aspects of the entry process 303
16.4. Endogenous Entry Costs 304
16.5. Solutions to exercises 309

VI. Advanced Commercial Strategies 315

17. Two-part tariffs 316
17.1. Monopoly, revisited 316

18. Second Degree Price Discrimination 320
18.1. Setting up a Simple Example: Cost and Demand . 322
18.2. Exogenous Segmentation Revisited 324
18.3. Endogenous Segmentation 328
18.4. More Examples of Screening 338

VII. Competition Policy 340

19. What is a market? 341
19.1. Substitution 343
19.2. Dimensions of market definition 346
19.3. Topics 350

20. Cartels 357
20.1. How a cartel can enforce its agreement 359
20.2. Factors Facilitating Cartels 368
20.3. Transparency, revisited 375
20.4. Fighting Cartels 381

21. Horizontal mergers 385
21.1. Puzzles and Controversies 385
21.2. Effects of Mergers 387
21.3. Merger Formation 410

VIIIGame Theory **427**

22.Static Games **430**

22.1. Game 1: Prisoners' Dilemma 430
22.2. Normal Forms 437
22.3. Game 2: Coordination game 442
22.4. Game 3: Chicken 446
22.5. Game 4: Stag hunt 448
22.6. Game 5: Football penalty game 450
22.7. Basic idea of an equilibrium 462
22.8. Summary 466
22.9. Additional Exercises 467

23.Dynamic Games **471**

23.1. Game 1: War & Peace 472
23.2. Game 2: Entry Deterrence 480
23.3. Summary 487
23.4. Exercises 490

Part I.

Why markets?

1. Exchange

1.1. Love of variety & returns to specialization

Yesterday, when you ate your dinner, the meat may have come from Argentina, the green beans from Kenya and a farmer in Vietnam may have produced the rice. The farmer probably planted the seedlings already last year and it may have been his son who harvested the crop once it was ripe. Someone packed the rice in plastic sacks and shipped it to Sweden. It was delivered by truck to your local store, where it was shelved for you to pick it up. Also the packaging, the ship and the truck have their history. While the truck may have been produced by German workers fairly recently, the German engineers may have done their job several years back. The steel for the ship and the truck may of course have a more local origin, in the northern parts of Sweden.

While we take all of this for granted today, it is actually quite amazing if we pause for a while and think about it:

> *How could all these complete strangers - working at various times and places - cooperate to produce that dinner of yours?*

The tension between variety and specialization...

Both as individuals, families and countries we must choose what to produce and what to consume. These choices are subject to a troublesome tension. On the one hand, we **love variety**: we wish to consume a large variety of products and services such as food, cloths, and haircuts and even a large variety of different foodstuffs including fish, meats and vegetables. On the other hand, there are enormous **gains from specialization**: most people become much more productive if they specialize in producing a single good, which they are relatively well suited for. In fact, most people specialize in a few tasks in the production of a single good. Some people may be physically strong others may have an innate gift for handling cattle. Some countries have an abundance of sun and fertile soil while others have long cost lines and rich fishing waters. With specialization follows repetition and thus knowledge, improving productivity even further. Recall Henry Ford's invention of the moving assembly line in 1913. The building of a Model T was divided into 84 steps. The production time was cut from 12 hours to around 90 minutes.

...is eased through exchange...

This tension between our love of variety and the gains from specialization is eased through **exchange**. We do not consume what we produce and we do not produce what we consume. Instead we trade. If Anderson can produce one kilo of apples in an hour but only half a kilo of pears, while Peterson can produce a kilo of pears but only half a kilo of apples, an exchange of apples for pears will allow them both to specialize in their respective areas of expertise while maintaining a mixed diet at the same time.

. . . which requires coordination. . .

But specialization and exchange does not arise all by itself. Anderson needs to know that it is apples that he is relatively suited for. He needs to know how many kilos of apples to produce, when and where to deliver them and to whom. And Anderson must also know what to require in return for these apples. Exchange calls for **coordination**.

If everyone would live alone on isolated islands, eating what they produce themselves, there would be no need for coordination. Robinson Crusoe did not have any difficulties in deciding how much corn and rice to grow or how much grapes to dry into raisins for the winter months. He knew what he liked and he quickly learned his proficiency in different activities. Also within a family it is often easy to divide the work and share the yields. Everyone knows every one else's strengths and weaknesses as well as what they like and not. It is easy to talk to one another face to face and to find a joint decision.

But, again, how did all those complete strangers, scattered around the globe, contributing at different times, align to provide the dinner you had yesterday?

. . . achieved by institutions such as markets

The core of economics is the study of exchange and how coordination may come about through various institutions such as firms, markets and money as well as even more basic institutions such as contracts and ownership. In this book we will primarily study the role of markets and take firms, contracts and other institutions as given.

While the word market used to mean a specific place where people met to exchange their goods at certain times, it is used in a slightly more abstract sense today. The rice market, for example, can be thought of as the collection of all rice transactions, possibly

restricted by some geographical boundaries and interval of time.

Markets are different. They work by many different procedures, usually referred to as **market forms**. In most consumer markets in Sweden exchange typically occur at terms decided by the seller only. These terms are decided in competition with other sellers. In intermediate goods markets, exchange more often occurs through bargaining and auctions. There are also more "institutionalized" exchanges for trading commodities and financial instruments where brokers simultaneously offer bid and ask prices according to well-defined rules.

The market outcome—who sells what to whom and at what price, and by implication, who produces what and who consumes what—clearly depends on people's preferences for different goods and services and the available technologies for producing them. But as we will see the market outcome also depends on many fine details of how markets work, such as the exact rules for bidding in a an auction, what information different buyers and sellers have, how many competitors there are, and so on.

1.2. Absolute advantage

But let us start from the beginning. It is quite easy to grasp that people should probably specialize in production and then exchange goods with one another. But one of the most fundamental economic problems is:

> *Under what conditions, more precisely, do people gain from specialization and exchange? And, who should then do what?*

To answer this question we will look at a simple, hypothetical, example. And, as it turns out, the scope for specialization and exchange is much wider than pure intuition might suggest.

Consider a village with two individuals, named Anderson and Peterson, who both work 1500 hours per year. Their diet consists of apples and pears only. Each person wants to consume as much fruit as possible, but variety is absolutely essential. They both wish to consume one apple for every pear and vice versa. Expressed differently, if Anderson eats five apples but only four pears, he derives no utility from the last apple. We may thus measure a person's well-being, or utility, in terms of the number of pairs of apples and pears he eats. Eating seven apples and four pears is as good as eating only four apples and four pears. Both fruit-bundles can thus be said to give a utility of four.

Anderson can produce two apples in an hour but only one pear, while Peterson can produce two pears but only one apple. These productivities are described in Table 1.1. We say that Anderson

Table 1.1.: Anderson's and Peterson's productivities, I

	Apples	Pears
Anderson	2	1
Peterson	1	2

has an **absolute advantage** in producing apples and that Peterson has an absolute advantage in producing pears, since Anderson is more productive than Peterson in producing apples and Peterson is more productive than Anderson in producing pears.

Before reading any further, please try to solve the following exercise. It is very important that you try to solve the problems as well as you can before looking at the solution which is available immediately below the exercise. (It may help to cover the solution with an empty sheet.) The main purpose of learning economics should be to learn how to analyze problems, not to learn replicating solutions. Expect to work one hour or so, before reading the solution.

Exercise 1. Absolute advantage

1. How many apples and pears would Anderson and Peterson produce and eat if they could not exchange fruit with each other?

2. How many pears would Peterson at most be prepared to give Anderson in return for one apple? How many pears would Anderson at least require from Peterson in exchange for one apple?

3. Anderson and Peterson agree (for some unexplained reason) to trade the fruit one for one. How many fruits do they trade with each other? Expressed differently, how many apples and pears will Anderson and Peterson produce and eat?

4. Compare the two outcomes.

Solution: Anderson always produces the same number of apples and pears in order to consume even pairs. He has to work 1.5 hours to produce one pair. Since he works 1500 hours, he will produce and eat 1000 apples and 1000 pears. Also Peterson will produce and eat the same number of apples and pears.

If Anderson produces an additional apple and gives it to Peterson, Anderson has to produce half of a pear less, as a result of his time constraint. But if Peterson would give Anderson half a pear in return for the apple, Anderson would still eat 1000 fruit-pairs. Thus, Anderson would demand at least half a pear in return for an apple.

If Peterson receives an apple from Anderson, Peterson may produce one apple less. Peterson can then use this freed-up time to produce two additional pears. If he gives those pears to Anderson, Peterson will still eat 1000 fruit-pears. Thus, Peterson would at most be willing to give Anderson two pears in return for one apple.

In sum, while Anderson only requires half a pear in return for an apple, Peterson is willing to give up as much as two pears. Anderson

and Peterson clearly have an opportunity to improve their well-being by collaborating with one another. By collaboration I mean specialization and exchange. It is not clear in what proportions they should exchange apples for pears. There is a range of possible agreements that would benefit both, but trading one for one appears to be a natural compromise. (As discussed later in the chapter on bargaining, this is also what economic analysis predicts, at least under certain conditions.)

If Anderson and Peterson specialize completely, Anderson can spend all his time producing apples, resulting in a total of 3000 apples, and Peterson can spend all his time producing pears, resulting in a total of 3000 pears. Exchanging 1500 apples and pears, both people will be able to eat 1500 fruit-pairs. Both people thus improve their utility by specializing in the production where they are more productive than their neighbor. They both benefit from eating 50% more fruit.◀

When the two farmers specialize and exchange goods with each other, the gross domestic product, GDP, of the village increases by 50%. This growth is achieved not as a result of technical progress or as a result of investments in machines and equipment, but rather as a result of a better organization of production.

1.3. Comparative advantage

Anderson suddenly becomes more productive in both apple and pear production and, as a result, he is now more productive than Peterson in both apple and pear production. Let's say that Anderson can produce eight apples or four pears in an hour. We say that Anderson has an absolute advantage in both fruits. When Anderson is more productive in producing both fruits, it is no longer obvious that the two neighbors could gain anything from cooperating with each other. Why should Anderson ask Peterson to produce pears

Table 1.2.: Anderson's and Peterson's productivities, I

	Apples	Pears
Anderson	8	4
Peterson	1	2

for him, when he is much more productive himself? Surprisingly, the answer is almost the same as in the previous case. Before reading any further, please try to solve the following exercise. Again, use an empty sheet to avoid ogling at the solution.

Exercise 2. Comparative advantage. Is there any reason for Anderson, who has an absolute advantage in the production of both activities, to take part in an exchange with Peterson? Hint: Answer the same four questions as above.

Solution: Anderson has to work 1.5 hours to produce four apple-pear-pairs. Absent trade, he will thus produce 4000 (= 4 * 1500 / 1.5) apples and pears. Peterson produces 1000 apples and pears.

If Anderson produces an additional apple and gives it to Peterson, Anderson has to produce half a pear less. But if Peterson would give Anderson half a pear in return for the apple, Anderson would still eat 4000 fruit-pairs. Thus, Anderson would demand at least half a pear in return for an apple.

If Peterson receives an apple from Anderson, Peterson may produce one apple less and instead produce two additional pears. If he gives those pears to Anderson, Peterson will still eat 1000 fruit-pears. Thus, Peterson would at most be willing to give Anderson two pears in return for one apple.

Thus, assume that Anderson produces one pear less. He will then be able to produce two extra apples. If Peterson at the same time produces one pear more and one apple less, they can exchange an apple for a pear. Both Anderson and Peterson will then end up with a surplus of one fruit-pair, i.e. they will consume 4001 and 1001 pairs.

Actually they may trade much more. They can continue with these reallocations until Peterson spends all his time on producing 3000 pears. After trading half of those with Anderson, Peterson ends up with consuming 1500 apples and 1500 pears. To produce the 1500 apples for Peterson, Anderson spends 187.5 hours of work. He also spends 187.5 hours to produce 1500 apples to eat in pair with the pears he received from Peterson. He therefore has 1125 hours left. During this time he can produce another 3000 apples and pears (3000 = 4 * 1125 / 1.5). Anderson thus ends up consuming 4500 apples and pears. ◀

Even if Anderson is more productive than Peterson in terms of the number of minutes it takes to produce a pear (15 minutes vs 30), Anderson is actually less productive than Peterson in terms of the number apples given up to produce a pear (2 apples vs 1/2). And

Table 1.3.: Opportunity cost

Cost of producing a pear	Time used	Apples forgone
Anderson	15 minutes	2
Peterson	30 minutes	1/2

it is the comparison in terms of apples that is relevant. If the two neighbors don't collaborate, they will both spend their full working day producing a mix of both fruits. When they produce pears, it is apples they actually give up, not time. Expressed differently, the opportunity cost of a pear is some number of apples and Anderson's opportunity cost of producing a pear is two apples while Peterson's opportunity cost is only half an apple.

Definition 1. The **opportunity cost** of an activity (here: production of pears) is the lost benefit of the best alternative that is not chosen (here: production of apples, rather than leisure time).

Even if Peterson is less productive than Anderson, he has a **comparative advantage** in producing pears. And, by specializing in

the fruit for which they have a comparative advantage and then exchange fruit with each other, both people can gain.

Conclusion 1. Law of Comparative Advantage Two individuals (or firms, or countries) both gain from exchange, if they have different *relative* productivities (i.e. different relative resource costs) for producing the same goods, even if one individual is more productive in the production of all goods (absolute advantage).

While some people or countries are more productive than others, it is probably impossible to find examples of people (or countries) with the same relative productivities in every pair of activities. Specialization and exchange are ubiquitous.

2. Prices

While the law of comparative advantage asserts that that people would gain from specialization and exchange, it does not explain how such collaboration may come about in practice. A second fundamental economic question is therefore:

> *How do individuals know that (i) somebody else produces what they wish to consume and that (ii) somebody else wants to consume what they produce?*

Expressed differently, how do people *coordinate* their plans in practice?

2.1. Problem 1: Coordination

If Anderson and Peterson live in the same small village, they probably know one another well. They can both actually watch the other person working and figure out their relative proficiencies. And they both know the other person's preference for eating fruits in pairs. They both understand that they would gain from specialization and exchange. It is also relatively easy for them to get together and jointly work out a common plan: (i) Anderson is to produce 3000 apples, (ii) Peterson is to produce 3000 pears, (iii) Anderson is to hand over 1500 apples to Peterson, (iv) Peterson is to hand over 1500 pears to Anderson, (v) Anderson is to consume 1500 fruit pairs, (vi) Peterson is to consume 1500 fruit pairs.

It is also easy for them to continuously monitor one another, to confirm that the other guy is actually working according to their

common plan. And if he does not, they can both quickly switch and start producing fruit for their own consumption only.

But what if they do not live in the same village? What if they live far apart and only have the opportunity to meet once a year, perhaps at the road junction - known as the marketplace - half way between their dwellings? Then, even if they agree to spend the coming year to produce fruits according to the same plan as above, they cannot be sure that the other guy will actually show up at the marketplace with the promised amount of the right kind of fruit next year.

Once back home, Anderson must decide how much of each fruit he actually wants to produce during the year. For simplicity, we may assume that he only considers three choices, namely to produce only apples, to produce equally many apples and pears or to produce only pears. If Anderson produces equally many apples and pears, he will simply eat his own fruit. He will then be able to consume 1000 fruit-pairs. There is no need for him to go to the marketplace. If Anderson specializes in apples, he will clearly go to the marketplace, hoping to meet with Peterson to exchange fruits. But the outcome of their meeting now depends on what Peterson has chosen to produce during the year. If Peterson has produced only pears, they will be able to trade. And, Anderson will end up consuming 1500 fruit pairs. If, however, Peterson has produced equally many apples and pears, or if he has specialized in apples, there will be no trade. Anderson will then not be able to consume any fruit-pairs during the year.

All possible outcomes are summarized in Table 2.1. The three rows represent Anderson's three possible choices. The three columns represent Peterson's three choices. In total, there are nine possible outcomes, represented by the cells. The number in each cell represents the number of fruit pairs that Anderson can consume after having exchanged fruits with Peterson. For instance, if Anderson has produced 3000 apples and Peterson has produced 3000 pears,

Table 2.1.: Fruit-pairs consumed by Anderson after exchange

Anderson produces	Peterson produces		
	1500 apples 0 pears	1000 apples 1000 pears	0 apples 3000 pears
3000 apples 0 pears	0	0	1500
1000 apples 1000 pears	1000	1000	1000
0 apples 1500 pears	750	0	0

we can assume that they will trade 1500 apples for 1500 pears. Then Anderson will be able to consume 1500 fruit-pairs. If Anderson has produced equally many apples and pears, there will be no trade. Anderson will then consume 1000 fruit-pairs.

Notice that Anderson would never chose to produce only pears. The reason is that it is better for him to produce equally many apples and pears than to specialize in the wrong fruit, independent of what Peterson choses to do. In fact, also Peterson would never specialize in the wrong fruit. One may say that specializing in the wrong activity is a dominated strategy. We can thus simplify the analysis by assuming that they both only consider two choices, namely to "produce for their own consumption" or to "specialize according to advantage," as described in Table 2.2.

Table 2.2.: Fruit-pairs consumed by Anderson after exchange

Anderson	Peterson	
	Not	Specialize
Specialize	0	1500
Not	1000	1000

Anderson's choice between the remaining two options is a difficult one. If he decides to produce for his own consumption he can be sure to consume at least some fruit. But, if Anderson specializes in apples and Peterson actually turns up at the marketplace with

pears, he will be able to consume 50 percent more. The snag is that Peterson may choose to produce for his own consumption. Then Anderson will end up without any fruit. Anderson's well-being depends not only on what he decides himself, but also on what Peterson decides to do.

In fact, Anderson's best choice depends on what he believes that Peterson will do. If Anderson believes that Peterson will specialize, then it is in Anderson's best interest to also specialize. But if Anderson believes that Peterson will not specialize, then it is in Anderson's best interest to also not specialize.

Peterson's situation is essentially the same. In fact, we may fill in Peterson's consumption of fruits in the same table. Thus, the first entry in each cell of Table 2.3 describes Anderson's consumption of

Table 2.3.: Fruit-pairs consumed by Anderson and Peterson after exchange

Anderson	Peterson	
	Not	Specialize
Specialize	0, 1000	1500, 1500
Not	1000, 1000	1000, 0

fruit-pairs and the second entry describes Peterson's consumption. Clearly, among the four possible outcomes, one is better than all the other, for both people. If both Anderson and Peterson specialize according to advantage, they will both end up consuming 1500 fruit pairs. This is the *efficient* outcome.

Before reading any further, think about the following problem.

Exercise 3. Is there in your view any reason to believe that Anderson and Peterson will be able to reach the efficient outcome, or will they both play it safe and produce for their own individual consumption? Is there even a risk that one chooses to specialize at the same time as the other chooses to not specialize.

There is no right or wrong answer to this question. Personally, I

would guess that people most often are able to reach the efficient outcome in situations of this kind, at least when only a few people are involved.

First notice that if Anderson for some reason believes that Peterson will bring 1500 pears to the marketplace next year, then it is in Anderson's best interest to specialize in producing apples. Similarly, if Peterson for some reason believes that Anderson will bring 1500 apples to the marketplace next year, it is in his best interest to specialize in pears. When they meet at the marketplace, their beliefs are confirmed, and they get to eat a lot of fruit.

Second notice that since Anderson and Peterson know each other well, i.e. they know each other's productivities and preferences for fruits, they can actually figure out that the other party does have an incentive specialize. They both understand that they could make an agreement to specialize beforehand. Equally important, they both understand that they both have an interest in keeping such a promise to specialize. Therefore they should both be willing to make such an agreement even absent a legal system and courts. One may say that an agreement to specialize and trade is *self enforcing*. (We will analyze situations of this kind more thoroughly in the lecture on game theory.)

Still, the main point here is that specialization and exchange cannot be taken for granted. While specialization and exchange may arise, it requires some form coordination. In a small village, it may suffice with an agreement between two neighbors.

2.2. Prices

But, what if we think about the entire world with seven billion people and who knows how many goods and services? What if they know neither how productive each and every neighbor is in every possible productive activity, nor what their neighbors like to

consume and in what proportions? What if they cannot all meet at the same time and agree on a common plan? Will cooperation in the form of specialization and exchange then break down? Not so, fortunately.

So, let us now consider such an economy with "many" people. All people wish to consume apples and pears in even pairs and they wish to consume as many pairs as possible. All people work 1500 hours which they can divide any way they want between producing apples and producing pears. People have different productivities. A person with productivity x can produce x apples per hour worked, and thus $1500 \cdot x$ apples if he specializes in producing apples. The more productive a person is in producing apples, the less productive he is in producing pears. A person with productivity x can produce $1 - x$ pears per hour and thus $1500 \cdot (1 - x)$ pears if he specializes in producing pears. If a person with productivity x spends h hours producing apples he can spend $1500 - h$ hours producing pears. His total production will then be $h \cdot x$ apples and $(1500 - h) \cdot (1 - x)$ pears. For simplicity we assume that there are four people and that they have the following x-values: 0.2, 0.4, 0.6, and 0.8. Everyone knows his own productivity and preference for consuming fruits in even pairs. They do not, however, know anything about other peoples' productivities or preferences.

Please try to solve each exercise before you look at the answer that follows immediately after the exercise. It may be helpful to cover the answer with a sheet of paper.

Exercise 4. Assume first that people cannot trade and that each person thus has to produce the fruit he wants to eat. How many apples and pears will a person with productivity x then produce and eat?

Answer: Since every person wants to eat equally many apples and pears, a person with productivity x will select the number h of

hours of apple-production to satisfy

$$h \cdot x = (1500 - h) \cdot (1 - x) .$$

This is an equation in one unknown, namely h. Solving the equation for h yields $h = 1500 \cdot (1 - x)$. Thus, a person with productivity x will produce $1500 \cdot (1 - x) \cdot x$ apples and equally many pears. You may notice that the more productive a person is in producing apples, the fewer hours that person will use for actually producing apples. ◀

Exercise 5. Assume now that people will be able to trade fruits on a market. For some (unexplained) reason, the only acceptable (relative) price is to exchange one apple for one pear. All people are informed about this price and, for some unexplained reason, they are all *convinced* that they can trade as many fruits as they wish at this price. They make up their individual production and consumption plans accordingly. How many apples and pears will each person *plan* to produce, trade and eat?

Answer: We will answer this question in two steps. First we will decide how much a person with productivity x will produce to maximize his market income. We will measure market income by the number of apples a person could buy if he would not buy any pears. (We could equally well have chosen to measure income in terms of pears. It does not matter which fruit we use as a so-called numeraire. But the analysis is easier if we do use one of them as a numeraire.) Second we will decide how much that person will plan to consume.

A person with productivity x expects to earn an income of x apples per hour spent on producing apples. The same person expects to earn an income of $1-x$ apples per hour spent on producing pears, since fruits can be exchanged one for one. Notice that $x > 1 - x$ if and only if $x > \frac{1}{2}$. Thus a person with $x > \frac{1}{2}$ will maximize his

income by specializing in apples. Similarly, a person with $x < \frac{1}{2}$ will maximize his income by specializing in producing pears. The total income of a person with $x > \frac{1}{2}$ is thus $1500 \cdot x$ and the total income of a person with $x < \frac{1}{2}$ is thus $1500 \cdot (1 - x)$.

Since the cost of buying an apple-pear-pair is 2 apples, a person with $x > \frac{1}{2}$ will thus plan to consume $750 \cdot x$ fruit pairs and a person with $x < \frac{1}{2}$ will thus plan to consume $750 \cdot (1 - x)$ fruit pairs.

To compute how much people plan to trade, we simply take the difference between their production and consumption plans. A person with $x > \frac{1}{2}$ produces $1500 \cdot x$ apples and plans to consume $750 \cdot x$ apples and pears. Thus, this person plans to "sell" $750 \cdot x$ apples and to "buy" equally many pears. A person with $x < \frac{1}{2}$ produces $1500 \cdot (1 - x)$ pears and plans to consume $750 \cdot (1 - x)$ apples and pears. Thus, this person plans to "sell" $750 \cdot (1 - x)$ pears and to "buy" equally many apples. ◀

Exercise 6. In total, how many apples and pears do people plan to "sell" in this economy?

Answer: The people with $x > \frac{1}{2}$ specialize in producing apples. In total, they plan to sell

$$A_s = 750 \cdot (0.8 + 0.6) = 1050$$

apples. The people with $x < \frac{1}{2}$ specialize in producing pears. In total, they plan to sell

$$P_s = 750 \cdot [(1 - 0.4) + (1 - 0.2)] = 1050$$

pears. ◀

Exercise 7. In total, how many apples and pears do people plan to "buy" in this economy?

Answer: The people with $x < \frac{1}{2}$ specialize in producing pears. In total, they plan to buy

$$A_b = 750 \cdot [(1 - 0.4) + (1 - 0.2)] = 1050$$

apples. The people with $x > \frac{1}{2}$ specialize in producing apples. In total, they plan to buy

$$P_b = 750 \cdot (0.8 + 0.6) = 1050$$

pears. ◀

Exercise 8. Can all people realize their plans at the same time?

Answer: Yes, all people can realize their plans at the same time since the total amount that people plan to "buy" equals the total amount that people have produced with the aim to "sell," i.e $A_b = A_s$ and $P_b = P_s$. ◀

Exercise 9. How many more fruit pairs can people expect to consume as a result of specialization?

Answer: All people gain from specialization. A person with productivity $x > \frac{1}{2}$ expects to consume $750 \cdot x$ fruit pairs instead of $1500 \cdot (1 - x) \cdot x$. The gain is thus

$$750 \cdot x - 1500 \cdot (1 - x) \cdot x = 1500 \cdot \left[x - \tfrac{1}{2}\right] \cdot x > 0.$$

A person with productivity $x < \frac{1}{2}$ expects to consume $750 \cdot (1 - x)$ fruit pairs instead of $1500 \cdot (1 - x) \cdot x$. The gain is thus

$$750 \cdot (1 - x) - 1500 \cdot (1 - x) \cdot x = 1500 \cdot \left[\tfrac{1}{2} - x\right] \cdot (1 - x) > 0.$$

◀

Exercise 10. What information did a person in this economy use to make his production, exchange and consumption plans? What information did he not have?

Answer: Everyone was assumed to know their own productivity and their own preferences. They where also assumed to know the (relative) price of apples and pears. They where not informed about other people's preferences or productivities. They where also not informed about other people's plans. ◀

Exercise 11. Assume now that the (relative) price is to exchange one apple for two pears. Assume also that all people are convinced that they can trade as many fruits as they wish at this price and that they make up their individual production and consumption plans accordingly. Will all people then be able to realize their individual production and consumption plans at the same time?

Answer: A person with productivity x expects to earn an income of x apples per hour spent on producing apples and an income of $(1-x)\cdot\frac{1}{2}$ apples per hour spent on producing pears. Thus a person with $x > \frac{1}{3}$ will maximize his income by specializing in apples and person with $x < \frac{1}{3}$ will maximize his income by specializing in producing pears. The total income of a person with $x > \frac{1}{3}$ is thus $1500 \cdot x$ apples and the total income of a person with $x < \frac{1}{3}$ is thus $750 \cdot (1-x)$ apples.

Since the cost of buying an apple-pear-pair is $\frac{3}{2}$ apples, a person with $x > \frac{1}{3}$ will thus plan to consume $1000 \cdot x$ fruit pairs and a person with $x < \frac{1}{3}$ will thus plan to consume $500 \cdot (1-x)$ fruit pairs.

A person with $x > \frac{1}{3}$ plans to sell apples and to buy pears. Such a person plans to sell $1500 \cdot x - 1000 \cdot x = 500 \cdot x$ apples and to buy $1000 \cdot x$ pears. A person with $x < \frac{1}{3}$ plans to sell pears and to buy apples. Such a person plans to sell $1500 \cdot (1-x) - 500 \cdot (1-x) = 1000 \cdot (1-x)$ pears and to buy $500 \cdot (1-x)$ apples.

In total people plan to sell

$$A_s = 500 \cdot (0.8 + 0.6 + 0.4) = 900$$

apples and to sell

$$P_s = 1000 \cdot (1 - 0.2) = 800$$

pears.

In total people plan to buy

$$A_b = 500 \cdot (1 - 0.2) = 400$$

apples and to buy

$$P_b = 1000 \cdot (0.8 + 0.6 + 0.4) = 1800$$

pears.

Thus, all people will not be able to realize their individual plans. There will be more apples produced for the market (900) than asked for (400), thus those who have produced apples will not be able to sell all apples that they wish. There are also fewer pears for sale (800) than people had planned to buy (1800). Thus, those who have planned to buy pears will not be able to buy as much as they had planned.

Only the one person who produced pears is able to realize his plan. He will exchange 800 pears for 400 apples (assuming that the pre-determined relative price indeed must be obeyed). As a result, he will consume 400 fruit-pairs. It is not clear with whom he will trade though.

One possibility is that all the apple producers are allowed to realize 4/9 of their plans. Then the person with $x = 0.8$ will sell $500 \cdot 0.8 \cdot \frac{4}{9} \approx 178$ apples and buy $1000 \cdot 0.8 \cdot \frac{4}{9} \approx 356$ pears. (This person will have to throw away 666 apples.) He will consume 356

fruit-pairs, to be compared to the 240 fruit pairs that he would consume absent the possibility to trade. The person with $x = 0.6$ will sell $500 \cdot 0.6 \cdot \frac{4}{9} \approx 133$ apples and buy $1000 \cdot 0.6 \cdot \frac{4}{9} \approx 266$ pears. (This person will have to throw away 501 apples.) He will consume 266 fruit-pairs, to be compared with the 360 fruit-pairs he would consume absent the possibility to trade. The person with $x = 0.4$ will sell $500 \cdot 0.4 \cdot \frac{4}{9} \approx 89$ apples and buy $1000 \cdot 0.4 \cdot \frac{4}{9} \approx 178$ pears. (This person will have to throw away 333 apples.) He will consume 178 fruit-pairs, to be compared with the 360 he would consume absent the possibility to trade. Thus, in the end, some of the apple producers regret that they did not choose to produce for their own private consumption only.◀

These exercises demonstrate how people sometimes may succeed to collaborate even if they know nothing about one another. They may all specialize according to their advantage even if they do not know how productive other people are in various activities. They may all be able to sell what they have produced and they may all be able to buy what they have planned to consume.

Instead of relying on detailed information about one another and on agreements beforehand, coordination may sometimes be achieved through prices.

The catch is that the (relative) prices must be carefully calibrated. If not, all people will not be able to realize their individual plans at the same time. They will not be able to sell what they have produced and they will not be able to buy the goods they have planned to consume. They may then be better off not taking part in this kind of collaborative scheme. They may prefer to become self-sufficient and produce what they wish to consume themselves. The question is then if there are any mechanisms for calibrating prices. We will return to this question in the lecture on perfect competition. Be warned, however, that this question still has only a partial answer.

3. Money

3.1. Problem 2: Double coincidence of wants

When the Swedish aerospace company Saab sold its multi-role fighter aircraft JAS 39 Gripen to Thailand, birds may have been traded for birds, according to some American embassy reports that where revealed by Wikileaks. The Thai government demanded that 80 000 tons of frozen chicken was accepted as part of the payment for the fighter jets. The Wikileaks documents do not reveal whether Saab, in turn, paid its workers and share holders in frozen chicken the same year, chickens that they, in turn, may have used to pay their rents. And perhaps the landlords, in their turn, used the chicken to pay for their winter holidays in Thailand.

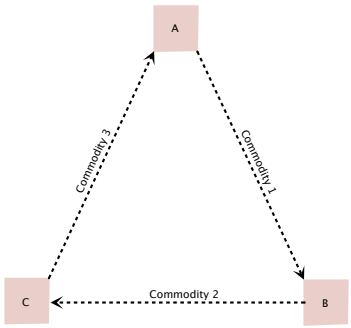
This example demonstrates that barter and other types of in-kind transactions are associated with severe problems. For you and I to barter, you need to have what I want and to want what I have. Such **double coincidence of wants** are rare. The solution may be to use money as a medium of exchange.

3.2. Commodity Money

Consider a hypothetical country in which nobody likes to consume what he can produce and where barter is difficult due to a lack

of double coincidence of wants.[1] To be more specific, assume that there are three types of individuals. People of type A can only produce a commodity called 1 that is only consumed by people of type B. People of type B can only produce a commodity called 2 that is only consumed by people of type C. People of type C can only produce a commodity called 3 that is only consumed by people of type A. This pattern is summarized in Figure 3.1.

Figure 3.1.: Double coincidence of wants fails



I this country trade must be *bilateral* since people only meet in pairs. To be more specific, assume that there are equally many people of all types and that, each day, every person meets with a randomly selected person from both of the other types, but at different times. In the morning, every person of type C meets with a randomly selected person of type B. Then, at noon, every person of type B meets with a randomly selected person of type A. Finally, in the evening, every person of type A meets with a randomly selected person of type C. At all these bilateral meetings, it is possible for people to exchange commodities with one another, it they so wish.

Trade must also be *quid pro quo* since nobody can be sure to meet the same person again. That is, to receive something from

[1] This section is based on Kiyotaki, N., & Wright, R. (1989). On money as a medium of exchange. The *Journal of Political Economy*, 97(4), 927-954.

somebody, one has to give something back at the same time and place.

Let us also assume that exchange would be *efficient*. To be more specific, we assume that any person who gets to consume his desired commodity will feel better off, but the producer will feel worse off due to the effort involved in production. Clearly, the best possible day for any individual would be a day with consumption and no production and the worst possible day would be to produce but not to consume. More crucially we assume that people would prefer a day with both consumption and work to a day without both consumption or work.

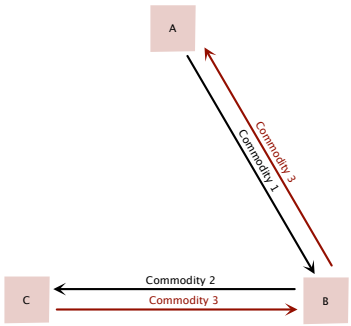
If the country would be governed by a benevolent dictator, there would be efficiency. This dictator could order all people of type B to provide a person of type C with a unit of their commodity in the morning, all people of type A to provide a unit of their commodity to a person of type B at noon, and all people of type C to provide a unit of their commodity to a person of type A in the evening. Unfortunately, dictators are often not benevolent and they typically lack information about peoples' preferences and abilities. If the dictator cannot say who is an A person, B person or C person, he could not enforce this plan. So the idea of the benevolent dictator would not work.

Also voluntary exchange is complicated to arrange, since trade will only occur if mutually agreeable and since in each meeting only one party wishes to consume what the other can offer.

There is a solution, however, namely indirect trade, as described in Figure 3.2. When C and B meet in the morning they may exchange commodities, a unit of commodity 3 for a unit of commodity 2. B will not consume commodity 3 but rather store it. Then, at noon, B and A may exchange a unit of commodity 1 for a unit of commodity 3. (The evening meeting be- tween A and C will then be superfluous.)

With this arrangement, people of type B will act as middle-men,

Figure 3.2.: Commodity money



transferring commodity 3 from people of type C to people of type A. And B will not hesitate to accept commodity 3 in return for commodity 2, even if it has no consumption value for them, since they know it will have a trading value for them later on. When an object (such as commodity 3 here) is accepted in trade not to be consumed or used in production, but to be used to facilitate further trade, it becomes a medium of exchange and is called **commodity money**.

Commodity money are problematic, however, as both storage and transportation is often costly. Such costs are examples of transaction costs. In order for the arrangement with indirect trade and commodity money above to work, a B-person must value consumption higher than the combined effort of production and storage and transportation. Otherwise, no one would be producing anything and no one would be consuming anything. There would be no exchange. Markets would not exist.

In sum: Direct barter does not work when there isn't a double coincidence of wants. Indirect trade with commodity money is plagued by problems of transportation and storage. Transaction costs may even cause markets to fail. That is, socially valuable

production and consumption may not occur due to the problems of organizing exchange. (And credit is not a solution in our situation either.)

3.3. Fiat money

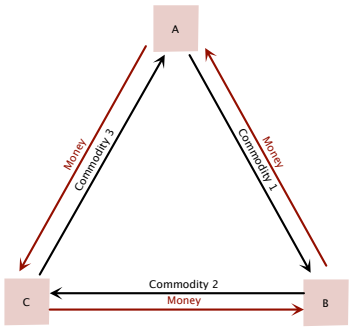
Assume now that someone provides the country with a load of small pieces of paper called one-euro bills. The bills are suddenly distributed during a night and each person of type C receives one bill. The bills cannot be used for anything and are completely worthless. Except, of course, that the creator of the bills suggest that anyone in possession of such a bill may leave it in exchange for a unit of service. And if the whole population accepts this idea, they may start to trade.

In the morning, B-people provide C-people with services. In exchange every producer receives a one-euro bill. At noon, A-people provide B-people with services. In exchange every producer receives a one-euro bill. In the evening C-people provide A-people with services. As a result, during the first day after the invention of Euro bills, everyone consumed and produced one unit of service. And, when the day ends, all C-people again holds a one-euro bill, which they can use to buy services on the morning of day 2. This pattern of exchange is described in Figure 1.5.

There are no transportation and storage problems.

But why do people of type C accept to provide people of type B with services in exchange for a piece of paper? The answer is, of course, that they expect that people of type A will accept the money in exchange for services later during the day. People of type A, in turn, accept money as they expect people of type C to accept the money in the evening. Thus, only if all parties are certain that everybody else will accept the paper money as payment for future services will this arrangement work.

Figure 3.3.: Fiat money



If an object with no intrinsic value becomes a medium of exchange, it is called **fiat money**. Fiat money takes on value, essentially as a self-fulfilling prophecy. Whether or not fiat money can serve as a medium of exchange hinges crucially on whether people believe that it will. In other words, the use of money involves a strategic element and aspects of social custom. In the modern world, governments contribute to guarantee the value of fiat money e.g. by accepting the money as payment for taxes.

Additional exercises

I hope that you take this course and read this book because it is fun. But I also hope that you wish to learn how to analyze economic problems at the same time. You will consequently be asked to solve problems and recapitulate the main points as you read. Let's start immediately:

1. Why do people trade goods and services with one another?
2. What is the difference between absolute and comparative advantage?
3. What is meant by opportunity cost?
4. What is meant by commodity money?
5. What problem does commodity money solve?
6. What problems does such money cause?
7. What is meant by fiat money?
8. What problems does such money solve?
9. What is the role of prices?

Wait, please answer the questions before reading any further. And answer them carefully. Taking tests improves learning. Newly published research shows that students who read a passage, then took a test asking them to recall what they read, retained 50 percent

more of the information a week later than students who used other methods, e.g. studying the material repeatedly.[2]

[2]The research proving that taking tests is a superior way of learning is published in: Jeffrey D. Karpicke and Janell R. Blunt. Retrieval Practice Produces More Learning than Elaborative Studying with Concept Mapping. Science, Published On- line 20 January 2011.

4. Aims of this book

The story about Anderson and Peterson demonstrates that both people would gain from specialization and trade, nothing more. Trade presumes agreement, which is not guaranteed. Anderson and Peterson are e.g. likely to disagree on the terms of trade, as both parties probably want to secure the best possible deal for themselves. And if they are too greedy, they may well end up making incompatible demands, and trade may be disrupted.

This book is concerned with how markets work: what the rules of the market are, how people behave when they exchange goods and services with one another under these market rules, and what the final outcome is. The question is if markets make people produce the goods they are relatively suited for, consume the goods they enjoy, and exchange the difference? In other words, are markets efficient? A first example of why markets may fail is the absence of a double coincidence of wants. A second example of why markets may fail is that prices do not clear the market.

The book is primarily focused on the details of individual markets rather than on how all the different markets in the economy interact. The book is also focused on final and intermediate goods markets leaving, in particular, labor and financial markets out. We will also disregard public provision of goods and services.

Stylized facts

One of the aims of this book is to present some important "stylized facts" about how markets work. Examples include:

1. Medical companies set very different prices for the same drug in different countries, even though the production costs are the same and the distribution costs are similar.

2. When a patent for a medical drug expires and new firms start producing copies, the producer of the original drug lowers its prices. Similarly, the hourly price for driving lessons is lower in municipalities with many driving schools than in municipalities with only a few schools.

3. Large retail chains often buy their products cheaper from their suppliers than smaller retail chains can.

4. Mobile phone operators allow their customers to choose from a menu of different pricing plans, typically including a flat-rate plan with a high monthly fee but with no additional costs for making calls.

5. Software companies often market versions of their programs with some features disabled. This practice may appear somewhat puzzling as disabling features increase the software companies' costs.

6. For a long time 7UP was marketed as the un-cola to set it apart from competing soft drinks. The commercial radio stations in Sweden do not appear to strive for differentiation. Most of them simply play the same hit lists over and over again.

7. The industries that are relatively concentrated in Sweden are also relatively concentrated in other countries, but some industries remain concentrated also in large economies such as in the US.

8. Mergers often reduce the merging firms' profits, but raise their share-prices. So, why do all these unprofitable mergers occur?

And how can share prices increase when profits are reduced?

Methodology

The second and main aim of the book is to introduce some methods that have proven useful for analyzing how markets work, to understand regularities such as those just mentioned.

Models = Imaginary economies

Understanding in economics often takes the form of constructing a model. An alternative name would be "imaginary economy." The exercise about Anderson, Peterson, apples and pears above is an example of such a model or imaginary economy. The stylized fact we wish to understand with that particular model is the drive for exchange itself. By constructing a simple imaginary economy consisting of only two people and only two goods and assuming (i) that both people wish to consume apples and pears in fixed proportions and (ii) that they have different productivities, we could demonstrate that Anderson and Peterson wish to specialize and trade with each other. Even if the example with absolute advantage may seem trivial, the second example with comparative advantage is much less transparent. But playing around with explicit productivity numbers clarifies that it is comparative advantage, not absolute advantage, that make them exchange apples and pears.

Still, many people are put off by this simplistic approach. Nobody cares about the imaginary Messrs. Anderson and Peterson. Apples and pears may strike you as rather silly metaphors for goods and services in general. And while most people love variety, fixed proportions is a blatantly false description of such preferences. Since when is caricature science?

We use imaginary economies for at least two reasons. The first reason is that real-world markets are far too complex for us to "see"

what is going on. Constructing an imaginary economy allows us to disregard many details that we may have reasons to believe are unimportant for the issues at hand. By only thinking about an economy with two people and two goods, rather than seven billion people and who knows how many goods, the analysis is easier. The second reason is that we can make experiments in imaginary economies. If we wish to investigate if a certain tax would hurt certain people, we can study this issue without actually increasing the tax and possibly hurting these people.

But simplification is also a weakness. Simplifying assumptions are by definition false descriptions of the world. We can never be sure that what is true in an imaginary world is also true in the real world. To make up our minds we need common sense: Is it likely that our conclusions (e.g. that comparative advantage drives trade) rely in some critical way on our simplifications (e.g. that people wish to eat apples and pears in fixed proportions)?

We may also believe more in the predictions if the same predictions can be generated in many different models that are characterized by different sets of simplifying assumptions. But, ultimately, our belief in the economic models should rest on empirical testing.

Game theory

Studying markets is studying peoples' exchange of goods and services with one another. Exchange is interactive. What one person does clearly affects the welfare of others. What one person does also affects what choices other people wish to make. Anderson can specialize only if Peterson does. Decisions are in other words interdependent. For this reason we will make heavy use of the theory of interdependent decision making, better known as game theory.

This is not a course in game theory, but as we repeatedly apply game theoretic analysis to economic problems, you will gradually learn how to use it. And, actually, we do start out with a quick

primer in **Part I**I.

Empirics and identification

Empirical work is important both to test if our models succeed to describe the way markets work and also to quantify important relationships.

We will look at one empirical issue. Since competition is one of the key factors determining how well markets work, we will introduce some empirical methods that have been used to measure competition in markets. The question is what data do we need to collect and how should we analyze the collected data if we wish to say how much the firms in a market compete with each other?

Public policy

The third and final aim of this book is to describe the role of some public policies such as competition policy and regulation that have been introduced to improve how market work.

Part II.

Basic Market Forms

5. Monopoly

The Swedish railway operator SJ is the only firm providing train services between most Swedish towns. On short distances, such as the 200-kilometer route between Stockholm and Linköping, air transportation is not a convenient alternative. And, at least for people who travel in their job, the buses are much too slow. So, if you don't have a car, you have to rely on SJ. With only one seller, the market might be characterized as a monopoly. Until recently also electricity, telephony and water markets where monopolies in many countries. But since many utility markets have been deregulated lately, allowing competing firms to enter, monopolies are much less frequent today than they used to be. Even though they are becoming less and less common, it is still important to study how monopolies work.

5.1. Why study monopoly?

Monopolies still exist An obvious reason for studying monopoly is that some markets still are monopolies. An important example is many pharmaceutical markets. The reason is that pharmaceuticals are often protected by patents. Thus, only the company inventing a drug has the right to manufacture and sell it for a long period of time. In most Western industrialized countries patents last for 20 years. As a result, many pharmaceutical markets are in effect monopolies.

When looking at pharmaceutical prices an interesting pattern is revealed. Prices differ largely between different countries. For in-

stance, Lipitor, the #1 cholesterol-reducing medicine in the world, was sold at USD 1.46 in the US and at only USD 0.94 in Sweden in 1998 (manufacturer prices per dosage, 10 mg tablets). The same year the price of Losec was USD 2.99 in the US and USD 1.74 in Sweden (manufacturer prices per dosage, 20 mg tablets). Also within Europe prices differ substantially. A comparison of 90 pharmaceutical products in 1998 revealed that prices were highest in Switzerland (25% above the European mean) and lowest in Greece (28% below the European mean) with Sweden approximately at the mean. The full comparison is presented in Table 5.1 .[1]

Table 5.1.: Average percentage deviation from European mean prices for 90 pharmaceutical products in 1998

Country	Deviation
Greece	-28
Spain	-20
Portugal	-13
Italy	-13
France	-10
Finland	-2
Austria	-2
Norway	-1
Sweden	-1
Belgium	-1
Netherlands	+2
Denmark	+3
Germany	+11
UK	+19
Switzerland	+25

Since medicines are often produced where it is cheap and then transported to the consumers around the world, the costs of supplying medicines are presumably rather similar in different European

[1]Keith Maskus (2001): Parallel imports in Pharmaceuticals: Implications for Competition and Prices in Developing Countries, Report to World Intellectual Property Organization.

countries. It is therefore unlikely that the differences in prices are caused by differences in costs. So, what might explain these price differences?

Recall that in perfectly competitive markets, prices are given by the average cost of production (the minimum long-run average cost, to be more precise). So, if the pharmaceutical markets in all countries had been perfectly competitive, the prices would have been almost the same in all countries. Thus the explanation for these price differences must have something to do with the fact that competition is limited in many pharmaceutical markets.

And, indeed, in monopoly markets, prices are not only determined by the cost of production. Also the consumers' willingness to pay matters. And, looking closer at the price differences above reveals that prices tend to be higher in countries with higher per-capita incomes where willingness to pay presumably is higher and lower in poorer countries where willingness to pay is lower. So, the question is then, why and exactly how monopoly prices depend on demand, i.e. the consumers' willingness to pay. To answer this question, we need to study the basic model of monopoly pricing.

Public policy Another reasons for why we need to understand how monopolies work is that there are several public policies actively preventing markets to become monopolized. Examples include the merger regulation, requiring firms to seek approval from competition authorities before they join together, and the Swedish press subsidy, providing financial support to the second largest newspaper in every local newspaper market. Other political interventions aim at deregulating monopoly markets to open them for competition. Examples are opening countries for trade or to allow private enterprises to be active in markets where only publicly owned firms previously have been active. To evaluate the effects of all these policies, we need to know how the market would work absent the policy, i.e. we need to know how monopolies work. The

money given to the second largest newspaper are well spent only if the quality of the press is higher or if the prices are lower than what they would have been absent competition.

Preparation for the study of oligopoly Finally there is a more practical reason. Studying how firms behave on monopoly markets is a good preparation for studying the much more difficult question, how firms behave when they have to compete with other firms. In some respects monopoly and oligopoly markets are similar. Firms face a trade-off between mark-ups and sales, irrespective of whether they have competitors or not. In other respects oligopoly markets are much more complex. Oligopoly firms have to predict their rivals' pricing in order to set their own prices. Monopoly firms only have to predict the consumers' purchasing decisions. It is much easier to first understand how to analyze the first problem (the trade-off between sales and mark-up) before also including the strategic interaction between different firms.

5.2. What is monopoly?

Definition The market for a certain product or service is a monopoly if (i) there is only one firm selling the product in that market, (ii) there are no close substitutes for the product, (iii) the barriers to entry are sufficiently high for other firms not to enter, and (iv) every individual customer (consumer or firm) buys a very small fraction of the firm's total sales.

When a market have these characteristics, the firm can set its price without thinking about other firms' behavior or about how individual customers behave.

In practice, however, it is often difficult to say if a certain market is a monopoly or not. Monopoly power is a matter of degree. The question is if a certain firm can raise its price high above production cost for an extended period of time without loosing so many

customers to some other firm that the increase is unprofitable. If it can, the firm has a high degree of monopoly power. Otherwise not.

Entry barriers In most cases, the ultimate reason why there is only one firm in the market and why there are no close substitutes is probably entry barriers. In this sense, all the conditions (i) - (iii) can be reduced to one - entry barriers. And entry barriers are also the ultimate reason why many other markets only have a few competitors.

Entry barriers come in many forms. There are **legal entry barriers** such as patents to provide incentives for R&D investments and copy rights to provide incentives for artistic work. Sometimes legal barriers are motivated by the government's wish to control consumption. In Sweden, alcohol (Systembolaget) is an example. In other cases the motive may be fiscal. Gambling (Svenska Spel) is a case in point.

But perhaps the most common entry barrier is that production is characterized by **economies of scale**, i.e. that production costs can be lowered the more a firm produces. In order for scale economies to prevent entry, the average production cost must be falling in the quantity produced even when the firm produces a large share of the total quantity demanded. Expressed differently, the minimum efficient scale must be large compared to the size of the market. In rural areas, food retailing is a monopoly. But in the city centers, there is room for many food retailers, and the market may be quite competitive. District heating, on the other hand, is a natural monopoly also in city centers.

Also exclusive access to **essential resources** constitute barriers to entry. In some cases, a single firm may own a natural resource. In other cases, firms may have so-called exclusive distribution agreements. For example, only Bilia can sell Volvo in Sweden. (Exclusive distribution agreement seldom lead to full monopoly, however. While there is no intrabrand competition, there is of course inter-

brand competition in the car market. Actually, exclusive distribution agreements my not even be legal in case there is no interbrand competition.)

Yet another type of entry barrier is **network effects**. In Sweden it would be difficult for anyone to start competing with Blocket. Those who want to sell a second-hand car will stay at Blocket since all the potential buyers use Blocket. And those who wish to buy a second-hand car will continue to use Blocket, since thats where all cars are advertised.

5.3. The monopoly model

Assumptions Consider a single market served by only one firm – called the monopolist. In this market, the monopolist is free to set whatever price it wishes. Once the price is chosen, the consumers are informed about the price and decide how much they wish to buy. The monopolist is required to produce and sell the quantity demanded at the price it offers.

The monopolist's production technology is summarily described by its cost function, $C(q)$, which tells us how high the monopolist's costs would be for any possible number of units, q, that the firm might choose to produce. The extra cost of increasing production by an infinitesimal amount, i.e. the marginal cost, is positive, $C_q(q) > 0$. Note that a sub-index on a function-symbol denotes a derivative. Sometimes we will assume that the cost function is given by $C(q) = c \cdot q + F$, where c is the marginal cost and F is a fixed cost. We say that the marginal cost is constant since it is the same, independent of the level of production.

The consumers' choices are not studied in detail here, but simply summarized by the demand function, $D(p)$. The demand function describes how many units of a product or service that the consumers would buy, for all possible prices that the monopolist might

charge. The monopolist knows demand, e.g. as a result of historical information about sales at different price levels. For the moment we will not describe the demand function in much detail. We only assume that consumers buy less at higher prices, or expressed differently, that demand is falling in price, i.e. $D_p(p) < 0$. Sometimes we also assume that the consumers do not buy anything in case the monopolist would charge a price above some price $\hat{p}$, i.e. $D(p) = 0$ for all prices $p \geq \hat{p}$. The consumers' behavior may alternatively be summarized by the inverse demand function, $P(q)$. The inverse demand function describes the consumers' maximum willingness to pay for an additional (infinitesimal) unit of the good, given that they already consume q units. As the name suggests, the inverse demand function is the inverse of the demand function. Before reading any further, please solve the exercise.

Exercise 12. Derive the inverse demand function if the demand function is linear and given by $q = D(p) = \alpha - \beta \cdot p$.

Answer: To find the inverse of a single demand function, solve the equation $q = \alpha - \beta \cdot p$ for the price. It is thus given by $p = P(q) = \frac{\alpha}{\beta} - \frac{1}{\beta} \cdot q$.◀

Finally we assume that the monopolist wants to maximize its profits, which is given by

$$\pi = p \cdot q - C(q).$$

In case the marginal cost is constant and there are no fixed costs, the monopolist's profit is given by the quantity sold times the mark-up, i.e.

$$\pi = (p - c) \cdot q.$$

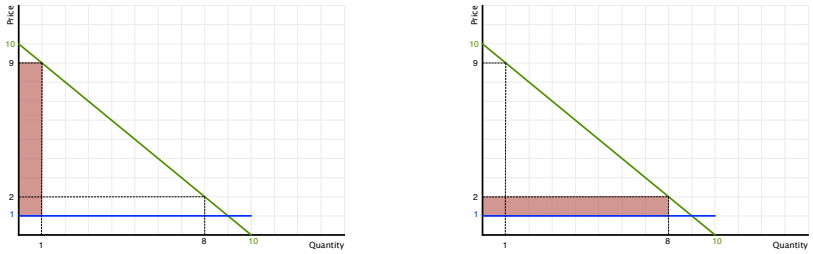
Notice that this monopoly model is defined by a set of assumptions about the market, namely the timing of decisions (monopolists sets price, then consumers decide how much to buy), the objectives of the monopolist (profit maximization) and a specification of the

external conditions, i.e. cost and demand conditions.

The monopolist's basic trade-off A monopolist is free to set its price without thinking about how other firms would react to its price. A monopolist can also ignore if any *individual* consumer would choose to leave the market. Each individual consumer is insignificant to the monopolist's profits. This does not mean, however, that it can set its price without thinking about the consumers as a group. Expressed differently, the monopolists' choice is constrained by the market demand function.

Consider the market described by Figures 5.2a and 5.2b. Demand

Figure 5.1.: The fundamental trade-off



(a) High mark-up (b) High sales

is linear. If the firm increases price by €1 the consumers buy one unit less. Marginal cost is constant and equal to one. If the firm wants to charge a price of €9, it will only sell one unit. And if the firm wants to sell eight units, it can only charge a price of €2. In fact, if the firm charges a high mark-up (say €8), its profit will be low as a result of low sales (in this case €8), as described by the colored rectangle in Figure 5.2a. And, if the firm sells a large quantity (say 8 units), its profit will be low as a result of a low mark-up (in this case €8), as revealed by Figure 5.2b. The monopolist thus faces a trade-off between mark-up and sales. Expressed differently, the monopolist needs to strike a balance between its mark-up and its sales.

5.4. Monopolist's choice

5.4.1. Quantity as a choice variable

The monopolist decides on a price, p, and a level of production, q, to maximize its profit,

$$\pi(p, q) = p \cdot q - C(q).$$

By looking at its indirect demand function, the monopolist can figure out the maximum price it can charge for any level of sales. Clearly, for a given quantity, q, the optimal price for the monopolist to set is $p = P(q)$. Only the optimal quantity remains to be chosen.

To find the optimal quantity, we first substitute the inverse demand function (i.e. the optimal price) into the profit function:

$$\pi(q) = P(q) \cdot q - C(q).$$

The profit is now written as a function of quantity only. Again, writing the profit in this form implicitly assumes that the monopolist will choose the maximum price that consumers are willing to pay for any quantity.[2]

Marginal revenue Before searching for the optimal quantity, it is instructive to note that profit is given by revenues, $P(q) \cdot q$, minus cost, $C(q)$. While the cost function conveniently describes how cost depends on the quantity, it is somewhat less clear to see how the revenues change with the amount sold.

Let us look at a simple numerical example, based on the linear demand function in Figure 5.2. If the monopolist wants to sell one unit, it can charge €9. Two units sell at maximum €8, and so

[2]Notice that I have abused notation slightly by using the same symbol π to describe two different functions, namely both the function describing profit as a function of both price and quantity and the composite function $\pi(P(q), q) = P(q) \cdot q - C(q)$.

Figure 5.2.: Linear demand

on. The same demand function is also presented in the first two columns of Table 5.2. The third column presents the monopolists'

Table 5.2.: How revenues change with increased sales

Sales	Price	Revenues	Change in revenue
0	10	$0 = 0 \cdot 10$	-
1	9	$9 = 1 \cdot 9$	$9 = 9 - 0$
2	8	$16 = 2 \cdot 8$	$7 = 16 - 9$
3	7	$21 = 3 \cdot 7$	$5 = 21 - 16$
4	6		
5	5		
6	4		
7			
8			
9			
10			

revenues associated with every level of sales. If the monopolist sells one unit, total revenues will be €9. Selling two units, generate a revenue of €16.

The fourth column describes how the revenues changed when the last unit of sales was added. This is called marginal (or incremental)

revenue.[3] When selling the first unit a price €9, revenues increase from €0 to €9. The marginal revenue is thus €9. When selling two units at a price of €8, revenues increase from €9 to €16. The marginal revenue is thus €7. The peculiar feature of the revenues is that while the additional unit is sold at the price of €8, revenues are only increased by €7. When selling three units at a price of €7, revenues increase from €16 to €21. The marginal revenue is this €5. Thus, while the additional unit is sold at the price of €7, revenues only increased by €5.

Exercise 13. Complete the table. Compare how revenues change when selling an additional unit with the price at which this unit is sold. Describe the pattern.

Answer: Marginal revenue is lower than price (at every quantity). In fact, marginal revenue may even be negative. ◀

Now to a more difficult question. Before reading any further, please solve the exercise.

Exercise 14. Why do revenues increase by less than the price paid for the last unit sold?

Answer: If the monopolist first sells q_0 units at price $p_0 = P(q_0)$, revenues are given by $R(q_0) = P(q_0) \cdot q_0$. If the monopolist increases sales to q_1 units and reduces price to $p_1 = P(q_1)$, the change in revenue is given by $\Delta R = R(q_1) - R(q_0) = p_1 \cdot q_1 - p_0 \cdot q_0$. By adding and subtracting the same term $p_1 \cdot q_0$, the change in revenues can be written as $\Delta R = p_1 \cdot q_1 - p_1 \cdot q_0 + p_1 \cdot q_0 - p_0 \cdot q_0$. After rearranging this expression slightly, the change in revenues is given by

$$\Delta R = p_1 \cdot (q_1 - q_0) + (p_1 - p_0) \cdot q_0. \tag{5.1}$$

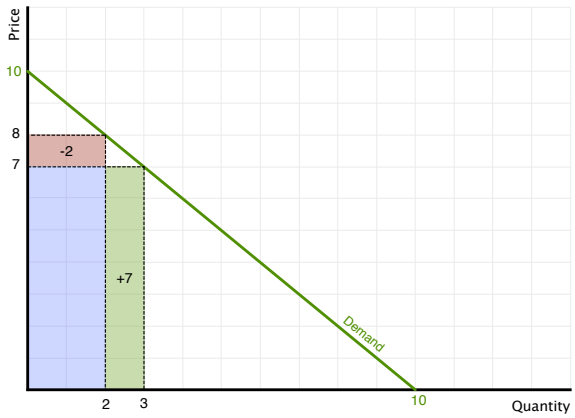
The change in revenues is thus composed of two terms, reflecting two different effects of sales on profit. According to the first term,

[3] The term marginal is more correctly used when the change in quantity is infinitesimal.

increasing sales by one unit, i.e. $q_1 - q_0 = 1$, increases revenues by the price paid for this last unit, i.e. p_1. But according to the second term, there is also a negative effect of increasing sales on revenues. To convince the consumers to buy an additional unit, the monopolist must lower price, i.e. $P(q_1) - P(q_0) < 0$. And, importantly, the monopolist must lower price on all units sold. That is, it is not only the last unit that must be sold at the relatively low price p_1. The monopolist must also lower its price on all the units already sold, i.e. on the so-called infra-marginal units, q_0. This loss is equal to $(p_1 - p_0) \cdot q_0$. ◀

The two effects are also described in the demand diagram in Figure 5.3. Selling two units at a price of €8 generates a revenue equal

Figure 5.3.: Marginal revenue



to the sum of the areas of the blue and red rectangles. Selling three units at a price of €7 generates a revenue equal to the sum of the blue and the green rectangles. Increasing sales from two to three, thus increases revenues by the green area, which represents the price paid for the last unit, but decreases revenues by the red area, which represents the reduction in the price of the infra-marginal units.

The change in revenues resulting from the increase in sales, given

by equation 5.1, can be rewritten as

$$\frac{R(q_1) - R(q_0)}{q_1 - q_0} = P(q_1) + \frac{P(q_1) - P(q_0)}{q_1 - q_0} \cdot q_0.$$

And, if we consider an infinitesimal increase in sales from q_0, we have defined marginal revenue, which is given by

$$MR(q_0) = P(q_0) + P_q(q_0) \cdot q_0,$$

where P_q is the derivative of the inverse demand function.[4] Since price must be reduced to sell more ($P_q < 0$), marginal revenue is lower than price, $MR(q_0) - P(q_0) = P_q(q_0) \cdot q_0 < 0$.

Exercise 15. Compute marginal revenue for a monopolist with the linear inverse demand function $p = \alpha - \beta \cdot q$.

Answer: Total revenue is given by $R(q) = p \cdot q = (\alpha - \beta \cdot q) \cdot q$. Thus, marginal revenue is given by $MR(q) = \alpha - 2 \cdot \beta \cdot q$. ◀

Notice that in the case of linear demand, also the marginal revenue function is linear in quantity, with the same intercept as the inverse demand function and twice the slope. The figure also shows that for any quantity q, the highest price consumers are willing to pay (given by the inverse demand function) is higher than the monopolist's marginal revenue.

The optimality condition: marginal cost = marginal revenue

Recall that the monopolist's profit is given by

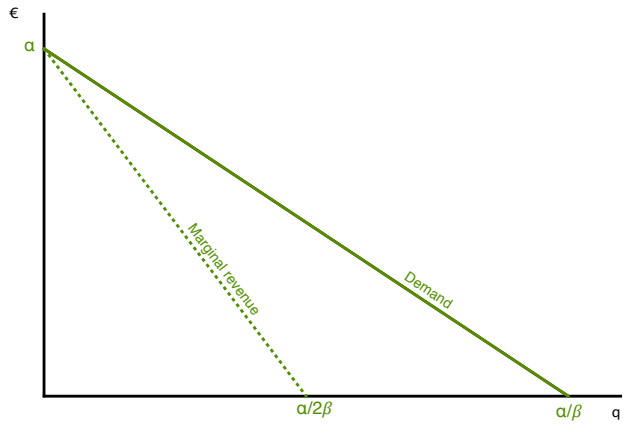
$$\pi(q) = P(q) \cdot q - C(q).$$

The monopolist's first-order condition is given by

[4] Recall that the derivative of a function f is in fact defined by

$$f_x(x_0) = lim_{x_1 \to x_0} \frac{f(x_1) - f(x_0)}{x_1 - x_0}.$$

Figure 5.4.: Marginal revenue



$$\pi_q(q) = P(q) - C_q(q) + P_q(q) \cdot q = 0.$$

The first-order condition may also be rearranged as

$$P(q) + P_q(q) \cdot q = C_q(q).$$

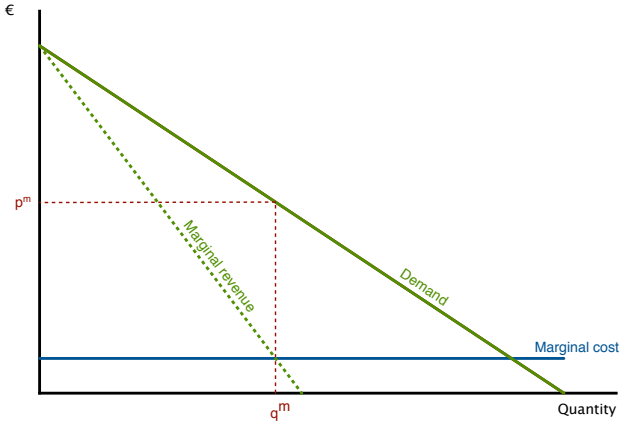
The right hand side is marginal cost, that is the change in cost as a result of a small increase in production. The left hand side we recognize as the marginal revenue, that is the change in revenues caused by a small increase in sales. The optimality condition thus requires the monopolist to produce the quantity that equates marginal cost and marginal revenue. The optimal price, is the highest price that the consumers are willing to pay for this quantity.

The monopoly price Notice that the monopolist will make the consumers pay a price that is higher than the cost of producing the good, that is

$$P(q) - c = -P_q(q) \cdot q > 0.$$

A firm's ability to charge a price above its marginal cost without losing all sales, is often referred to as **market power**.

Figure 5.5.: Marginal revenue equal to marginal cost.



The monopoly quantity Also notice that the monopolist will produce and sell a quantity, which is smaller than the quantity equating price to marginal cost. The quantity q^c equating price and marginal cost, i.e. $P(q^c) = c$, may be called the competitive quantity.[5]

The result that the monopoly quantity is lower than the competitive quantity is just another way to state the result that the monopolist charges a price above marginal cost. In fact, it is only by producing and selling a small quantity (below the competitive level) that the monopolist can charge a high price (above marginal cost).

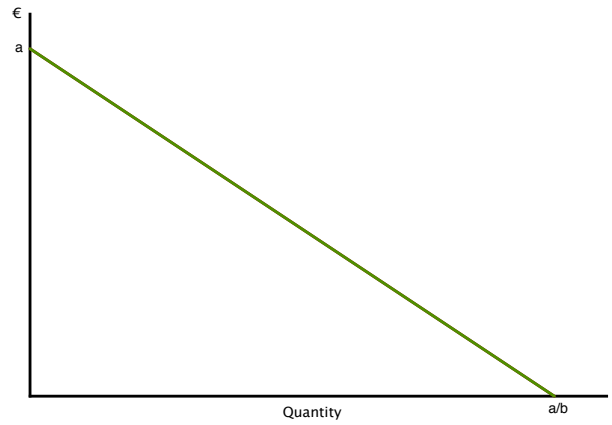
Exercise 16. Consider a monopolist with cost function $C(q) = F + c \cdot q$. Assume that inverse demand is linear and given by $P(Q) = a - b \cdot Q$, where $a > c > 0$ and $b > 0$, as described in Figure 5.6.

1. What is the interpretation of a and b?

2. Why do we assume $a > c$ and $b > 0$?

[5] To see that the monopoly quantity q^m is smaller, rewrite the first order condition as $P(q^m) + P_q(q^m) \cdot q^m = P(q^c)$. It follows that $P(q^m) - P(q^c) = -P_q(q^m) \cdot q^m > 0$ and hence that $P(q^m) > P(q^c)$ which implies that $q^m > q^c$ since the inverse demand curve has a negative slope.

Figure 5.6.: Linear demand



3. Set up the monopolist's profit as a function of quantity!
4. Derive the monopolist's first-order condition for the optimal quantity!
5. Solve for the optimal quantity!
6. Find the monopoly price!

Hint: The monopoly price is given by $(a + c)/2$. ◀

5.4.2. Price as choice variable

Now let us derive exactly the same result, but in a slightly different manner, by highlighting the monopolist's choice of price instead of highlighting the quantity. Recall that the monopolist must decide both what price to set and how much to produce. There is no reason to produce more than can be sold at a given price. That would simply add to cost. Nor is there any reason to produce less. Then it would have been better to raise price. The optimal level of production is therefore simply given by the quantity demanded at the chosen price, i.e. $q = D(p)$. Only the optimal price remains to be chosen.

To find out what price the monopolist will charge, we need to study how profits depend on price. How is the monopolist's profit changed if the monopolist would increase its price slightly by dp? To answer this question, we focus on cost function $C(q) = F + c \cdot q$, and write the monopolist's profit as a function of price only, i.e.,

$$\pi(p) = (p - c) \cdot D(p) - F$$

by substituting in the demand function. Notice that the monopolist's profit varies with the price charged for two reasons. On the one hand, a higher price means that the firm will earn a higher mark-up, $p - c$. On the other hand, a higher price means that the monopolist will sell fewer units, $D(p)$.

The next step is to differentiate the profit function, to describe how profit changes, $d\pi$, if the monopolist changes its price by a small amount, dp:

$$d\pi = \pi_p(p) \cdot dp = D(p) \cdot dp + (p - c) \cdot D_p(p) \cdot dp.$$

This expression reveals that the fundamental trade-off in monopoly pricing is a trade-off between high sales and a high mark-up:

- The positive effect of increasing price ($dp > 0$) is that the monopolist earns a higher mark-up on the goods actually sold, i.e. $D(p) \cdot dp > 0$. This effect is captured by the first term on the right-hand side.

- The negative effect of increasing price is a loss of sales and the loss of the corresponding mark-up, i.e., $(p - c) \cdot D_p(p)\, dp < 0$. This effect is captured by the second term on the right-hand side.

The advantage with studying the differentiated profit function (rather than the profit function directly) is that it allows us to say something more about how large the two effects are.

Consider first the case when the price is very low and close to marginal cost, i.e. $p \approx c$. Then, a small price increase ($dp > 0$) is profitable, i.e. $\pi_p(c) > 0$.[6] The reason is that the loss due to the loss of sales is small (i.e. $(p-c) \cdot D_p(p) \approx 0$) while the gain from a higher mark-up is large since so much is sold at such a low price (i.e. $D(c) = \text{large}$). Thus, the monopolist does not want to charge price equal (or very close to) to marginal cost; at that level a price increase would be profitable.

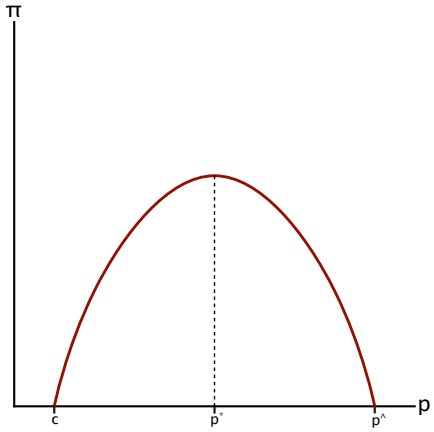
Consider next the case when the price is very high and equal to the price $\hat{p}$ where there is no demand, i.e. $D(\hat{p}) = 0$. Then, a small price reduction ($dp < 0$) is profitable, i.e. $\pi_p(c) < 0$. The reason is that a price reduction does not cost anything since nothing is sold anyway, i.e. $D(\hat{p}) = 0$. The gain from the additional sales (i.e. $(p-c) \cdot D_p(p) < 0$) may be substantial since the mark-up is high.

Clearly, the monopolist wants to charge a price somewhere in-between marginal cost and the price where no one would buy the good. But how do we determine exactly how high the optimal, i.e. profit-maximizing, price is? Think about a series of price increases starting from $p = c$. At first, such price increases lead to large increases in profit, as described by Figure 5.7. But, as price is increased further and further, the losses due to lower sales are increasing and the gains due to the increasing mark-up are falling. Eventually, the monopolist reaches the price, p^* , where the gains and losses from a price increase exactly offset each other, i.e. $\pi_p(p) = 0$. Additional price increases would lead to a net loss, i.e. $\pi_p(p) < 0$. Since profit is increasing in price at prices below p^* and decreasing in price at prices above p^*, it follows that p^* maximizes the monopolist's profit.

In the figure the (red) profit function reaches its highest point over p^*. Also notice that the profit function is increasing to the left and decreasing to the right of this price. Precisely at the optimal

[6] Recall that $\pi_p(c)$ is the first derivative of the profit function evaluated at the point $p = c$.

Figure 5.7.: Monopoly profit



price, however, the slope of the profit function is zero.

This reasoning shows that the profit-maximizing price can (sometimes) be found using the so-called first-order condition.[7] The first-order condition is $\pi_p(p) = 0$, i.e. the derivative (slope) of the profit function with respect to price set equal to zero. In particular, the first-order condition is given by

$$D(p) + (p - c) \cdot D_p(p) = 0.$$

Note that the first-order condition is an equation in one variable, namely the price, p. It is sometimes possible to solve the equation for p, to find an explicit expression for the optimal price as a function of all the parameters such as the marginal cost, c.

Exercise 17. Consider a monopolist with constant marginal cost $c > 0$. Assume that inverse demand is linear and given by $P(Q) = a - b \cdot Q$, where $a > c$ and $b > 0$, as described in Figure 5.6.

1. Derive the (direct) demand function, $D(p)$!

[7] The first-order condition can be used when, as assumed here, profit is first increasing and then decreasing in price, as depicted in the figure above. This is true e.g. if $D_p(p)$ is a constant, i.e. the same for all prices. This is the same as saying that demand is linear.

2. What is the slope of the demand function, $D_p(p)$?
3. Set up the monopolist's profit as a function of price!
4. Derive the monopolist's first-order condition for the optimal price!
5. Solve for the optimal price!
6. Find the monopoly quantity!
7. Compare to the solution when the monopolist chooses quantity (see above exercise)!

5.5. The determinants of price

In a perfectly competitive market, the firms know that there is a price at which they can sell whatever quantity they wish. They also know that if they would try to charge a price higher than this *competitive price* all their customers would spend their money elsewhere. The firms maximize their profits simply by adjusting their production and sales to equate their marginal cost to the competitive price. Moreover, if the existing firms would be able to earn a profit at the competitive price, new firms would enter the market. The new firms have access to the same production technology and they can buy inputs at the same prices as the incumbents. As a result of the increase in production, the competitive price would fall. And it would continue to fall until profits are eliminated and no additional firm would enter the market. That profits are eliminated must mean that the price is equal to the average cost of production. A more detailed analysis reveals that price is equal to the so-called minimum long-run average cost of production. Expressed differently, in competitive markets, prices are determined

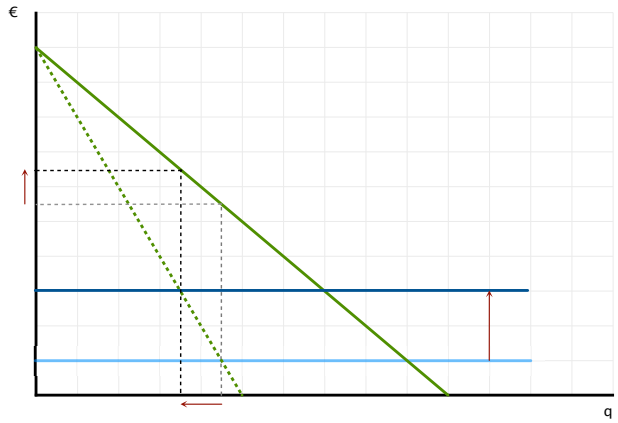
by the cost of production only, at least in the long run.[8]

While cost is a determinant of price also in monopoly markets, it is not the only one. But, let us start with cost.

5.5.1. Cost

Marginal cost Figure 5.8 shows that when it becomes more costly

Figure 5.8.: Increased marginal cost



to produce the good, the monopolist will produce less and increase its price. In this example, when the marginal cost is increased by €2, the monopolist increases its price by €1. We say that pass-through is a half, since the monopolist increases the price by a half of the increase in marginal cost. Symmetrically, if the monopolist's marginal cost is reduced, the monopolist also reduces its price. With a linear demand, a reduction of cost by €1, leads to a reduction in price by a half.

[8] In the short run, an increase in demand would lead to an increased price. But this price hike would be eliminated by the entry of new firms. Demand thus only affects the quantity traded at the competitive price, but not the price itself. However, if the firms make use of some specialized input, also the price of this input would increase as a result of the entry and increased production. That would imply that the minimum average cost of production actually depends on the level of production, which would imply that also the level of demand would affect the competitive price.

Also the monopolist's first-order condition,

$$D(p) + (p - c) \cdot D_p(p) = 0,$$

reveals the same relation between the monopolist's marginal cost and its behavior. The first-order condition is an equation in one unknown, namely the price, p. As the first order condition is written here, it has one exogenous parameter, namely the firm's constant marginal cost, c. If c would change, the monopolist would have to change p to satisfy the first order condition. And along with the change in price, also the quantity would have to be adjusted. Expressed differently, the firm's marginal cost matters for what price the monopolist will charge and also for the how much is will produce and sell. We say that the first-order condition defines price as an implicit function of the marginal cost.

To describe how price changes as a result of a change in the marginal cost, we can totally differentiate the first order condition, i.e.

$$2 \cdot D_p(p) \cdot dp + (p - c) \cdot D_{pp}(p) \cdot dp - D_p(p) \cdot dc = 0.$$

The total differential may be written as

$$[2 \cdot D_p(p) + (p - c) \cdot D_{pp}(p)] \cdot dp = D_p(p) \cdot dc.$$

Notice that the expression within the square brackets on the left hand side is negative. That follows from checking the second-order condition for maximization. Rewriting this equation shows how price depends on the constant marginal cost:

$$\frac{dp}{dc} = \frac{1}{2 + \frac{p-c}{p} \cdot \frac{D_{pp}(p) \cdot p}{D_p(p)}} > 0. \tag{5.2}$$

Exercise 18. Consider a monopolist with constant marginal cost $c > 0$. Assume that inverse demand is linear and given by $P(Q) =$

$a - b \cdot Q$, where $a > c$ and $b > 0$, as described in Figure 5.6.

1. By how much would the monopolist increase price if marginal cost is increased by €1?

2. Would it have been possible to draw this conclusion quickly by inspecting the more general expression 5.2?

Hint: If the demand function is linear, then $D_{pp}(p) = 0$. ◀

Expressed differently, the monopoly pass-through rate is always a half, whenever the demand function is linear. In other cases, the pass-through rate may actually be close to zero. And, in yet other cases, the the monopolist's pass-on rate may be higher than one. That is, the monopolist may increase its price by more than the increase in cost. The pass-on rate depends on the shape of the demand function.

Fixed cost

Exercise 19. What happens to the monopoly price if fixed cost, F, is increased?

Answer: Notice that the firm's optimality condition does not depend on the fixed cost. This means that the monopolist will not change its price or quantity. ◀

That price is unaffected by the fixed cost is probably surprising to most people. But, if the fixed cost would be increased by so much that the profit would actually turn negative, then the firm would actually discontinue its operations, at least in the long run, when the fixed cost can be avoided.

A reduction in the fixed cost can also affect the market price in a more indirect way in the long run. A lower fixed cost may actually lead to entry into a market and, thus, to increased competition. But this is a topic saved to another chapter.

5.5.2. Demand

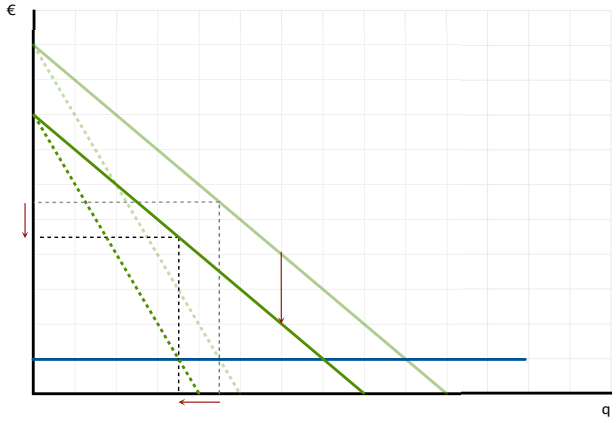
To study how changes in demand affects a monopolist is more complex. A quick glance at the first-order condition

$$D(p) + (p - c) \cdot D_p(p) = 0,$$

shows that both the level of demand, D, and the consumers' price sensitivity, D_p, matters.

Level of demand We start out by focusing on a linear demand function. Figure 5.9 shows that if the intercept of a linear demand

Figure 5.9.: Willingness to pay



function is lowered, the monopolist will lower its price and also reduce production. Such a reduction in the intercept of a linear demand function can be described as a uniform reduction in the quantity demanded (i.e. the same reduction in quantity demanded at every price). But it can also be described as a uniform reduction in the consumers' willingness to pay for the good. In particular, when the consumers' willingness to pay is (uniformly) reduced by €2, the monopolist will reduce its price by €1.[9]

[9]One may in fact demonstrate that such *uniform* reductions in demand always lead to a lower price. To see this, think of the demand function

Price sensitivity To highlight the role of the consumers' price-sensitivity, it is common to rewrite monopolist's first-order condition as

$$\frac{p-c}{p} = -\frac{D(p)}{D_p(p) \cdot p}.$$

Exercise 20. Rewrite the first-order condition into this form!

Hint: The above equation is derived from the first-order condition from the monopolist's choice of price. The same formula expressed in terms of the indirect demand function can be derived from the monopolist's choice of quantity. ◀

To interpret this revised formulation of the monopolist's first order condition, two definitions are required. First, recall that the price elasticity of demand is a commonly used measure of how responsive the consumers are to price increases. The elasticity answers the question: "If price is increased by one per cent, by how many per cent is quantity demanded reduced?" More formally the price elasticity is defined as

$$\eta \equiv -\frac{\left(\frac{dq}{q}\right)}{\left(\frac{dp}{p}\right)} = -\frac{p}{q} \cdot \frac{dq}{dp}.$$

Since quantity demanded is given by the demand function (i.e., $q = D(p)$) and since the change in quantity demanded is given by the first derivative of the demand function (i.e., $dq/dp = D_p(p)$), the price elasticity of demand can also be expressed as $\eta(p) \equiv -p \cdot D_p(p) / D(p)$. Note that the right hand side of the monopolist's

$\delta(p)$ and then uniformly add ε units to the demand function, at all levels of price, to obtain the new demand function $D(p) = \delta(p) + \varepsilon$. Then the monopolist's profit is given by $\pi(p) = (p-c) \cdot [\delta(p) + \varepsilon]$ and the first-order condition is $\delta(p) + \varepsilon + (p-c) \cdot \delta_p(p) = 0$. The change in price resulting from a change in ε is computed from a total differentiation $\frac{dp}{d\varepsilon} = -1/[2 \cdot \delta_p(p) + (p-c) \cdot \delta_{pp}(p)] > 0$. Recall that the expression within the square brackets is negative.

first order condition happens to coincide with the inverse of the price elasticity of demand (i.e., $1/\eta$).

Second, a key concept in economics is *market power*. Market power is defined as the firm's ability to set a price above marginal cost (i.e. a positive mark-up) without loosing all customers. A firm's market power is often measured by the so-called Lerner index,

$$L \equiv \frac{p-c}{p},$$

i.e. the mark-up as a percentage of price. If price is equal to marginal cost, the Lerner index is given by $L = 0$, which indicates the complete absence of market power. If p is larger than c, the Lerner index will be positive. If p would be much larger than c, the Lerner index would be close to 1, indicating very high market power. Note that the left hand side of the (revised version of the) monopolist's first order condition happens to coincide with the Lerner index.

Using these two definitions, we can rewrite the first-order condition as:

$$L = 1/\eta\,(p)$$

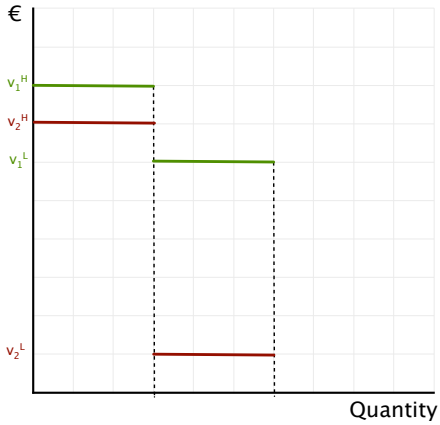
which is the so-called inverse-elasticity rule. It shows that market power in a monopoly market is determined by how price sensitive the consumers are. In particular:

Conclusion 2. The monopolist has more market power, the lower is the price elasticity of demand.

That is, if the consumers are reluctant to leave the market and buy something else when the price is increased, the monopolist will be able to set a high mark-up. On the other hand, if consumers are highly price-sensitive, then the a lack of competitors probably does not have serious consequences. But only if the consumers are extremely price sensitive ($\eta \to \infty$) will the monopolist have almost no market power and charge a price close to marginal cost.

More about the importance of price sensitivity A somewhat surprising implication of the importance of price-sensitivity is that price does not need to be higher in a market with higher demand, if the consumers are more price-sensitive in the market with high demand. This point is illustrated in Figure 5.10. In Country 1

Figure 5.10.: "The slope effect"



(marked by green) consumers are willing to pay a very high price, v_1^H, for the first unit of the good and a relatively high price, v_1^L , also for the second unit. Demand is high in Country 1: consumers buy the good even if price is high. But also note that the consumers in Country 1 are *price-sensitive*: a small reduction in price from v_1^H to v_1^L is sufficient to increase sales from one to two units.

In Country 2 (marked by red) consumers are willing to pay a relatively high price, v_2^H, for one unit of the good but only a very low price, v_2^L, for the second unit. Demand in Country 2 is clearly lower in than in Country 1. (Consumers have a lower willingness to pay for both the first and the second unit. Expressed differently, at all possible prices, people buy fewer or the same number of units as the consumers in Country 1.) But demand in Country 1 is also less price-sensitive. The monopolist would have to lower price by a large amount to motivate the consumers to buy a second unit.

It is easy to verify that the monopolist would charge a lower price in the market with high demand, Country 1, than in the market with low demand, Country 2. Assume that the marginal cost is zero. Then, the monopolist charges $p_1 = v_1^L$ in Country 1 if $2 \cdot v_1^L > v_1^H$. The monopolist charges $p_2 = v_2^H$ in Country 2 if $2 \cdot v_2^L < v_2^H$. Thus, $p_2 = v_2^H > v_1^L = p_1$.

5.6. Welfare

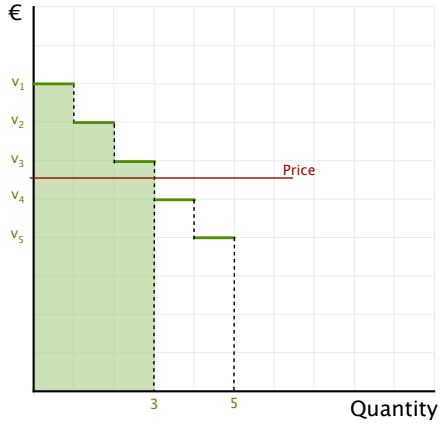
5.6.1. Consumer surplus

Consumers derive utility from consuming the goods they buy. It is common to measure this utility in monetary terms by the so-called *consumer surplus*. That is to say, it is common to measure utility in e.g. Euros or Dollars. The *gross* consumers' surplus is defined as the consumers' own valuation of the goods they consume. More precisely, the consumers' valuation is equal to their willingness to pay for the goods they consume.

Under certain conditions, the gross consumers' surplus is given by the area under the (Marshalian) demand curve.[10] To see that the consumers' welfare (or utility) measured in monetary terms is equal to the area under the demand curve, it is best to use a simple example. Assume that each consumer either consumes one or no unit of the good. Let's say that consumers are numbered in order of their valuations of the good (or willingness to pay), so that $v_1 > v_2 > ... > v_n$. With 5 consumers demand is described by the green line in Figure 5.11. If price is set between the third and the fourth highest valuations, three units of the good will be sold and the gross consumer surplus is given by $CS^G(3) = v_1 + v_2 + v_3$, which is equal to the green colored area in the figure, i.e. the area under the demand curve up to three units.

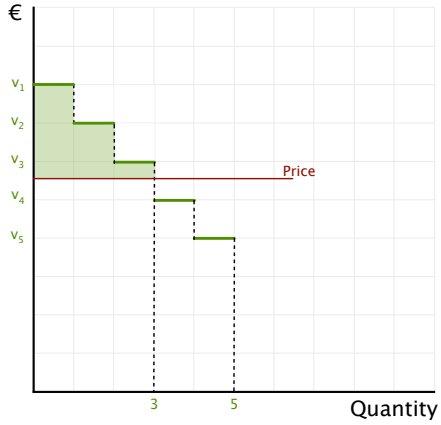
[10] This is exactly true if demand is independent of income and it is a good approximation if the income elasticity of demand is small.

Figure 5.11.: Gross consumer surplus



The *net* consumers' surplus is defined as the consumers' willingness to pay for the goods they buy minus the amount they actually have to pay, that is $CS^N(3) = v_1 + v_2 + v_3 - 3 \cdot p$. The net consumers' surplus is illustrated by the area below the demand curve, above the price line in Figure 5.12.

Figure 5.12.: Net consumer surplus



More generally the gross consumer surplus derived from consum-

ing q units is given by the sum of the values of these units

$$CS^G(q) = \sum_{i=1}^{q} v_i$$

Recall that the consumers' valuations of the good is given by the indirect market demand function, i.e. $v_1 = P(1), ..., v_q = P(q)$. When the units are infinitely small, integration replaces normal sums, and the gross consumer surplus can be written

$$CS^G(q) = \int_0^q P(x) \cdot dx.$$

As already mentioned, the net consumer surplus is defined as the difference between what consumers are willing to pay and what they actually pay, i.e.

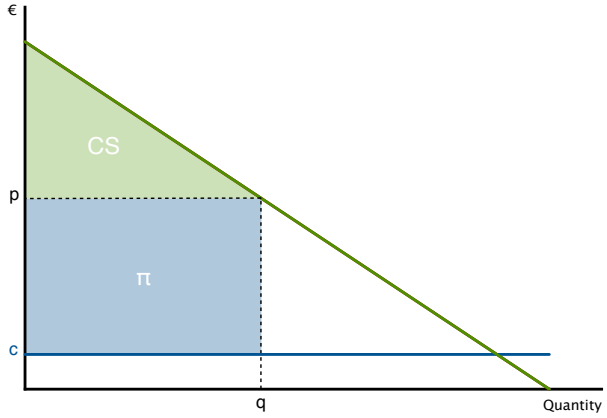
$$CS^N(q) = \int_0^q P(x) \cdot dx - P(q) \cdot q.$$

If demand is income-elastic, the consumers' utility is still defined as their willingness to pay for the goods they consume minus what they actually have to pay. But in this case, consumer utility is not equal to the consumers' surplus (the area under the Marshalian demand curve). Instead one would have to work with the so-called equivalent variation or compensating variation. But that will not be necessary here.

5.6.2. Profits

Not only the consumers derive a surplus from being active in the market. The firm derives a profit. For example, if the firm has constant marginal cost the profit is given by $\pi(q) = [P(q) - c] \cdot q - F$. The variable part of the profit, i.e. the first term of the profit, is described by the blue rectangle in Figure 5.13. The height of the rectangle is given by the mark-up and the length of the rectangle is

Figure 5.13.: Welfare



given by the quantity produced. The area of the rectangle is simply the height times the length, i.e. mark-up times quantity.

5.6.3. Social welfare

It is common to measure the social welfare created in the market, by the so-called *total welfare*, which is defined as the sum of the net consumer surplus and the firm(s') profits, i.e.

$$W(q) = CS^N(q) + \Pi(q).$$

The consumers' surplus and profits can be added since they have the same unit of measurement. This is one of the reasons why it is so convenient to measure consumer utility in monetary terms.

The key ethical assumption behind this measure of social welfare is that the distribution of welfare between consumers and the firm is unimportant. If profits are increased by one Euro at the same time as the consumer surplus is decreased by one Euro, social welfare is unchanged.[11]

[11] To give consumers and firms equal weight in the welfare function is by no means the only possibility. For example, the norm among competition authorities around the world is to only care about the consumers' welfare.

It is easy to see that social welfare, defined this way, is also equal to the gross utility of consuming the goods minus the cost of production, that is

$$W(q) = CS^G(q) - C(q)$$

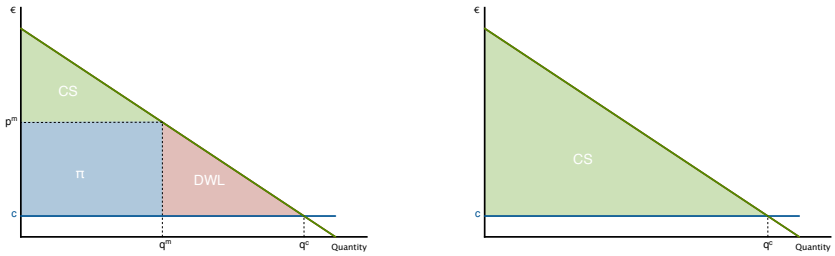
since the consumers' expenses and the firm's revenues cancel. Again, the reason why the payments from the consumers to the firms do not matter, is that we simply have decided to disregard them.

Exercise 21. Rewrite social welfare defined in terms of the net consumers' surplus and profits as a function of the gross consumer's utility and production costs.

5.6.4. The dead-weight loss

A key result in economics is that monopoly pricing reduces welfare compared to the competitive situation, when price is equal to marginal cost and q^c is consumed. The difference in welfare between the competitive outcome and the monopoly outcome is called the dead weight loss. In Figure 5.14 it is indicated by the red trian-

Figure 5.14.: Welfare



(a) Monopoly (b) Competition

gle denoted DWL. There are some consumers who wont buy the good despite being willing to pay more for the good than it costs to produce. This is a loss to society.

Then, the social welfare function would simply coincide with the net consumers' surplus.

Exercise 22. Consider a monopolist with constant marginal cost $c > 0$. Assume that inverse demand is linear and given by $P(Q) = a - b \cdot Q$, where $a > c$ and $b > 0$, as described in Figure 5.6.

1. How much would it cost the monopolist to product one unit more?
2. How much would the consumers value the consumption of one unit more?
3. How much would the monopolist's revenues increase if one unit more is sold?
4. Compute the dead weight loss.

5.7. Third-degree price discrimination

One of the main conclusions from the monopoly model is that in monopoly markets, price depends not only on the cost of production but also on demand. A monopolist sets a higher price, the more the consumers are willing to pay and the less price-sensitive they are. It follows that a monopoly firm will set different prices for the same product in different geographical regions, if the demand for the product differs between the regions. This practice is often called second-degree price discrimination.

It is possible that the differences in the prices of on-patent medicines between different countries (see page 45), at least partly, is a result of differences in the willingness-to-pay between different countries. Remember that cost differences is an unlikely explanation since the medicines often are produced in the same factory and that transportation costs are often low compared to the price at which the medicines sell. What is striking is that the countries with the lowest prices are Greece, Spain and Portugal and that the countries with the highest prices are Switzerland, U.K. and Germany. It seems

reasonable to assume that the latter group is willing to pay more than the former as a result of a higher income. The pattern is not that simple, however. That a rich country like Norway is on the European average reveals that also other aspects matter. An example is the various forms of regulations in the pharmaceutical markets (which is discussed below).

An important restriction on a firm's ability to charge different prices in different countries is that consumers sometimes can travel to make their purchases where the prices are low. In the pharmaceutical market this may be difficult for most people, since the price difference of many medicines are too small to pay the ticket, or because they are too ill to travel. It is probably also more difficult for most people to get a prescription in other countries than their own.

Within Europe, there are, however, firms that specialize in buying medicines where they are cheap and then sell them where they are expensive. This practice, called *parallel import*, is possible as a result of the European Union's Internal Market Program. The Internal Market Program stipulates that all goods and services can move freely in the Union. But even if parallel import is a whole industry in Europe, the price data clearly reveal that such parallel import is not enough to completely wipe out price differences.[12]

Welfare effects of second-degree price discrimination Second-degree price discrimination means that a firm charges different prices in different geographical markets. A monopolist, which is not constrained by any arbitrage (parallel trade), sets a markup in each market, related to the price-sensitivity of the consumers in

[12] In north america, the prices of medicines are much lower in Canada than they are in the U.S. There is nevertheless no organized parallel trade in medicines from Canada to the U.S. since such trade is illegal. The reason why the U.S. government does not allow parallel imports from Canada is that it would reduce the pharmaceutical companies ability to price discriminate and to earn a profit well-aligned with the value of the products that the companies have invented.

that market. With two countries A and B, characterized by the demand-elasticities $\eta_A < 0$ and $\eta_B < 0$, the prices are characterized by

$$\frac{p_A - c}{p_A} = -\frac{1}{\eta_A}$$
$$\frac{p_B - c}{p_B} = -\frac{1}{\eta_B}.$$

If the monopolist would not be allowed to charge different prices in different countries, the price would be set somewhere in-between the two country-specific monopoly prices. In particular, one may show that, a monopolist that must charge the same price would set

$$\frac{p - c}{p} = -\frac{1}{\eta_A \cdot s_A + \eta_B \cdot s_B}$$

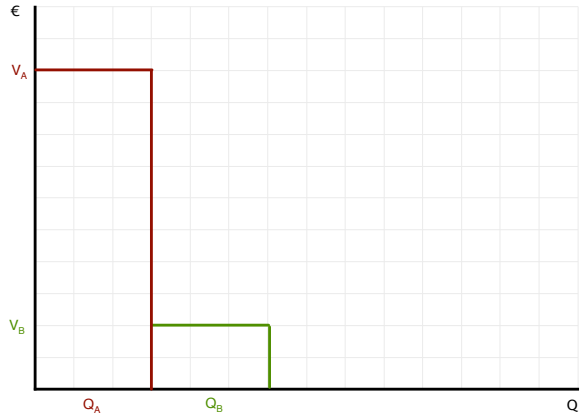
where $s_A = \frac{q_A}{q_A + q_B}$ and $s_B = 1 - s_A$ reflect the relative size of the two countries. Expressed differently, second degree price discrimination is beneficial for the consumers in the country where consumers are relatively price sensitive and harmful for the consumers in the country where the consumers are relatively insensitive to prices.

Exercise 23. Consider a monopolist selling the same good in two countries A and B. Everyone citizen in each country wants to have one unit of the good. In country A everyone is willing to pay V_A for a unit and in country B everyone is willing to pay V_B. It is assumed that V_A is much higher than V_B. Nobody is interested in a second unit. There are equally many people in both countries, $Q_A = Q_B$. For simplicity the marginal cost is equal to zero. This market is described in the Figure.

1. Assume that arbitrage is impossible between the two countries. What prices and quantities will the monopolist choose?

2. Assume that arbitrage is costless and that the monopolist must charge the same price in both countries. What price and quantities will the monopolist choose?

Figure 5.15.: Price discrimination, example



3. What happens to global output and global welfare if the monopolist can price-discriminate?

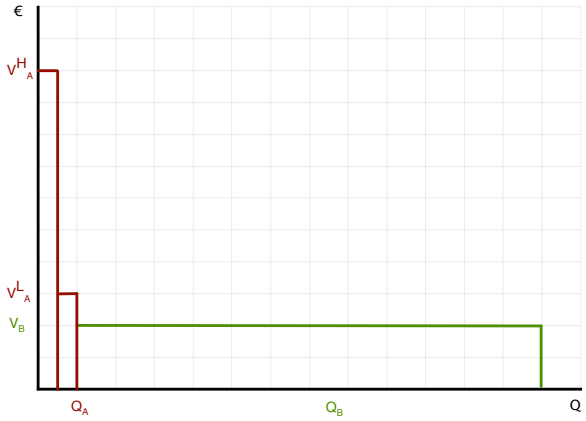
Exercise 24. Consider a monopolist selling the same good in two countries A and B. Everyone citizen in each country wants to have one unit of the good. In country A half of the population is willing to pay V_A^H for a unit while the others are willing to pay a much lower amount V_A^L. In country B everyone is willing to pay V_B which is slightly lower than V_A^L. There are many more people in country B. For simplicity the marginal cost is equal to zero. This market is described in the Figure.

1. Assume that arbitrage is impossible between the two countries. What prices and quantities will the monopolist choose?

2. Assume that arbitrage is costless and that the monopolist must charge the same price in both countries. What price and quantities will the monopolist choose?

3. What happens to global output and global welfare if the monopolist can price-discriminate?

Figure 5.16.: Price discrimination, example



5.8. Pros and cons of price regulation

Monopolists create inefficiencies. By raising their prices above their marginal costs, monopoly firms inflict losses to society. Goods and services that would have been more valuable to consume than they cost to produce never reach the market. The most obvious political remedy would be to regulate prices, to force monopoly firms to charge prices based on their costs, and not on their consumers' willingness to pay. Such regulation is, however, the exception rather than the rule.

Current regulations The european competition rules prohibits so-called *abuses of dominant positions*.[13] The first example of such an abuse is "imposing unfair purchase or selling prices". Any European who believes that he has to pay too much for some good can thus report this expected abuse to the relevant european or national authority. In Sweden, this would be Konkurrensverket, the Swedish competition authority. The Swedish competition authority

[13]This prohibition is currently (February 2016) called Article 102 of the Treaty on the Functioning of the European Union. Corresponding provisions are also included in the member states' national laws. In Sweden, the counterpart is included in Konkurrenslagen (the competition act).

can then investigate the allegations and, if appropriate, bring the case to court. If the firm would be found guilty of charging unfair prices, the court would impose fines and damages. In practice, however, the prohibition of abuse of dominance has almost never been used to eliminate "unfair prices." This particular provision may even be obsolete. That is, in practice, also monopoly firms are free to set whatever prices they wish.

There are, however, also special laws regulating prices in some specific markets. But these markets are characterized by specific problems in addition to a lack of competition. One examples is the telecom markets. The European rules (in Sweden implemented by the Electronic Communications Act/Lagen om elektronisk kommunikation) stipulates that telecom firms must charge cost-based prices when selling "interconnection services" to one another. Interconnection means that Telia will put through calls from Telenor customers. One of the particularities with this market is that the telecom firms, by the same law, are obliged to provide one another with interconnection services. The normal market mechanism is thus short-circuited. When telecom firms sell their services to the final users, that is you and I, they are free to set their own prices. Another example of a market with some elements of price regulation is pharmaceuticals, which will be described in some more detail below.

But what is the reason for this relaxed attitude among the competition and regulatory authorities?

Problems with price regulations That markets fail to deliver efficient outcomes is not sufficient for instituting corrective measures. Regulations come with their own costs. The most immediate problems of price regulation are five.

To eliminate the dead weight losses, prices should be set equal to marginal costs. Many markets are, however, characterized by huge investments and negligible variable costs. The cost of rolling

out a network for mobile communication is staggering. But once the network is in place, the additional cost of a call is even hard to measure. Requiring telecom firms to set prices equal to their marginal costs would therefore quickly destroy the whole market. The fixed costs for the networks must be financed somehow. One possibility is to allow the firms to charge prices above their marginal costs, thereby accepting the dead weight losses. Another possibility would be to cover the fixed costs via subsidies financed by taxes in other markets. But collecting taxes in other markets would impose a wedge between the consumer price and production cost in those other markets. The dead weight loss would simply be moved from one market to another. And paying out subsidies to cover any fixed cost claim would eliminate the firms incentives to keep their costs low and also create incentives to claim that costs are higher than they really are.

Furthermore, an effective regulation requiring firms to set prices equal to their (marginal or average) cost requires authorities to actually know what these costs are, which they do not. If costs are over-estimated, there will still be dead-weight losses. If costs are under-estimated, the firms may even be forced into bankruptcy. And, if firms are allowed to raise their prices as soon as their costs are increased, their incentives to keep costs low and to report them truthfully are compromised.

Perhaps the most harmful result of price regulation is that incentives for innovation are put to sleep. One of the key drivers for inventing new products is presumably the ensuing monetary rewards for the inventor. To focus inventive resources on the projects creating the highest value to consumers, the inventors' gains of success must be metered out in proportion to the value created. But that is just a complicated way of saying that the prices of newly invented products must be allowed to reflect the consumers' willingness to pay. Market power is often defined as an ability to charge a price above cost without losing all customers. A more precise

definition might be that it is an ability to set prices reflecting the buyers' willingness to pay. Market power thus fills an important positive role in a dynamic economy. This is the economic rationale for various policies of granting exclusive rights, e.g. by patents.

To allow sensible prices (e.g. with respect to innovation incentives and the coverage of fixed costs) any price regulation must be rather complex. It will therefore be difficult for regulated firms to say what prices they are allowed to charge. Price regulation creates risks. A case in point is that the Swedish Market Court (the appellate body in competition policy cases) finally determined what prices TeliaSonera was allowed to charge for certain services delivered during the fall of 2000 in April 2013. Expressed differently, TeliaSonera and its customers, had to wait 12 years after delivery to find out what the price was.[14]

Price regulations also comes with administrative costs. To regulate price, there must be an authority with officials who can collect and interpret information about e.g. cost conditions. The regulated firms must, themselves, collect, compile and supply the authorities with the relevant information. And, for reasons of legal certainty, the firms must be given the right to appeal every decision to reduce their price. And that is exactly what they do. As the TeliaSonera case illustrates, some such cases may drag on for years due to their complexity. During this time clever business lawyers work thousands of hours. Judging from the verdict, TeliaSonera's cost for external legal advice in this case must have been at least a million Euro. To this should be added TeliaSonera's internal costs, the Swedish competition authority's work, as well as the work done by the complaining firms and by the courts Sweden (as well as in Europe, where some of the issues were settled).

[14]Marknadsdomstolens dom i mål nr A 8/11, 2013-04-12.

5.9. Case Study: Value-Based pricing of pharmaceuticals

Pharmaceutical pricing is a jumble of dilemmas, one of which is the conflict between providing incentives for developing new drugs and the proper use of those that already exist.

Developing new medicines requires huge investments in research and development (R&D). To provide the pharmaceutical companies with appropriate incentives to undertake such investments, they are granted patents when they succeed to develop a new medicine. Thus, the company that develops a new medicine is granted an exclusive right to produce and sell that medicine for a period of 20 years. During this period the company should be able to earn a profit which is related to the value created.

The practice of granting patents clearly comes at a cost to society, namely the dead-weight loss. Not all the people who could have benefitted from the existing medicines will get to do so. As the costs of producing medicines are often low, once the huge R&D costs have been taken, the wedge between price (the consumption value) and cost is probably often substantial. Expressed differently, one may suspect that the dead-weight loss is substantial.

Effect of subsidy In part for this reason, medicines are heavily subsidized. On average Swedes pay approximately 20 percent of the price for subscribed medicines directly out of their pockets. The rest we pay via taxes.

Exercise 25. Consider a monopolist with constant marginal cost $c > 0$. Assume that demand is linear and given by $q = a - p$, where $a > c$. The product is subsidized, meaning that the consumers only play the price $p = \lambda \cdot r$ where $\lambda \leq 1$ and r is the producer price (i.e. the price that the producer receives per unit sold).

1. Set up the monopolist's profit as a function of the producer

price.

2. Compute the monopoly consumer and producer price in case there is no subsidy (i.e. $\lambda = 1$) assuming $a = 100$ and $c = 1$.

3. Compute the monopoly consumer and producer price in case consumers only have to pay 20 percent of the producer price (i.e. $\lambda = 0.2$) assuming $a = 100$ and $c = 1$.

4. What is the effect of the subsidy on the producer price, the consumer price and consumption?

This example reveals the risk of combining a subsidy with free pricing in a market with market power. The subsidy which is meant to provide the citizens with an insurance making it possible for us to buy the medicines we need when we fall ill, is in fact turned into a gift to the medical companies instead. The industry gets a subsidy. Sick people don't get an insurance.

For this reason the pharmaceutical companies are not allowed to set whatever prices they wish if they want to sell their products in Sweden. More precisely, they are not allowed to set whatever price they like if they wish to be included in the subsidy scheme. In some countries, including Sweden and in the U.K., medicines that are still protected by a patent must adhere to so-called value-based pricing.[15]

Value-based pricing For safety reasons, medicines must be approved by the Swedish Medical Products Agency (Läkemedelsverket) or the European Commission in order to be sold in Sweden. Once approved, the producer can apply to be included in the pharmaceutical benefits scheme. Being included in the subsidy is clearly

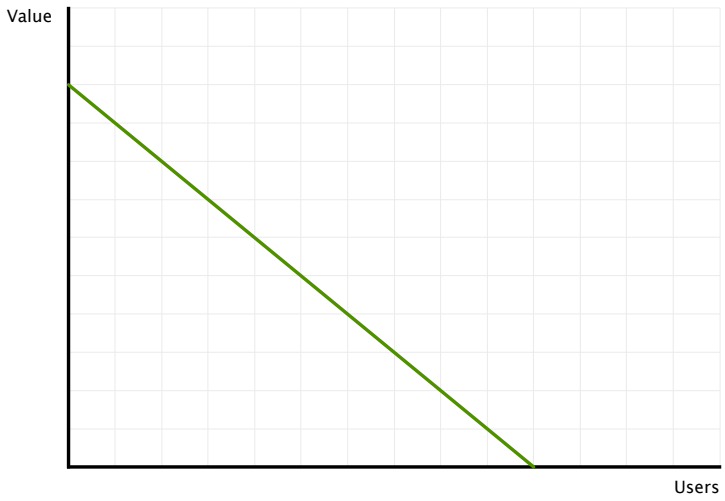
[15] Other countries use other pricing principles for patented medicines. For example, Norway uses so-called external reference pricing, which means that the price of a medicine must not exceed the price of the same medicine sold in certain other countries.

very attractive as it means that the customers will only have to pay a low share of the full price.

To be included in the subsidy, an application must be sent to the the Dental and Pharmaceutical Benefits Agency (Tandvårds- och läkemedelsförmånsverket, TLV). The application includes a considerable amount of information about the effects of the new medicine. The purpose is to establish the value of the new medicine for different patients with different diagnoses and for different patients who may suffer from different side effects.

Clearly, the value of the medicine may be higher for people with serious conditions than for people with less severe conditions. The value is also higher for people who do not suffer from any negative side effects. Ordering different (potential) users from the highest to lowest value creates a "value-curve" such as the one in Figure 5.17.

Figure 5.17.: Valuation of medicines



It is the pharmaceutical companies that must produce the research necessary to prove that their products indeed have the claimed positive effects. They may e.g. have to prove to TLV's medical ex-

perts how many additional years a new cancer drug will give for various types of patients. Or they may have to prove by how much a drug against obesity reduces weight for people with different BMI. All these positive effects are then translated into a an overall measure called QALY which is short for Quality Adjusted Life Years. A medicine that allows a patient to live one extra year in perfect health is said to give 1 QALY. An additional year with lower health gives a lower QALY. Also medicines that increase the quality of life, but have no effect on the expected life span, are given some QALY-value. Finally, to translate the health effects into a monetary value, the QALY of a medicine is multiplied by an (approximate) value per QALY. While it may vary from case to case, the value of one QALY in Sweden is probably around €50000. The value of a QALY is the same for every person with the same condition, independent of that person's income.

Once the pharmaceutical company has succeed to convince the TLV experts about the value of the new medicine, it will state the price at which it wants to sell its product. TLV then decides which groups of potential users that will be included in the subsidy. The rule is to include the potential users for which the value of the medicine is higher than the price, as shown in Figure 5.18. If the new medicine provides a treatment for obesity, the potential users could for example be defined in terms of their BMI. And TLV's decision may then be to subsidize the use of the new medicine by all patients who have a BMI higher than a certain threshold value.

Exercise 26. Consider a pharmaceutical company with constant marginal cost $c > 0$. Assume that TLV's valuation of the new medicine for the q'th potential user is given by $v = a - q$, where $a > c$. That is, the value of the medicine for the first user is a and the value for the 10th potential user is $a - 10$.

1. Which potential users will be subsidized if the company wants to charge a price of p.

Figure 5.18.: Valuation of medicines

2. Assume that all patients who get their medicines at the subsidized price will buy the medicine but that all patients who are not allowed the subsidy refrain from buying it. What price will the medical company charge?

Value-based pricing is thus akin to normal monopoly pricing. The idea is to allow the companies with patent-protected medicines to charge a "normal" monopoly price related to the value of the product. This is meant to provide appropriate incentives for R&D. The companies are not, however, allowed to grab the subsidy.

5.10. Exercises

Basics

List the most important barriers of entry.

Short questions

1. Why are monopoly firms selling fewer units than would be sold in a competitive market?
2. Define market power.
3. How is market power measured?
4. How is the price-elasticity of demand defined?
5. What does the inverse-elasticity rule tell us?
6. What is meant by the gross and net consumers' surplus?
7. What is the relations between the consumers' surplus and the demand function?
8. What is meant by the dead-weight loss?

Monopoly pricing

Consider a monopoly market where each individual buyer has demand function $q = p^{-\eta}$ where $\eta > 1$. The cost of production is given by $C = F + c \cdot q$.

1. Derive the market demand function if there are N buyers.
2. Compute the elasticity of market demand.
3. What price will the monopolist charge?
4. How is price and quantity changed if marginal cost is increased?
5. How is price and quantity changed if the fixed cost is increased?
6. How is price and quantity changed if more people move in to the market?

7. How much market power, as measured by the Learner index, does the monopolist have?

Discussion

1. What are the pros and cons of the European Union's Internal Market Program in the context of the pharmaceutical market?

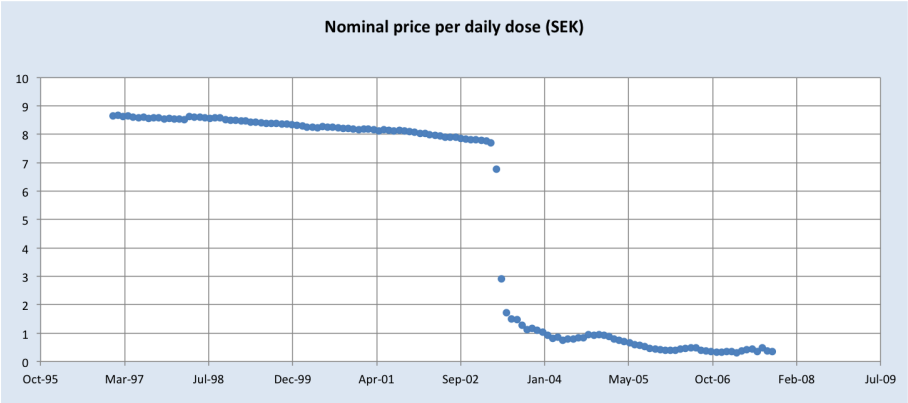
2. What are the pros and cons of value-based pricing of pharmaceuticals?

6. Oligopoly

6.1. Why study oligopoly?

In April 2003 Merck & Co's cholesterol-reducing drug Zocord went off patent in Sweden. Soon after, other companies started to produce and sell perfect copies, containing exactly the same active ingredient Simvastatin. Within a few months Merck slashed the price down to a quarter of its original level, from approximately € .8 SEK to approximately € .2 per daily dose. This remarkable development is described in Figure 6.1.

Figure 6.1.: Price of Zocord in Sweden



A patient using Zocord on a daily basis thus saves more than two hundred Euros per year, as a result of competition. Patients who have reached the ceiling for yearly medical costs under the public medical insurance scheme don't gain directly, but the public budget gains in their place.

Although this example is extreme, it clearly demonstrates what a radical effect competition can have, at least in some markets. In this chapter we will investigate how price competition works. We will also start addressing the question what determines the strength of price competition in different markets? Why did the expiry of the patent have such a dramatic effect in the market for Simvastatin? Could we expect the same effects in other markets? Finally, we will study what the effects of competition on social welfare are.

To characterize price competition we will use a simple model of a market and compare monopoly with oligopoly. How does the outcome in a given market differ depending on whether there is just one monopoly firm serving the market or if there are two or more firms competing to serve the market? To study competition, we will focus on oligopolies, i.e. markets with only a few sellers. We will not study perfectly competitive markets, since they are quite rare in reality. Studying oligopoly is more complicated than studying monopoly, however. In oligopoly markets the optimal pricing and output decision by one firm depends on the actions taken by their rivals. To predict firm behavior, we will need to formulate the market interaction as a normal form game. This is a good time to revise your game theory skills. An introduction is available in Part VIII on page 427.

Actually, there are several different oligopoly models. Part of the reason is that different oligopoly markets are organized in different ways. In some markets, such as the market for generic drugs, the firms set their prices and sell the amount demanded at that price. Such markets can be described by the so-called Bertrand model (discussed in section 6.2 on the next page). In other markets, such as the wholesale market for fish, the firms first decide how much to sell and then a market clearing price is determined through a centralized auction. Such markets can be described by the so-called Cournot model (discussed in section 6.4 on page 102).

6.2. Bertrand Oligopoly

Basic market data Consider the monopoly market discussed above. Assume now that an additional firm has entered the market, turning it into a duopoly. The firms are called 1 and 2. They produce identical goods, or at least the consumers cannot tell the difference between the goods. Market demand is described by the same demand function $D(p)$ as above. Both firms also have the same cost function $C(q_i) = c \cdot q_i$, where c is the marginal cost and q_i is the quantity produced by firm i.

The market works as follows: First, the firms simultaneously set their prices, independently of each other. Second, the consumers decide how much to buy and which firm they want to buy from.

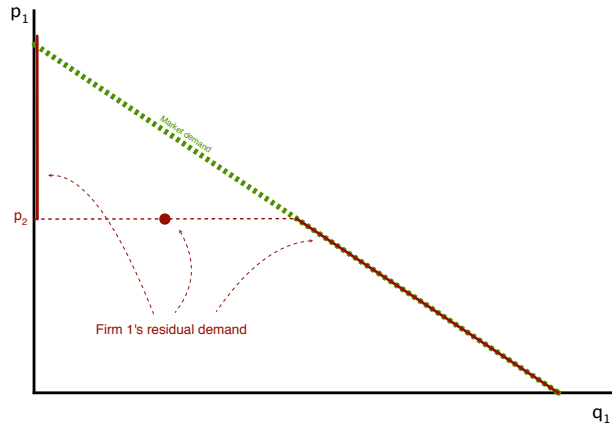
Residual demand The first step in the study of oligopoly is to derive the firms' residual demand functions. The residual demand function describes how much a firm will sell depending on the price it charges, given the price charged by the other firm. Thus the residual demand function describes an oligopoly firm's trade-off between price and sales, assuming that the competitor's price is fixed.

Since, in our case, the goods are homogenous, the consumers will only buy from the firm charging the lowest price. If firm 1 charges the lowest price, $p_1 < p_2$, firm 1's residual demand will be the whole market demand, that is $D_1(p_1, p_2) = D(p_1)$, and firm 2 will not have any residual demand, that is $D_2(p_1, p_2) = 0$. If $p_1 > p_2$, the roles will be reversed. Then, $D_1(p_1, p_2) = 0$ and $D_2(p_1, p_2) = D(p_2)$. It is also assumed that if the firms charge the same price $p_1 = p_2$, the two firms will share the market equally, i.e. $D_1(p_1, p_2) = D(p_1)/2$ and $D_2(p_1, p_2) = D(p_1)/2$. In sum:

$$D(p_1, p_2) = \begin{cases} D(p_1) & p_1 < p_2 \\ D(p_1)/2 \quad if & p_1 = p_2 \\ 0 & p_1 > p_2 \end{cases}$$

When market demand is linear, firm 1's residual demand is described by the Figure :

Figure 6.2.: Residual demand



Exercise 27. Write down firm 2's residual demand and display in a figure!

Note that each firm's residual demand depends on the prices charged by both firms. This is the reason why the firms' choices are interdependent and why we need game theory to study oligopoly markets.

The Bertrand game Before reading any further, please solve the exercise:

Exercise 28. Describe the the market as a normal form game.

Answer: A game in normal form has three ingredients, the players, their strategy sets and their payoff functions. The players of the market game is the two firms. For simplicity, we do not treat the consumers as players. Their behavior is instead conveniently summarized by the two residual demand functions.[1] Each firm's

[1] An alternative and equivalent methodology would have been to define an extensive form game with the consumers' choices as a second period. Deriving demand would then have been made part of the game.

strategy is the price it wants to charge. (The strategy sets are thus the set of real numbers.) Since we assume that firms want to maximize their profits, the profit functions are the payoff functions. Firm 1's profit is given by $\pi_1(p_1, p_2) = (p_1 - c) \cdot D_1(p_1, p_2)$. Firm 2's profit is given by a similar expression. ◀

This game is often referred to as the Bertrand model after its inventor.

Equilibrium A pair of prices (p_1^*, p_2^*) constitutes a Nash equilibrium if,

- $\pi_1(p_1^*, p_2^*) \geq \pi_1(p_1, p_2^*)$ for all possible prices p_1, and
- $\pi_2(p_1^*, p_2^*) \geq \pi_2(p_1^*, p_2)$ for all possible prices p_2.

We will now show that there exists a unique equilibrium in the Bertrand model. In fact, we will use two different methods to prove this result.

Testing different possible outcomes The first method of finding the equilibrium is to check if different possible outcomes, i.e. strategy profiles (p_1, p_2), are Nash equilibria. That is, for every strategy profile, we will investigate if any of the firms have an incentive to deviate from the prescribed behavior.

First, note that any pair of prices such that one is larger than the other and the lower price is larger than marginal cost, i.e., $p_i > p_j > c$ could not be an equilibrium. The reason is that the firm with the higher price, firm i, makes zero profit, despite the possibility of earning a positive profit by deviating to charge a price slightly below the rival. Exercise: Does firm j have a profitable deviation?

Second, note that both firms charging the same price above marginal cost, i.e. $p_i = p_j > c$, could also not be an equilibrium. The reason is that firm i could almost double its profit by slightly undercutting

the rival. Undercutting the rival's price slightly would essentially double sales without hardly reducing the mark-up.

Third, note that one firm charging price equal to marginal cost and the other firm charging a higher price, i.e. $p_i > p_j = c$, could not be an equilibrium, since firm j could increase its price at least slightly and earn a positive profit.

Finally, note that $p_i = p_j = c$ is an equilibrium. Both firms earn zero profits. By reducing price below c, a firm could corner the market, but all sales would be at a loss. Increasing the price above c would leave profit unchanged. Consequently, no firm has an incentive to change its behavior.

The conclusion is thus that there exists a unique equilibrium in the Bertrand game, namely that both firms charge the same price equal to marginal cost.

One argument in the proof is worth pointing out. In the third step it was shown that it could not be an equilibrium that one firm charges price equal to marginal cost while the other firm charges a higher price. The surprise is that it is the firm actually following the equilibrium strategy that has a profitable deviation in this case. If the other firm charges a higher price, the firm setting price equal to cost should increase its price to earn a positive mark-up.

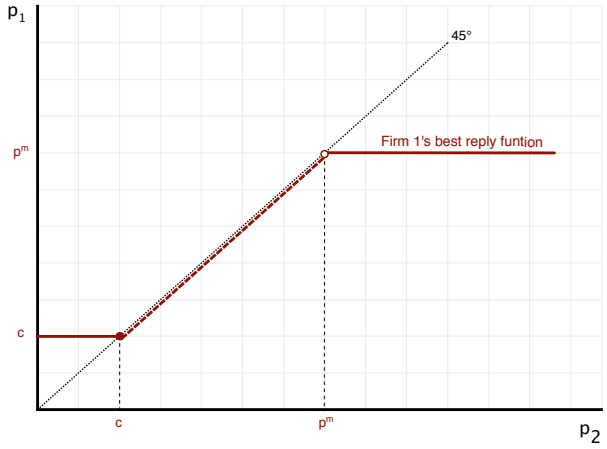
Another noteworthy aspect of this proof is that we did not find the exact price that the deviating firms should choose optimally. When rejecting an "equilibrium candidate" it is sufficient to show that one player could increase his payoff by some deviation. It is not necessary to find the best possible deviation.

Finally, recall that Nash's theorem only guarantees the existence of an equilibrium in finite games. The firms in the Bertrand game have infinitely many strategies, since each firm can choose any real number. Still, as we just learned, there actually exists an equilibrium also in the Bertrand model. In fact, there is a unique equilibrium. This example demonstrates the difference between sufficient and necessary conditions. Finiteness is sufficient for a game to have

an equilibrium. Finiteness is not, however, necessary.

Best-reply analysis The second method to find the equilibrium starts out by deriving the two firms' best reply functions. Recall that firm 1's best reply function describes the best (i.e. profit maximizing) price for firm 1, for every price that firm 2 could charge. And, as it turns out, firm 1's optimal choice is different depending on what firm 2 is assumed to do. The best reply function is described by the red line in Figure 6.3.

Figure 6.3.: Firm 1's best-reply function



First, assume that firm 2 charges some price above the monopoly price, i.e. $p_2 > p^m$. Then, firm 1's best choice would be to set the monopoly price, $p_1 = p^m$, and serve the whole market. In this case firm 1 would earn the monopoly profit. This is described by the upper right horizontal segment in Figure 6.3.

Second, assume that firm 2 charges some price below marginal cost, i.e. $p_2 < c$. Then, if firm 1 would undercut or match the rival's price, i.e. $p_1 \leq p_2$, it would make a loss, by selling a positive quantity at a price below its own cost. However, any price above the rival's price, i.e. $p_1 > p_2$, would lead to a zero profit. This means that there does not exist a unique best reply (profit-maximizing

price, given firm 2's price). This will not create any problems for the analysis, however. We can simply select $p_1 = c > p_2$, which is one of all the best replies. This is described by the lower left horizontal segment in Figure 6.3.

Similarly, if firm 2 would set price equal to marginal cost, $p_2 = c$, then any price $p_1 \geq c = p_2$ would lead to a zero profit, which is better than undercutting. Again, we can simply select $p_1 = c = p_2$ which is one of the best replies. This is described by the filled red dot in Figure 6.3.

Third, assume that firm 2 charges a price at or below the monopoly price but above marginal cost, i.e. $p_2 \in (c, p^m]$. Then, firm 1's best strategy is to set its price 2 slightly below the competitors' price, i.e. $p_1 = p_2 - \varepsilon$ where ε is a small number, say one Cent. Charging exactly the same price as the competitor would make the two firms share the market equally and lead to a positive profit, $\pi_1 = (p_2 - c) D(p_2)/2$. But, undercutting the rival's price slightly would essentially double sales hardly without reducing the mark-up. This is described by the dashed red line, just below the 45°-line in Figure 6.3.

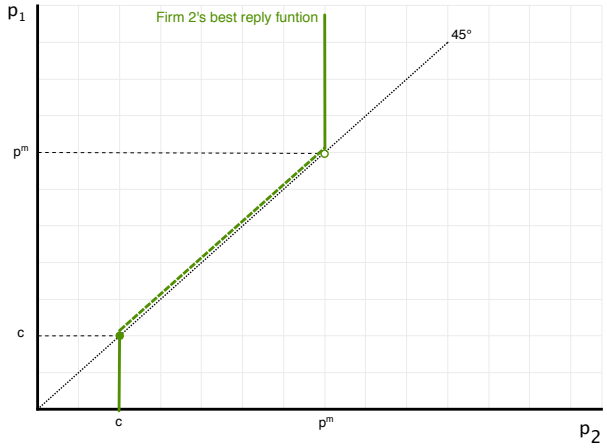
Strictly mathematically speaking, however, there does not exist a best-reply, since we cannot determine the exact amount by which the firm should undercut the rival. It must be a strictly positive number. So, let us say that it is $\varepsilon > 0$. But then $\varepsilon/2 > 0$ would lead to an even higher profit. This does not cause any serious problems for the analysis, however. It is a technicality. The reason is if when $\varepsilon > 0$ is very small, then $p_2 - \varepsilon$ and $p_2 - \varepsilon/2$ result in almost identical outcomes. The reason why the best-reply function in the figure is dashed is to indicate that we cannot determine the best reply exactly.

Exercise 29. Derive firm 2's best reply function and draw it in a figure with firm 2's price on the horizontal axis.

Answer: Likewise, firm 2's best reply function describes the best

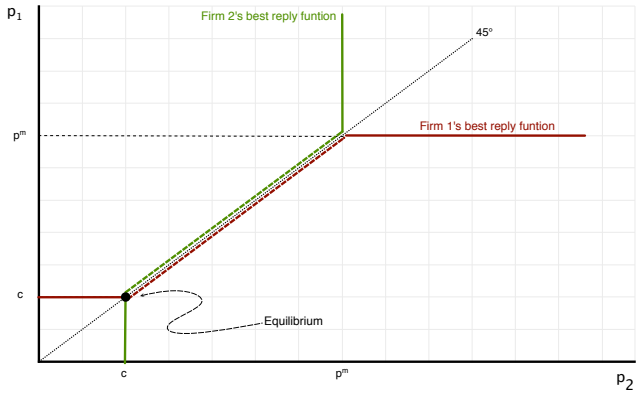
price for firm 2, given any price charged by firm 1. This best reply function is described by the green line in Figure 6.4. .◀

Figure 6.4.: Firm 2's best-reply function



The equilibrium is a pair of prices (p_1^*, p_2^*) such that p_1^* maximizes firm 1's profit if firm 2 charges p_2^* and vice versa. That is, both prices are mutual best replies, which means that any intersection of the two best-reply functions must be an equilibrium. As revealed by Figure 6.5, the two best-reply functions intersect at one point only,

Figure 6.5.: Bertrand equilibrium



namely (c, c) which means that there exists only one equilibrium,

which prescribes that both firms should set their prices equal to their common marginal cost.

6.3. What is price competition

Price competition arises when two or more firms wish to sell the same or similar goods to the same set of consumers and therefore set their prices to convince at least some of the consumers to buy from them rather than from the other firms. As a result prices will be lower than they would have been with only one firm serving the whole market. We have thus arrived at a *prediction* about how market price relates to market concentration:

Conclusion 3. The more firms there are in the market, the lower is price.

This is one of the most fundamental results of industrial economics. This prediction is also captured in a more formal way by comparing the outcome of the Bertrand model with the outcome in the monopoly model, keeping demand and technology the same in the two models.

The Bertrand model portrays competition as extremely fierce. With only two firms, price is brought down to cost, and profits to zero. The reason why competition is so fierce is that *even if a firm undercuts the competitor's price by just one cent, all customers will switch to the low-price firm.* As a result, as long as the rival's price is above cost, undercutting the rival is always profitable.

Actually, the radical drop in the price of Zocord when the patent expired is probably due to the fact that the firm charging the lowest price can count on serving a very large share of the Swedish market. Since October 2002 new rules govern the consumption of pharmaceuticals in Sweden. The pharmacies have to substitute a prescribed pharmaceutical with the cheapest available equivalent copy containing the same substance (so-called generics) in stock.

Thus, similar to the Bertrand model, any Simvastatin manufacturer undercutting the rivals' prices by only a small amount is rewarded with a very substantial share of the market.

In most markets, however, also duopoly firms are usually able to set price above cost, and earn a profit. Also duopoly firms typically have some market power. There are limits to competition also in duopolies. We will return to this discussion in 102 and more fully in 189.

6.3.1. Does competition reduce price?

The prediction that competition lowers price is one of the most fundamental predictions of industrial economics. But is it true? To answer this question we need to look at empirical studies. Actually, there are many empirical studies that seem to confirm the theoretical prediction. We only discuss this briefly here since we will return to empirical issues in later chapters.

Studying changes over time Some studies have investigated how prices change over time, when the number of firms in the market changes. The number of firms may, in turn, change e.g. as a result of mergers and acquisitions. Another source of variation is of course that patents expire, as in the case of Zocord discussed in the introduction to this chapter.

Comparing different geographical markets Other studies compare prices in different geographical markets. Comparing the price of the same product in different geographical areas with different number of firms reveal that price is often lower in markets with more firms. One study examined driving schools in 250 Swedish municipalities. Each municipality was considered to be a separate market, i.e. it was assumed that people only go to driving schools in the municipality in which they live. The municipalities in the Stockholm

area were excluded from the study since it is questionable if these municipalities, which all belong to the same commuting area, can be viewed as separate markets. The municipalities included in the study displayed substantial variation, with the number of driving schools ranging from zero to thirteen. The study found that all driving schools within a municipality charged approximately the same price. Between the different municipalities, prices varied a lot, however. In one municipality the price was SEK 4.50 per minute and in another it was 8.00 per minute. In all other municipalities the price was between these two extremes. The study then computed the average price among the municipalities with only one driving school and among municipalities with five driving schools. The price was found to be 0.30 higher in the monopolistic municipalities. The conclusion, then, is that the data is consistent with the theoretical prediction. It should be said, however, that market concentration appears to have a rather modest effect on driving school prices. In some other empirical studies, using the same methodology, the effect of competition appears to be larger.

6.4. Cournot oligopoly

6.4.1. Gothenburg Fish Auction

Every night fishing boats arrive in Gothenburg's fishing harbor.[2] The fish is gutted and put on ice already at sea. And when the fishermen are nearing the quay they call for immediate service to keep the fish cold at all times. Fish also arrive in lorries from other parts of the country. All the fish is registrated during the night before the Gothenburg fish auction starts around seven in the morning. Afterwards, the wholesalers take the fish to their storage rooms for filleting and further distribution to food stores, restaurants and schools.

[2] http://www.gfa.se/fiskauktionen.html

The individual fishermen determines how much time they spend fishing and therefore how much fish they wish to sell, but they do not set their own prices. The pricing decision is delegated to the auction mechanism. The more fish brought by one, the lower the price will be for all. By fishing more, each fisherman thus reduce the competitors' profits. But, not only their profits are affected. A reduced price also changes how much the competitors wish to fish. That is, the fishermen's decisions on how much to fish is interdependent.

A market in which all the producers first decide how much to sell and where the price is subsequently set to clear the market is called a Cournot market. We will analyze the Cournot model in two steps. First we consider a duopoly market and focus on the qualitative features and analyze the model largely using graphs only. Then we solve the model algebraically focusing on the importance of market concentration.

6.4.2. The duopoly model

Consider a homogenous goods market with two firms. The duopolist simultaneously and independently decide how much to produce. Both firms know that they will be able to sell at the market-clearing price, which is given by the inverse demand function, i.e. $p = P(Q)$ where $Q = q_1 + q_2$ is the total amount of goods sold in the market and q_i is the quantity produced by firm i. It is assumed that both firms wish to maximize their own profits. Firm i's profit depends on both firms' quantities and is given by $\Pi_i(q_1, q_2) = P(q_1 + q_2) \cdot q_i - c \cdot q_i$.

We have now specified a game in normal form - the Cournot model. The players are the two firms. The firms' strategies are their production quantities. The firms' payoffs are their profits. Note that in order to specify the profit function we need to know some market fundamentals, namely the demand function and the

cost function.

A pair of quantities (q_1^*, q_2^*) constitutes a Nash equilibrium, also called the Cournot equilibrium, if:

$$\begin{aligned} \Pi_1(q_1^*, q_2^*) &\geq \Pi_1(q_1, q_2^*) \qquad for\ all\ q_1, \\ \Pi_2(q_1^*, q_2^*) &\geq \Pi_2(q_1^*, q_2) \qquad for\ all\ q_2. \end{aligned}$$

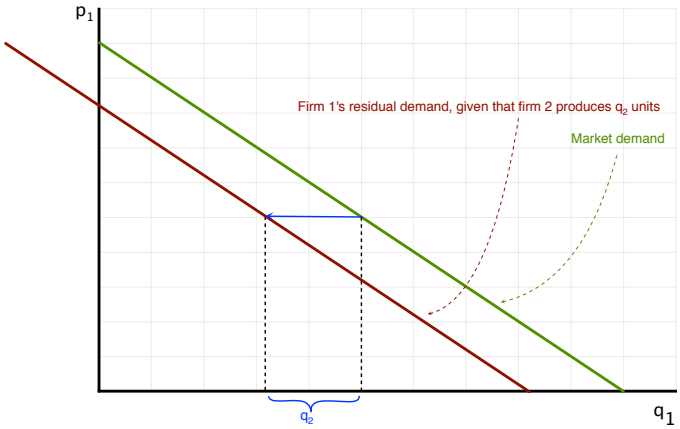
Without going into the details, one may show that there exists a unique equilibrium it the demand and cost functions satisfy certain regularity conditions. We will concentrate on deriving the equilibrium in a simple case when these conditions are fulfilled. In particular, we will assume that demand is linear, i.e. that demand is given by $P(Q) = a - b \cdot Q$. Before solving the model algebraically, we will rely mainly on graphs.

6.4.3. Residual demand

The first step of analysis is to derive the firms' residual demand functions. A firm's residual demand function shows what price the firm will get for every possible quantity the firm may choose to produce, taking as given both market demand and the quantity produced by the other firm. Market demand is described by the green line in Figure 6.6.

Firm 1's residual demand is described by the red line. Firm 1's residual demand is thus simply market demand minus the quantity produced by firm 2. That is, residual demand is equal to market demand shifted inwards by the amount produced by the other firm. It is easy to see that a firm's residual demand is lower the more the other firm produces. That is, the more one firm sells, the less the other firm will be able to sell. This is the reason why the firms' choices are interdependent and why we need to use game theory to predict the firms' behavior.

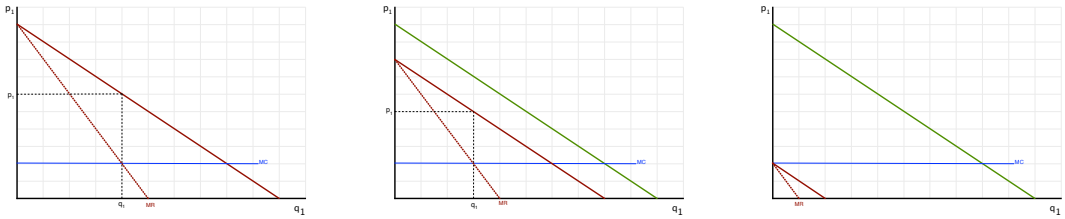
Figure 6.6.: Residual demand in Cournot



6.4.4. Best replies

A firm's best-reply function shows how much the firm should produce to maximize its profits, for every possible quantity that the other firm might choose to produce. Firm 1's optimal quantity is determined by equating marginal revenue to marginal cost. Since residual demand varies with the amount produced by firm 2, marginal revenue will vary with q_2 . As a consequence, also firm 1's optimal quantity varies with q_2. Figure 6.7 shows firm 1's optimal quantity

Figure 6.7.: Firm 1's optimal quantity



(a) $q_2 = 0$ (b) $q_2 \in (0, Q^c)$ (c) $q_2 = Q^c$

for three different levels of q_2.

Panel (a) shows that if firm 2 does not produce anything, then firm 1's residual demand is the same as market demand, meaning

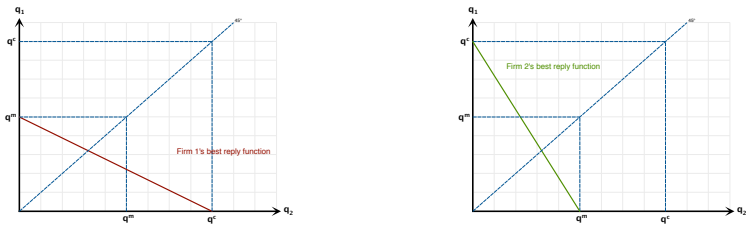
that firm 1 will produce the monopoly output, i.e. if $q_2 = 0$, then $q_1 = q_1^m$. On the other extreme, panel (c) shows that if firm 2 produces the competitive output (the output where demand and marginal cost intersect), then the best thing for firm 1 to do is to not produce anything, i.e. if $q_2 = Q^c$, then $q_1 = 0$. The middle panel (b), shows that if firm 2 produces a volume between the monopoly level and the competitive level, then firm 1 will optimally produce a positive amount below the monopoly level.

Thus, firm 1's best-reply function requires firm 1 to produce less, the more firm 2 produces. Looking carefully, one can also see that if firm 2 increases production by one unit, then firm 1 will reduce output by less than one unit.

Exercise 30. Show that in a figure, by shifting firm 1's residual demand inward by one unit!

In sum, firm 1's best reply function is illustrated in the left panel of Figure 6.8. Recall that firm 1 produces the monopoly level if

Figure 6.8.: Cournot Best replies



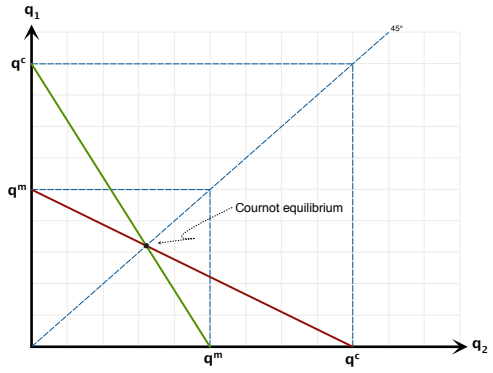
(a) Firm 1 (b) Firm 2

firm 2 does not produce anything and that firm 1 produces nothing if firm 2 produces the competitive output. Likewise, the right panel depicts firm 2's best reply function, i.e. the best quantity for firm 2 to produce, for all possible quantities that firm 1 might produce.

6.4.5. Equilibrium

The Nash equilibrium of the Cournot model is given by the intersection of the two best reply curves in Figure 6.9. The question

Figure 6.9.: Cournot equilibrium



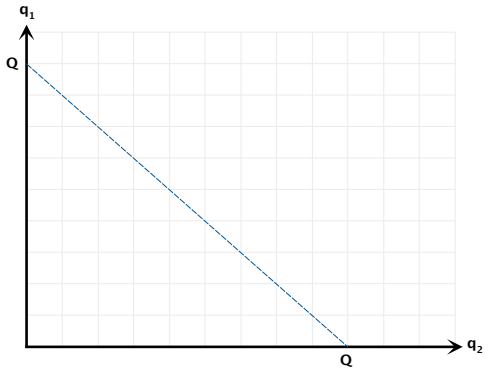
is now how the Cournot equilibrium compares with the Bertrand equilibrium and the monopoly outcome. How much competition is there in Cournot?

6.4.6. Comparisons

To compare the outcomes of the different models it is convenient to focus on aggregate output, $Q = q_1 + q_2$. Note that we can rewrite this equation as $q_1 = Q - q_2$. This equation is illustrated in Figure 6.10 as the dotted straight line with intercept Q and with slope equal to -1. Each point on the dotted straight line represents a different pair of quantities, but in each case the two quantities sum to the same aggregate output. Clearly both $(Q, 0)$ and $(0, Q)$ as well as $\left(\frac{Q}{2}, \frac{Q}{2}\right)$ for example are points on this line.

Figure 6.11 has three such aggregate output lines. They represent the output under monopoly, Cournot and perfect competition. Now it is easy to see that the aggregate output in a Cournot equilibrium,

Figure 6.10.: Combinations of firm outputs summing to the same aggregate output



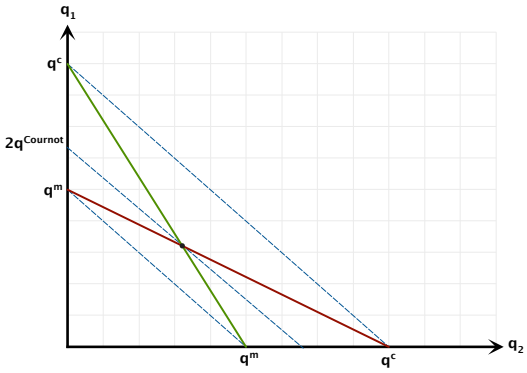
i.e. $Q^{Cournot} = q_1^{Cournot} + q_2^{Cournot}$, is larger than the monopoly output, Q^m, but smaller than the competitive output, Q^c.

Exercise 31. How does the equilibrium price in a Cournot market compare to the the monopoly price and the price under perfect competition?

Answer: Recall that the equilibrium price under Cournot is the price that clears the market, given the amount produced. That is $p^{Cournot} = P\left(Q^{Cournot}\right)$. Similarly, a monopolist charges the highest price possible, that is $p^m = P\left(Q^m\right)$. Finally, the price under perfect competition is given by $p^c = P\left(Q^c\right)$. Demand slopes downward. Expressed differently, to clear the market, the price must be lower, the higher aggregate production is. Thus: $p^m > p^{Cournot} > p^c$.◀

The first conclusion is that competition matters, also when firms have capacity constraints. Comparing the Cournot duopoly outcome with the monopoly outcome, we see that more firms in the market leads to lower prices, even if they are capacity constrained. The second conclusion is that competition is less intense with capacity constraints than without capacity constraints. Comparing the Cournot duopoly outcome with the Bertrand duopoly outcome,

Figure 6.11.: Comparison of Cournot to monopoly and perfect competition



we see that prices are higher under Cournot.

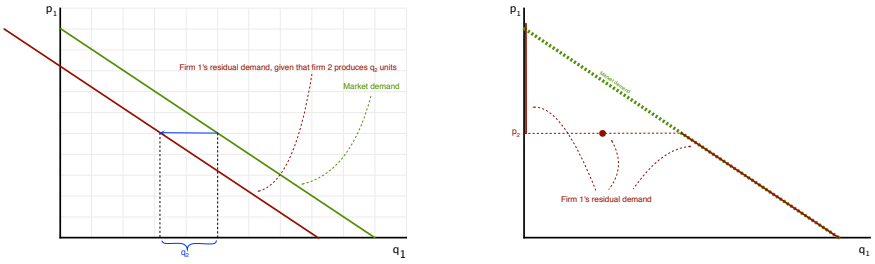
6.4.7. Explaining the difference between Cournot and Bertrand

The Cournot model shows that there is less competition in a market when capacity constraints are binding. Intuitively, it is less profitable for a firm to reduce price if the firm cannot serve all the consumers who wish to buy at the lower price.

One may get some additional intuition for the difference between Bertrand and Cournot by comparing the residual demand functions of the two models. In particular, we need to compare the price elasticity of residual demand at equal prices ($p_1 = p_2$). Recall that firm 1's residual demand function describes the combinations of price and quantity (p_1, q_1) that firm 1 can choose from. The two red lines in Figure6.12 display the residual demand functions in Cournot and Bertrand, given the same (green) market demand function.

To see why firm 1's residual demand curves are different in Cournot and Bertrand markets, assume that we start out from a situa-

Figure 6.12.: Comparing Cournot and Bertrand



(a) Cournot (b) Bertrand

tion where the two firms charge the same price $p_1 = p_2 = p^*$ and where they both sell half the market demand at this price, $q_1 = q_2 = D(p^*)/2$. Let is now consider the effect of a change in firm 1's behavior. Assume in particular that firm 1 contemplates increasing price (p_1) and reducing its quantity (q_1). To determine if such a change is profitable or not, we need to study how much the quantity is reduced in relation to how much the price is increased, i.e. we need to determine how price-elastic firm 1's residual demand curve is at the common price.

Let us start with the Bertrand model. If firm 1 increases its price slightly above p^*, firm 2's price remains fixed at p^*. As a consequence, the quantity that firm 1 can sell drops to zero, since all customers want to buy from the cheapest firm. This implies that firm 1's residual demand is perfectly elastic at p^*.

Next consider the Cournot model. If firm 1 reduces its output slightly below $D(p^*)/2$, firm 2's quantity remains fixed at $D(p^*)/2$. Note that a small reduction in firm 1's output below $D(p^*)/2$ causes the market-clearing price to increase as a result of the reduction in the total quantity supplied. Expressed differently, firm 1's price will increase slightly above p*. But firm 2's price will also be increased at the same time and to the same level. Since the two firms' prices are the same also in the new situation, consumers will not switch from firm 1 to firm 2 when firm 1 initiates the price

increase. Firm 1 will thus be able to continue to sell its now slightly lower production. This implies that firm 1's residual demand is not so elastic at p^*.

It is thus more profitable for a firm to initiate a price increase in a Cournot market (by cutting back production) than it is in a Bertrand market. In a Bertrand market the price increase will cause a huge drop in sales. In a Cournot market the same price increase can be accomplished by a much smaller loss of sales.

6.5. The role of market concentration

Consider now a market with linear inverse demand $P(Q) = a - b \cdot Q$ and N identical firms, all producing at constant marginal cost c. Recall that firm i's profit is given by $\Pi_i = [P(Q) - c] \cdot q_i$ where aggregate production is the sum of all firms' quantities, i.e. $Q = q_1 + q_2 + ... + q_N = \sum_{j=1}^{N} q_j$. In this section we derive the equilibrium of the linear Cournot model analytically.

Best replies First, we derive firm 1's optimal choice of quantity, given any possible quantities produced by the competitors. To do this, we use the first-order condition:

$$\frac{\partial \Pi_i}{\partial q_i} = [P(Q) - c] + \frac{\partial P(Q)}{\partial q_i} \cdot q_i = 0.$$

The first-order condition states that marginal revenue ($P(Q) + \frac{\partial P(Q)}{\partial q_i} \cdot q_i$) and marginal cost ($c$) should be equated. With linear demand, the condition can be written as:

$$[a - b \cdot Q - c] - b \cdot q_i = 0$$

or

$$a - b \cdot \sum_{j=1}^{N} q_j - c - b \cdot q_i = 0.$$

Solving this equation for q_i gives firm 1's best reply function:

$$q_i = \frac{(a-c)}{2 \cdot b} - \frac{1}{2} \cdot \sum_{j \neq i} q_j.$$

Exercise 32. Do these calculations yourself!

Note that each firm wishes to produces less the more is produced by the competitors.

Equilibrium Recall that in order for the quantities $(q_1, q_2, ..., q_N)$ to be an equilibrium, they need to satisfy all the firms' best-reply functions at the same time. This means that the equilibrium quantities $(q_1, q_2, ..., q_N)$ must satisfy a system of linear equations:

$$\begin{aligned} q_1 &= \tfrac{(a-c)}{2 \cdot b} - \tfrac{1}{2} \cdot \textstyle\sum_{j \neq 1} q_j \\ q_2 &= \tfrac{(a-c)}{2 \cdot b} - \tfrac{1}{2} \cdot \textstyle\sum_{j \neq 2} q_j \\ &\vdots \\ q_N &= \tfrac{(a-c)}{2 \cdot b} - \tfrac{1}{2} \cdot \textstyle\sum_{j \neq N} q_j \end{aligned}$$

Thus, to find the equilibrium, we may solve this system of equations for the equilibrium quantities.

Luckily, there is a simpler procedure. To derive the equilibrium it is actually more convenient to first rewrite firm 1's first order condition on the following form

$$q_i = \frac{(a-c)}{b} - Q.$$

Exercise 33. Do that!

Second, since all firms' first order conditions can be written exactly the same way, we can find the *total quantity* in equilibrium by summing all these equations:

$$\sum_{i=1}^{N} q_i = \sum_{i=1}^{N} \left[\frac{(a-c)}{b} - Q \right].$$

Since the left hand side is equal to total output and the right hand side is a sum of N identical terms:

$$Q = N \cdot \left[\frac{(a-c)}{b} - Q \right].$$

We now have one equation in one unknown, the total equilibrium output. Solving the equation yields

$$Q = \frac{N}{N+1} \cdot \frac{a-c}{b}.$$

Third, the firm i's individual quantity is found by using the (rewritten) first-order condition and substitute for the total quantity, i.e.

$$q_i = \frac{(a-c)}{b} - Q = \frac{(a-c)}{b} - \frac{N}{N+1} \cdot \frac{a-c}{b},$$

i.e.

$$q_i = \frac{1}{N+1} \cdot \frac{(a-c)}{b}.$$

Finally, we compute the equilibrium price by substituting the aggregate production into the residual demand function:

$$P(Q) = a - b \cdot Q = a - b \cdot \frac{N}{N+1} \cdot \frac{a-c}{b},$$

i.e.

$$p^{Cournot} = \left(\tfrac{1}{N+1}\right) \cdot a + \left(\tfrac{N}{N+1}\right) \cdot c.$$

That is, the equilibrium price is determined by the consumers' willingness to pay (as captured by their willingness to pay for the first unit, a), the marginal cost of production (c) and the number of firms in the market, N. In fact, the equilibrium price lies in the interval (c, a) and the exact location, in this interval is determined by the number of firms.

Exercise 34. Compute the price for monopoly ($N = 1$) and com-

pare with the monopoly model above. Compute the price for duopoly ($N = 2$) and triopoly ($N = 3$). Compute the price for $N = 150$ and $N = 1500$, which is the approximate number of licensed fishing boats in Sweden.

These numerical examples demonstrate that the equilibrium price is lower, the more firms there are in the market.

Exercise 35. Prove that price is lowered when the number of firms is increased? What happens to price when the number of firms becomes very large (i.e. goes to infinity).

Answer: That price is lowered by an increase in the number of firms is proved by the derivative:

$$\frac{\partial p^{Cournot}}{\partial N} = -\left(\tfrac{1}{N+1}\right)^2 \cdot (a-c) < 0.$$

When the number of firms becomes very large

$$\tfrac{1}{N+1} \approx 0$$

and

$$\tfrac{N}{N+1} \approx 1$$

which means that

$$p^{Cournot} \approx c.$$

◀

6.6. Exercises

6.6.1. Definition of the Cournot game

Who are the players in a Cournot game? What are their strategies? What are their payoffs? What market fundamentals do we need to know to specify the payoff functions?

6.6.2. Best reply

Consider a Cournot duopoly. How much would firm 1 produce if firm 2 does not produce anything? How much would firm 1 produce if firm 2 produces the competitive output? Will a Cournot firm prefer to produce more or less if the rivals increase their production? Why? (Hint: what happens to residual demand?)

7. Bargaining

7.1. Why study bargaining?

Food retailers and their suppliers Food retailers have become much larger over time. Some retailers have sell an enormous amount of goods and operate on a global basis. Wal-Mart, which is the world's largest retailer, has recently also become the largest grocer in the world.[1] Its food sales in 2003 was about half of the Swedish GDP the same year. The top thirty companies account for around

Table 7.1.: World's largest food retailers 2003 (US$mn)

Company	Food sales
Wal-Mart	121 566
Carrefour	77 330
Ahold	72 414
Tesco	40 907
Kroger	39 320
Rewe	36 483
Aldi	36 189
Ito-Yokado	35 812
Metro Group ITM	34 700

one third of global grocery sales, and in many countries including the UK and Sweden, the top five supermarkets account for 70% or more of grocery sales.

Many food producers are also large firms. Some are large multinational conglomerates, with worldwide sales of over £25 billion.

[1] UK Department for International Development: Concentration in Food Supply and Retail Chains, August 2004.

Others operate on a national scale. On the other extreme there are specialist cottage industry providers, with sales of under £1 million.[2]

Not only are both food producers and food retailers quite large, they are also mutually dependent on each other. The retailers are particularly dependent on the suppliers of so-called "must-stock" brands. That is, most retailers probably have to carry certain brands such as Coca Cola or consumers will go somewhere else. But suppliers are also dependent on the retailers. Suppliers of branded goods typically want wide distribution of their products and generally seek to trade with all retail chains.

The UK Competition Commission asked large suppliers of groceries to provide information about their sales to retailers in the UK. The data indicate that the suppliers are quite dependent on their largest customers. The sales to the suppliers' largest three customers accounted for on average 56 per cent. And some suppliers only deal with one retailer.

Given such high concentration on both sides and mutual dependence, it is not surprising to learn that wholesale prices are often *negotiated bilaterally* between the parties. Day-to-day negotiations particularly on price and quantity are usually conducted orally. As a result, prices and other conditions are often determined on a relation-specific basis, in contrast to the universal price lists posted by sellers in consumer markets.

A key characteristic of the wholesale market is so-called *buyer power*. Buyer power means that the retailer chains are often able to obtain lower prices from their suppliers as a result of their size and importance to the supplier's business. In fact, the largest chains are able to secure better deals than smaller chains, even though there are only small differences in the cost to serve most retailer chains.

[2]UK Competition Commission: Supermarkets: A report on the supply of groceries from multiple stores in the United Kingdom, 2000. See especially chapter 11 and Appendix 11.5.

That is, the suppliers' order processing and physical distribution costs are essentially the same, independent of which retailer chain they are dealing with.

The UK Competition Commission asked 26 large suppliers of groceries to provide the prices for their top five branded product lines paid by the different retailers. The larger retailers pay less than small ones, with Tesco, Sainsbury, Asda, Somerfield and Safeway paying the lowest prices. The table below shows the average (over different groceries) wholesale prices relative to Tesco, which pays the lowest price. Other, smaller buyers, pay even higher prices.

Table 7.2.: Average wholesale prices

Retailer	Market **Share** (CC Table 5:3, p. 44)	Price (CC Table 5, p. 435)
Tesco	24.6	100.0
Sainsbury	20.7	101.6
Asda	13.4	102.3
Somerfield	8.5	103.0
Safeway	12.5	103.1
Morrison	4.3	104.6
Iceland	0.1	105.3
Waitrose	3.3	109.4
Booth	0.1	109.5
Netto	0.5	110.1
Budgens	0.4	111.1

The table reveals a clear negative relationship between the retailer size and price. Given that the costs of the goods for resale constitutes around 85% of total cost in food retailing (CC, p. 210) the wholesale price differences between the different retailers displayed in the table are considerable. A key question is thus why large buyers succeed to obtain much better prices?

Unionization in labor markets Swedish employees are heavily organized. Most nurses, for example, belong to Vårdförbundet,

a part of the Swedish Confederation for Professional Employees (TCO). Since most health care services are publicly provided in Sweden, most nurses work for the county council or the municipality where they live. They have typically only a few potential employers. The market for nurses in any Swedish municipality is thus highly concentrated on both the sellers' side and on the buyers' side.

The nurses' wages and other working conditions are determined in a collective agreement negotiated between Vårdförbundet, on behalf of their members, and the employers. This means that the nurses' wages are not set by one party only, with the other party passively accepting whatever level is offered. Since the nurses and hospitals are mutually dependent on each other, lacking alternative partners, they will both have a say over the agreement. Both sides have market power.

Commercial radio's reliance on music The commercial radio stations in Sweden rely almost exclusively on music to attract listeners. Without the right to play music, they would have few listeners, earn little advertising revenues, and most of them would probably not even exist. Even though the stations belong to two competing commercial groups (MTG and SBS) they have formed a joint committee (Förhandlingsgruppen för reklamradioföretagen) to negotiate conditions with the music producers. Also the music producers are heavily organized. STIM negotiates conditions on behalf of all the authors, including both composers and text writers. SAMI negotiates on behalf of all the performing artists, including both singers and musicians. IFPI represents the record companies. These organizations do not only represent their Swedish members, but they also act on behalf of all the music producers around the world. The price of music in commercial radio in Sweden is set through a negotiation between monopoly sellers and a monopoly buyer.

Relation-specific assets in car manufacturing A key to the success of the automobile industry in Japan is an extensive use of relation specific investments. Japanese auto-part suppliers send engineers to work at the customer's site, locate plants near the customer and invest in customized tools. These dedicated assets provide the Japanese manufacturers with substantial competitive advantages in the form of lower costs and higher quality.[3] The flip side of the arrangement is that the manufacturer and the subcontractors are looked-in with each other. As switching to another partner would be expensive, they will have to agree on prices, quantities, delivery times and so on with each other. Absent competition, it is the bargaining power of the two parties that will determine the outcome.

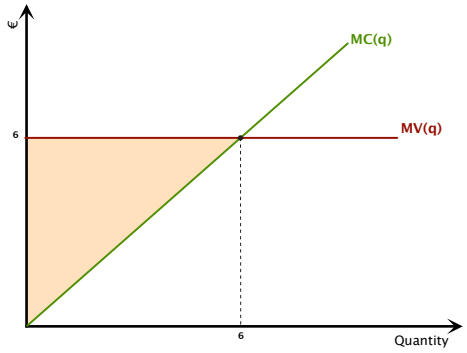
A simple example of bilateral monopoly In markets where there are either only a few buyers and sellers or where the buyers and sellers are heavily organized or where each buyer and seller is looked in with a specific partner, no single party can dictate the conditions for trade but all have a say and can influence the conditions at which they trade. There is market power on both sides.

To analyze the price negotiations we will, for simplicity, think of a bilateral monopoly, i.e. a market with only one buyer and one seller. The seller may be a producer of some food product. The seller can produce this product with a technology represented by the marginal cost curve (MC) in Figure 7.1 . The buyer may be a food retailer. The buyer's marginal valuation for the product is marked MV.

For simplicity it is assumed that the buyer's marginal valuation is constant. This would for example be true in case the buyer is a retailer with constant returns to scale and without any market

[3] Jeffrey H. Deyer. Dedicated Assets: Japan's Manufacturing Edge, Harvard Business Review, 1994.

Figure 7.1.: Bilateral monopoly



power when reselling the product to the final consumers. For example, if the retailer can sell the product to the final consumers at the constant price p and if the retailer incurs the constant marginal cost c for every unit, the retailer would earn $MV = p - c$ for every unit of the product it buys and resells.

Exercise 36. What would the MV-curve look like if there would be decreasing returns to scale in retailing? What would the MV-curve look like if the retailer would have market power in the final goods market?

The question now is how much the buyer and the seller will agree to trade and at what price?

It may appear natural that the two parties would agree to trade the efficient quantity, equalizing marginal cost and marginal valuation (i.e. $MV(q^*) = MC(q^*)$), which means six units in the example given in the figure. Agreeing on anything less, they could supplement the contract with agreement to produce a little bit more and share the extra surplus.

The question of price is trickier. If they agree to trade q^* units, the total surplus from the transaction is represented by the triangle formed by the MC- and MV-curves and the y-axis. The seller's total

cost of producing the six units would then be €18 and the buyer's total valuation of the six units would be €36. The total social surplus is thus €18 in this example. Then any price for the six units[4] covering the seller's cost and which is lower than the buyer's total valuation seems possible.

So what, if anything, could force the two parties to agree on some specific price in the range between €18 and €36? One important force is that haggling takes time and thus is costly. If one party demands "too much" the other party will reject the offer and make a counter-offer. Offers and counter-offers could go on forever. This means that the gains from trade could be delayed and that the parties have to spend time at the negotiation table. But, again, exactly what price will the parties agree on? To answer that question we need to analyze the situation in a more formal way, using the so-called bargaining theory. As it turns out, we will have to study the negotiation process in quite some detail, before we arrive at a prediction about the price (or wage) and the quantity traded in a bilateral monopoly.

Other applications Although this chapter focuses on markets with bilateral market power, bargaining theory is useful for studying a very large and varied set of phenomena. In fact, bargaining is everywhere. Family members bargain over the division of household duties. Firms and employees bargain over wages and working conditions. Firms bargain with their sub-contractors over prices, qualities and delivery times. Governments bargain over trade agreements, climate protocols and military pacts.

[4]Note that the price is here assumed to be for all the 6 units.

7.2. An extensive form bargaining game

Bargaining is a complex phenomenon. Some people are skilled at it; others are not. In some cases the parties understand each other's objectives; in other cases not. Some negotiations are bilateral; others are multilateral, involving a large number of parties. And so on.

Here, we will only provide a first glimpse of the economic analysis of bargaining, by presenting the most used extensive-form model of bargaining. As all models, it is an abstraction. A host of important aspects of real-life bargaining are not represented in the model. It is e.g. assumed that the players are fully rational and that they have complete information about their rival's payoff. This game is often referred to as the Rubinstein-Ståhl bargaining game to credit the two most important early contributors to this theory.

7.2.1. Ultimatum bargaining

Let us start with the simplest version of the bargaining model, which is often referred to as ultimatum bargaining. There are two players, called the seller and the buyer, who can share a surplus if they agree to trade. To simplify the analysis, they either trade the efficient quantity or not at all. To simplify notation we assume that the surplus is €1. The seller is allowed to propose whether to trade or not and (in case of trade) a split of the surplus, say σ_S for himself and σ_B for the buyer. The buyer can either accept the offer or reject it. If the buyer accepts the offer, the players get σ_S and σ_B according to the proposal. If the buyer rejects the offer, bargaining stops and they both receive zero. The players are selfish and only care about the amount of money they receive themselves.

Equilibrium Since this is a dynamic game (with two periods) and since information is both perfect (the buyer knows what the seller

proposes) and complete (the seller and the buyer know the other player's payoff function), the relevant solution concept is subgame perfection, which can be found by backwards induction.

Backwards induction requires us to start from the end, with the buyer's decision to accept or reject the seller's offer. If the buyer rejects the offer, he gets zero. If he accepts the offer, he gets σ_B. Thus, the buyer accepts any offer $\sigma_B \geq 0$. If the offer is exactly zero the buyer is actually indifferent between accepting and rejecting, and we simply assume the buyer to accept the offer.

The next step is to analyze the seller's decision. Being able to predict the buyer's future behavior, the seller maximizes his utility by (i) suggesting to trade and (ii) taking all the money for himself and offering zero to the buyer.

That is, there is a unique subgame perfect equilibrium with agreement on $\sigma_B = 0$ and $\sigma_S = 1 - \sigma_B = 1$.

Properties of the equilibrium Before going on to analyze a slightly more realistic bargaining game, we need to reflect on the properties of this equilibrium. First note that the outcome is efficient, i.e. the seller proposes trade, thereby generating a surplus of €1. More generally, the seller would propose the efficient quantity, maximizing the total surplus. That is, in the example in the figure above, the seller would propose to trade six units.

Next note that the outcome gives the proposer the whole surplus and nothing to the respondent. The distribution of wealth is thus highly asymmetric in the ultimatum bargaining game. The proposer has all the bargaining power.

Finally recall that the buyer is assumed to only care about his monetary payoff and to act rationally. Thus, the buyer does not get upset by the seller's selfishness or the unfairness of the situation.

7.2.2. Alternating offers

Two periods

Assume now that there is time for two rounds of negotiations. The negotiation starts by the buyer proposing and the seller responding. If the two reach an agreement, the negotiation is over. If they fail to reach and agreement, the negotiation continues a second period. In the second round it is the seller's turn to make the offer and the buyer will respond. To prepare for more periods, we will number the last period T and the second to last (i.e. first) period T-1. The deal suggested in period t is denoted (σ_B^t, σ_S^t), where the first number is the buyer's share of the surplus and the second number is the seller's share of the surplus.

The players are assumed to be impatient. They do not only care about the amount of money they receive, but also when they receive the money. The seller's discount factor is $\delta_S < 1$ and the buyer's discount factor is $\delta_B < 1$.This means that the seller is equally happy if he receives €1 in period two or the lower amount €δ_S already in period one. To fix ideas we assume that the time between the offers is short. Then, getting the money the next period is almost as good as getting it now. In this case, the discount factors are close to one.

Equilibrium Again, we can find a prediction using backwards induction. We already know what would happen if the last period (period T) is reached. In that case the two players find themselves in the ultimatum game we analyzed above. Thus, the buyer accepts any offer giving him at least zero and the seller will therefore propose to trade and:

$$\sigma_S^T = 1$$
$$\sigma_B^T = 0.$$

In the first period (period T-1), the buyer gets to propose a deal

and it is the seller that responds. Using backwards induction, we start with analyzing the seller's behavior. If the seller rejects the offer, he will earn $\sigma_S^T = 1$ the next period. The value of that deal in period T-1 is $\delta_S \cdot \sigma_S^T = \delta_S < 1$. If the seller accepts the buyer's offer, he will get σ_S^{T-1}. Thus, the seller accepts any offer such that $\sigma_S^{T-1} \geq \delta_S$. Since $\sigma_S < 1$, the seller is willing to accept an amount that is slightly lower than the amount he would get in the last period. A slightly smaller amount now is better than having to wait for the whole surplus the next period. The buyer will therefore propose trade and the following division of the surplus

$$\begin{aligned} \sigma_S^{T-1} &= \delta_S \\ \sigma_B^{T-1} &= 1 - \delta_S. \end{aligned}$$

which is accepted.

Properties of the equilibrium Again, the two parties agree to trade. The outcome is efficient in this sense.

The two parties reach an agreement already in the first period (period T-1). There is thus no delay in equilibrium. Since both parties prefer an early agreement to a later agreement (on the same split), the fact that there is immediate agreement is also a form of efficiency.

Note that when the negotiation can continue for two rounds, the parties split the money somewhat more equally between themselves. The seller has the advantage by being allowed to make the last proposal. But the buyer can appropriate a small share of the surplus by exploiting the seller's impatience, i.e. the seller's preference for an early agreement. The size of the buyer's share is determined by the seller's aversion against delay. If it is really important for the seller to get his money immediately rather than to wait until the next period (δ_S is low) the seller will only get small share of the money ($\sigma_S = \delta_S$) and the buyer gets a lot ($\sigma_B = 1 - \delta_S$). But if the seller is patient and doesn't care much about when he receives

his money (δ_S is high) the seller will get a large share of the money and the buyer has to accept a modest share. That is bargaining power is partly determined by patience.

Three periods

Assume now that there is time for three rounds of negotiations. The players take turns proposing an agreement: the seller proposes in the last period T, the buyer proposes in the second to last period T-1, and the seller proposes in period T-2.

Equilibrium Using backwards induction, we already know what will happen if the bargaining would continue to period T-1. The outcome would then be

$$\sigma_S^{T-1} = \delta_S$$
$$\sigma_B^{T-1} = 1 - \delta_S.$$

In period T-2, the buyer would accept any offer giving him at least $\sigma_B^{T-2} = \delta_B \cdot \sigma_B^{T-1} = \delta_B - \delta_B \cdot \delta_S$. The seller therefore proposes

$$\sigma_S^{T-2} = 1 - \delta_B + \delta_B \cdot \delta_S$$
$$\sigma_B^{T-2} = \delta_B - \delta_B \cdot \delta_S.$$

Properties of the equilibrium Again, the parties agree to trade. Again, there is no delay. The only news is that now the split is determined by both players' discount factors, i.e. both player's patience.

More periods

With four periods, the buyer makes the first offer. The parties agree immediately to trade and to share the surplus in the following way:

$$\sigma_S^{T-3} = \delta_S - \delta_B \cdot \delta_S + \delta_B \cdot \delta_S^2$$
$$\sigma_B^{T-3} = 1 - \delta_S + \delta_B \cdot \delta_S - \delta_B \cdot \delta_S^2.$$

With five periods, the seller makes the first offer. The parties agree immediately to trade and to share the surplus in the following way:

$$\sigma_S^{T-4} = 1 - \delta_B + \delta_B \cdot \delta_S - \delta_B^2 \cdot \delta_S + \delta_B^2 \cdot \delta_S^2$$
$$\sigma_B^{T-4} = \delta_B - \delta_B \cdot \delta_S + \delta_B^2 \cdot \delta_S - \delta_B^2 \cdot \delta_S^2.$$

With six periods, the buyer makes the first offer. The parties agree immediately to trade and to share the surplus in the following way:

$$\sigma_S^{T-5} = \delta_S - \delta_B \cdot \delta_S + \delta_B \cdot \delta_S^2 - \delta_B^2 \cdot \delta_S^2 + \delta_B^2 \cdot \delta_S^3$$
$$\sigma_B^{T-5} = 1 - \delta_S + \delta_B \cdot \delta_S - \delta_B \cdot \delta_S^2 + \delta_B^2 \cdot \delta_S^2 - \delta_B^2 \cdot \delta_S^3.$$

Exercise 37. Describe the equilibrium if there are seven periods.

Answer: The parties agree immediately to trade and to share the surplus in the following way:

$$\sigma_S^{T-6} = 1 - \delta_B + \delta_B \cdot \delta_S - \delta_B^2 \cdot \delta_S + \delta_B^2 \cdot \delta_S^2 - \delta_B^3 \cdot \delta_S^2 + \delta_B^3 \cdot \delta_S^3$$
$$\sigma_B^{T-6} = \delta_B - \delta_B \cdot \delta_S + \delta_B^2 \cdot \delta_S - \delta_B^2 \cdot \delta_S^2 + \delta_B^3 \cdot \delta_S^2 - \delta_B^3 \cdot \delta_S^3.$$

◀

Finding an expression that can be interpreted We could continue, but the pattern is now clear. It is time to rewrite the equilibrium shares on a form that can be interpreted. Let us focus on the buyer's share, when there are seven periods, which can be rewritten as:

$$\sigma_B^{T-6} = \delta_B \cdot (1 - \delta_S) \cdot \left[1 + \delta_B \cdot \delta_S + \delta_B^2 \cdot \delta_S^2\right].$$

Following the same procedure, one can show that when T is odd (so that the seller makes the bid in the first period), the buyer's share in equilibrium is given by

$$\sigma_B^1 = \delta_B \cdot (1 - \delta_S) \cdot \left[1 + \delta_B \cdot \delta_S + \delta_B^2 \cdot \delta_S^2 + ... + \delta_B^{\frac{T-3}{2}} \cdot \delta_S^{\frac{T-3}{2}}\right].$$

Exercise 38. Show this for the case when there are five periods. (Hint: rewrite $\sigma_B^{T-4} = \delta_B - \delta_B \cdot \delta_S + \delta_B^2 \cdot \delta_S - \delta_B^2 \cdot \delta_S^2$.)

The expression in the square bracket is a so-called geometric series and can be simplified.[5] The buyer's share can be rewritten as:

$$\sigma_B = \frac{1-\delta_S}{1-\delta_B \cdot \delta_S} \cdot \delta_B \cdot \left[1 - (\delta_S \cdot \delta_S)^{\frac{T-1}{2}}\right],$$

where the superscript, indicating that this is the agreement in the first period, has been dropped.

7.2.3. Predictions

The expression above reveals that the parties will split the money depending on how patient they are, i.e. depending on the two discount factors δ_S and δ_B. In fact, as we will soon demonstrate, the more patient a party is in relation to the other party, the larger share of the surplus he will be able to capture.

The split also depends on the number of periods that the parties could go on haggling (T) and also on some finer details of the bargaining procedure, namely who gives the last and first bids. These factors are probably not very important in many real-world negotiations when the negotiators sit around the same table and offers and counter-offers can be made very quickly and when the negotiation can go on for a very long time. And, in fact also this comes out as a prediction.

Advantage of being the last proposer The last factor (i.e. $1 - (\delta_S \cdot \delta_S)^{\frac{T-1}{2}} < 1$) captures the fact that the buyer has a disadvantage since seller has the right to make the last proposal. To see that this term comes from the fact that it is the seller that makes the last proposal, one simply has to analyze the bargaining game again, assuming that it is the buyer who has the right to make the

[5] Recall that $1 + x + x^2 + ... + x^N = \frac{1-x^{N+1}}{1+x}$.

last bid. However, if haggling could go on for a very large time, i.e. if T is very large, the significance of the last period vanishes and this factor will be approximately equal to one.[6] Then, the buyer's share will approximately be given by

$$\sigma_B = \frac{1-\delta_S}{1-\delta_B \cdot \delta_S} \cdot \delta_B.$$

Relative patience The main insight of the bargaining game is that the players' relative patience determines the split. This is most easily seen in case one of the players is impatient while the other is very patient. First, note that if the buyer is impatient (i.e. $\delta_B < 1$) while the seller is very patient (i.e. $\delta_S \approx 1$) then the buyer does not get anything (i.e. $\sigma_B \approx 0$). On the other hand, if the buyer is very patient (i.e. $\delta_B \approx 1$) while the seller is impatient (i.e. $\delta_S < 1$) then the buyer takes everything (i.e. $\sigma_B \approx 1$). More generally, one may show that the more patient a player is, the larger share he will get.

Exercise 39. Show that σ_B is increasing in δ_B and decreasing in δ_S by taking derivatives. Interpret!

Advantage of being the first proposer In case the two players are equally patient, $\delta_S = \delta_B = \delta$, the buyer's share is given by $\sigma_B = \frac{\delta}{1+\delta} < \frac{1}{2}$. The reason why the buyer receives less than half is that the seller has an advantage by being the one who has the right to give the first offer. But if the players are very patient or if the time between offers is very short, then $\delta \approx 1$ and the two negotiators get approximately equal shares.

7.2.4. Appendix: Infinite horizon

Consider now the possibility that the two parties could continue bargaining forever, without any end. In all other respects, however,

[6] Multiplying the same number $\delta_A \cdot \delta_B < 1$ to itself a large number (T) of times, gives a number close to zero. To see this note that $.9^2 = .81$, $.9^3 \approx .73$, $.9^4 \approx .66$, $.9^5 \approx .59$,..., $.9^{10} \approx .35$,..., $.9^{20} \approx .12$,..., $.9^{30} \approx .04$.

the game is as above. The two players, the seller and the buyer, can share €1 if they agree on how to split the money between themselves. In every odd-numbered period, the seller can make an offer and the buyer can either accept or reject the offer. If the offer is accepted, the game ends and the money is split according to the agreement. If the offer is rejected, the game continues to the next round. In every even period, the buyer can make an offer and the seller either accepts or rejects. The game thus starts in period 1 with the seller giving a bid. As long as no agreement is reached, the bargaining continues. The players only care about the amount of money they receive themselves and, being impatient, the time at which they receive the money. The seller's discount factor is $\delta_S < 1$ and the buyer's discount factor is $\delta_B < 1$. This means that the seller is equally happy if he receives €1 the next period or the lower amount €δ_S already today. There is complete and perfect information. That is, the players know each other's payoff functions and they can observe the complete history of bids and responses at any time.

Subgame Perfect Equilibrium

Finding a subgame perfect equilibrium is more difficult when bargaining could go on forever. When there is no last period, we cannot use backwards induction the way we are used to. As it turns out, however, we can work around this problem.

Efficiency First, we note that the bargaining must come to immediate agreement and that the agreement must be efficient.

Claim. The game must end with an agreement at some point.

To prove this claim, we assume that the claim is false and demonstrate that absence of an agreement would lead to a contradiction. Thus, assume that there is no agreement in equilibrium. Then, both players would receive zero payoffs. In any odd period, the

buyer would then accept any bid offering him at least zero. The seller could thus profitably deviate from the equilibrium by offering exactly that and take all the money himself. The game would thus end with an agreement, contrary to the assumption.

Claim. The game must come to an immediate agreement.

Assume not. Say that the agreement is reached in period $t > 1$. Then, the player making an offer in period $t-1$ could deviate by offering exactly the same agreement, but one period earlier. Since also the respondent is impatient, the offer would be accepted already at that time. The same logic demonstrates that an agreement would have to be reached a time $t - 2$, and so on.

Claim. Any agreement must be efficient.

Assume not. Say that the parties agree to take €1/4 each, leaving €1/2 on the table. Then, the bidder could then profitably deviate, continuing to offer €1/4 to the respondent and taking €3/4 for himself. The respondent would have the same incentives to accept this offer.

Split of surplus Next, we have to characterize exactly how the two parties divide the money. To do this, we need to contemplate what a rejection would imply, even if it will not occur in equilibrium. It is the respondent's ability to reject the offer that forces the bidder to give a reasonable bid. Unreasonable bids result in delays that impatient players wish to avoid. An offer must be constructed in such a way that the respondent prefers to accept it rather than to reject it. To compute acceptable offers we consequently need to compute the respondent's payoff in case of rejection. As it turns out, it is convenient to consider two rejections.

Assume that for some reason two offers have been rejected and that the players find themselves in period 3. Assume that an agreement in period 3 would give the seller σ_S^3 leaving the buyer with $\sigma_B^3 = 1 - \sigma_S^3$. Then, the buyer could already in period 2 offer

$\sigma_S^2 = \delta_S \cdot \sigma_S^3$. The seller would accept this offer since it is at least as good as rejecting the offer and reaching an agreement at time 3. The buyer would then take the rest of the money, i.e. $\sigma_B^2 = 1 - \delta_S \cdot \sigma_S^3$. But, then, the seller could already in period 1 offer $\sigma_B^1 = \delta_B \cdot \sigma_B^2 = \delta_B - \delta_B \cdot \delta_S \cdot \sigma_S^3$. The buyer would accept, leaving the seller with $\sigma_S^1 = 1 - \sigma_B^1 = 1 - \delta_B + \delta_B \cdot \delta_S \cdot \sigma_S^3$. The crucial point here is that if σ_S^3 is an equilibrium payoff for the seller in a subgame starting at time 3, we have now computed the equilibrium payoff for the seller in the game as a whole (the subgame starting at time 1). In particular,

$$\sigma_S^1 = f\left(\sigma_S^3\right) = (1 - \delta_B) + \delta_B \cdot \delta_S \cdot \sigma_S^3.$$

We must allow for the possibility that there are multiple equilibria with different payoffs. Since the function f is increasing, the highest equilibrium payoff to the seller in a subgame starting at time 1, denoted M_S^1 , and the highest equilibrium payoff to the seller in a subgame starting at time 3, denoted M_S^3, must satisfy $M_S^1 = f(M_S^3)$. Likewise, the lowest equilibrium payoffs must satisfy $m_S^1 = f(m_S^3)$.

Note also that the subgame starting in period 3 (where the seller makes the bid) is actually identical to the game as a whole, which starts with the seller making a bid in period 1.[7] Therefore, the set of equilibrium payoffs to the seller must be the same in the subgame as it is in the game as a whole. That is, $M_S^1 = M_S^3 = M_S$ and $m_S^1 = m_S^3 = m_S$. Both M_S and m_S are fixed points of f (i.e. solutions to the equation $x = f(x)$). But since f is strictly increasing, it can only have one fixed point, implying that $M_S = m_S$. Expressed differently, the seller's (and therefore the buyer's) payoff is uniquely determined in equilibrium.

[7] With a *finite* number of time periods, the subgame would be different since there would be fewer rounds until the game ends.

Solving the equation $\sigma_S = f(\sigma_S)$ for σ_S yields:

$$\sigma_S = \frac{1 - \delta_B}{1 - \delta_B \cdot \delta_S}.$$

The division of surplus is completely determined by the discount factors, i.e. the players patience.

Properties of the equilibrium First, note that if the two parties are equally impatient, i.e. if $\delta_S = \delta_B = \delta$, then the seller would get

$$\sigma_S = \frac{1}{1 + \delta}.$$

The reason why the seller gets more is that the seller has the opportunity to make the first bid. The buyer accepts a little less than half to avoid the delay. The seller exploits the buyer's weakness. If the two parties would become very patient, $\delta \to 1$, the first mover advantage would be of negligible effect, however.

Recall that the discount factor can be written $\delta = e^{-r\Delta}$ where Δ is the length between two offers and r is the continuous-time interest rate. There are thus two reasons why the players may divide the cake equally. Either bids and counter-offers can be made very quickly (i.e. $\Delta \to 0$) or the players' do not care about when they receive the money (i.e. $r \to 0$).

Second, note that the more patient the seller is, the larger will be his share of the cake. That is, a high δ_S leads to a high σ_S . Reversely, the more patient the buyer is, the smaller the seller's share will be.

Let r_S , r_B be the two players subjective continuous-time interest rates. Then, using l'Hôpital's rule, one may show that

$$\sigma_S \approx \frac{r_B}{r_S + r_B}$$

if the time between offers is very short.

Summary There is a unique subgame perfect equilibrium in the sequential bargaining game. An agreement is reached immediately and it is efficient. A party receives a larger share of the cake the more patient he is.

7.3. Nash Bargaining Solution

The extensive form bargaining game with alternating offers discussed above is instructive because it gives a stylized description of the bargaining process that resembles negotiations that all of us have been involved in many times. The main insight that we get out of studying that model is how important impatience is. It is only because people are impatient that they will reach an agreement rather than continue haggling forever. People are willing to accept a smaller amount today to avoid haggling. The model also describes in detail why a person's bargaining power is determined by how patient he is in relation to the other party. A negotiator who has an extra strong preference for receiving money today rather than tomorrow is willing to accept an extra low offer today to avoid haggling. The opponent will exploit this weakness and consequently also offer a lower amount to an impatient negotiator.

The drawback with the extensive form bargaining model is that it is quite tedious to analyze. It would be rather inconvenient to do all the calculations every time we want to analyze some bargaining situation. Luckily we do not have to. There is a reduced-form model that we can use instead. A reduced-form model is a model that can be used to find the prediction of a complicated model, but in a much simpler way. This reduced-form model is called the Nash bargaining solution.

To use the Nash bargaining solution, the first step is to describe the bargaining situation. Such a description must contain the following ingredients:

1. Who are the two players?
2. What contracts can they agree upon?
3. What payoffs would the two players receive from every possible agreement they could make?
4. What payoff will the players earn during the negotiation, before they reach an agreement? These payoffs are often referred to as the parties' threat points.
5. What is the relative bargaining power (i.e. relative impatience) of the two players?

Let us look at a couple of examples.

Bilateral monopoly - an example

Consider a negotiation in an intermediate goods market between a monopoly manufacturer and a monopoly retailer. Let us say that the retailer can sell any quantity of the product at the same price, v, and that for simplicity the retailer does not have any costs in addition to the cost of buying the goods from the manufacturer. The manufacturer's cost of producing the goods is given by $C(q) = c \cdot q$ up to some capacity constraint, k. Let us assume that both firms' objectives are to maximize their own profits.

Bargaining situation The first step in finding the Nash bargaining solution is to describe the bargaining situation:

1. The players are the retailer and the manufacturer.
2. They can agree on any contract (p, q) that specifies a price, p, per unit and a quantity below the manufacturer's capacity constraint, $q \leq k$.

3. The retailer's payoff function is $\pi_R(p,q) = (v-p)\cdot q$. The manufacturer's payoff function is $\pi_M(p,q) = (p-c)\cdot q$.

4. Before an agreement both parties earn zero profit: $\tilde{\pi}_R = 0$ and $\tilde{\pi}_M = 0$. That is, the threat point is zero profit for both parties.

5. The two parties are equally patient and therefore have the same bargaining power.

Nash product The second step to find out what contract (p,q) the two parties will agree upon, is to define a function called the Nash product. In our example it is given by

$$N(p,q) = [\pi_R(p,q) - \tilde{\pi}_R] \cdot [\pi_M(p,q) - \tilde{\pi}_M].$$

Note that if the two parties agree on a contract (p,q) the retailer's gain would be $\pi_R(p,q) - \tilde{\pi}_R$ which is the first factor on the right-hand side. The manufacturer's gain would be $\pi_M(p,q) - \tilde{\pi}_M$ which is the second factor. Thus the Nash product is simply the product of the two parties' gains from making the agreement. (If one party would be more patient that the other, the Nash product would be a little bit more complicated but not much.)

Deriving the prediction The third step is surprising. The contract that would be predicted by the extensive form bargaining model can be found by simply maximizing the Nash product with respect to the contract, i.e. the price (p) and the quantity (q). That is, instead of solving a complicated game, we solve a much simpler maximization problem. Let us try.

Plugging in the payoff functions and the threat points, the Nash product is given by:

$$N(p,q) = [(v-p)\cdot q - 0] \cdot [(p-c)\cdot q - 0],$$

which can be simplified to

$$N(p,q) = (v-p)\cdot(p-c)\cdot q^2.$$

Let us start with predicting what quantity the two parties will agree upon. Clearly the two parties will make an agreement on some price between v and c so that both parties make some profit. Consequently $(v-p)\cdot(p-c) > 0$ which means that the Nash product, $N(p,q)$, is increasing in the quantity, q. Thus, the Nash product is maximized by setting the quantity as large as possible. That is $q = k$. This means that the Nash bargaining solution predicts that the two parties will agree to trade the efficient quantity. This is also what we know is predicted by the extensive form bargaining game. The two predictions are the same.

Let us now predict what price the two parties will agree upon. To find the price we use the first-order condition, which is given by

$$\frac{\partial N(p,q)}{\partial p} = -(p-c)\cdot q^2 + (v-p)\cdot q^2 = 0.$$

Dividing the first-order condition by $q > 0$ and rearranging the terms gives $v - p = p - c$ which simply means that the two parties will agree upon the price that will make the retailers gain from the agreement $v-p$ equal to the manufacturer's gain $p-c$. The reason why they split the gains equally is of course that they are equally patient so that they have equal bargaining power. Again, we see that the Nash bargaining solution finds the equilibrium agreement in the extensive form bargaining game. More specifically, the agreed price is given by

$$p = \frac{v+c}{2}.$$

That is, the retailer hands over half of what he earns when he sells the product to the final consumer. In addition, the retailer also pays half of the manufacturer's costs.

The conclusion is that by maximizing the Nash product we pre-

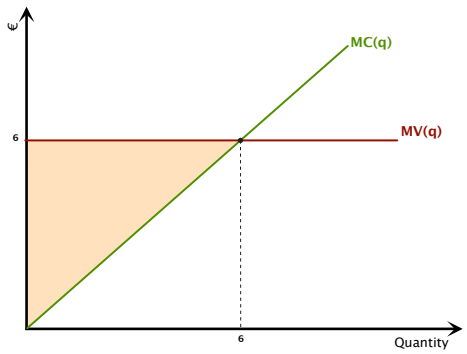
dict that the two firms will agree on the quantity that maximizes the sum of the two firms' profits and then they will split this maximum surplus equally. This is the same prediction that would have come out from analyzing the full extensive form game.

7.4. Bilateral market power

7.4.1. Bilateral monopoly

Let us now look at a somewhat more realistic but slightly more complicated example of bilateral monopoly. The only difference is that the manufacturer's cost of producing the goods is given by $C(q) = \frac{c}{2} \cdot q^2$. That is the manufacturer's marginal cost is increasing and given by $MC(q) = c \cdot q$. The average cost of producing q units is given by $AC(q) = \frac{c}{2} \cdot q$. It is also instructive to define the retailer's revenue function $R(q) = v \cdot q$. Clearly the marginal revenue is constant and given by $MR(q) = v$. This market is described by Figure 7.2 assuming that $v = 6$ and $c = 1$.

Figure 7.2.: Buyer power



Bench mark If the manufacturer and the retailer had been two divisions within the same vertically integrated firm, the integrated

firm's profit would have been given by $\pi_I(q) = R(q) - C(q)$. In order to maximize its profits, such an integrated firm would produce a quantity equalizing marginal cost and marginal revenue, i.e. $\frac{\partial \pi_I(q)}{\partial q} = MR(q) - MC(q) = 0$. Plugging in the cost and revenue functions, the first order condition is given by $v - c \cdot q = 0$ which means that $q = v/c$. In the numerical example of the figure, $q = 6/1 = 6$.

Bargaining situation But since the manufacturer and the retailer are two independent firms, they have to bargain about how much to produce and about the price that the retailer should pay the manufacturer. We describe this negotiation by the Nash bargaining solution.

1. The players are the retailer and the manufacturer.
2. They can agree on any contract (p, q) that specifies a price, p, per unit and a quantity, q.
3. The retailer's payoff function is $\pi_R(p, q) = R(q) - p \cdot q = (v - p) \cdot q$. The manufacturer's payoff function is $\pi_M(p, q) = p \cdot q - C(q) = p \cdot q - \frac{c}{2} \cdot q^2$.
4. Before an agreement both parties earn zero profit: $\tilde{\pi}_R = 0$ and $\tilde{\pi}_M = 0$. That is, the threat point is zero profit for both parties.
5. The two parties are equally patient and therefore have the same bargaining power.

Nash product and predictions Now the Nash product is given by:

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

The first-order condition for the price is given by

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

Dividing the first-order condition by $q > 0$ gives

$$[p \cdot q - C(q)] = [R(q) - p \cdot q].$$

This condition says that the price should be chosen to make the retailer's and the manufacturer's profits equal. That is the two parties split the surplus from trade equally. Solving the first-order condition for the price reveals that

$$p = \frac{R(q)/q + C(q)/q}{2} = \frac{v + AC(q)}{2}.$$

The retailer will thus pay a price per unit that is equal to half of the average revenues plus half of the average cost. This is essentially the same conclusion as in the previous example. The only difference is that in the current example, average cost depends on the quantity produced.

The first-order condition for the quantity is given by

$$\frac{\partial N(p,q)}{\partial q} = [MR(q) - p] \cdot (p \cdot q - C(q)) + (R(q) - p \cdot q) \cdot [p - MC(q)] = 0.$$

Recall that the price will be set to guarantee that $p \cdot q - C(q) = R(q) - p \cdot q > 0$. Thus, the expressions within the rounded brackets are actually equal. Thus we can divide the first-order condition by this amount. And the first-order condition can be written

$$MR(q) = MC(q).$$

That is, the two parties will choose the same quantity as an integrated firm would have chosen. This is the efficient quantity from the point of view of the two firms.

In the example of the figure, the producer and the retailer will thus agree to trade six units and to split the corresponding surplus of €18 equally. This implies that the retailer must pay the producer a total price of €27 for the six units. The price per unit is €4.50 (= 27/6). The total cost of producing the 6 units is $18 = \frac{1}{2} \cdot 6^2$ and the average cost is $3 = 18/6$. Thus the price is given by $p = (v + AC)/2 = (6 + 3)/2 = 4.5$.

Note that the retailer pays a price per unit that is below the Walrasian price, which is €6, that would arise under perfect competition. This may be viewed as a definition of buyer power, i.e. market power exercised by a buyer.

Also note that at a per unit price of €4.50, the seller would prefer to sell less than the six units. Such a reduction in the quantity would not be allowed, however, since the seller is bound to sell six units according to the contract with the retailer.

7.4.2. Large buyers have more buyer power

An even more interesting exercise is to compare the prices paid by two different retailers buying different amounts from the same seller. To make this comparison we assume for simplicity that both retailers have the same constant marginal valuation equal to €6 and that the producer's marginal cost is the same as above.

When there is one seller and two buyers, the seller will have to negotiate with both buyers separately. We will not go through this analysis here, but simply assume that the two negotiations end up with the following outcome:[8]

1. The total quantity traded is efficient, i.e. six units.

2. One retailer buys more than the other, perhaps because they sell in two different towns of different sizes. Lets say that

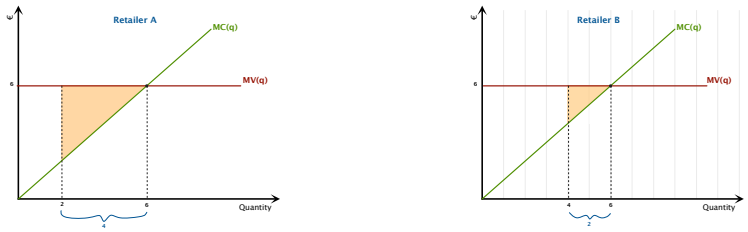
[8] One may demonstrate that two parallel negotiations will indeed result in this outcome under certain conditions. Proving this is beyond this course, however.

retailer A buys four units and that retailer B buys two units.

3. The manufacturer splits the *bilateral surplus* equally with each retailer. The bilateral surplus is simply the surplus created by trading with that particular retailer, taking all other sales as given.

The negotiation with retailer A can then be summarized by the left panel and the negotiation with retailer B can be summarized by the right panel of the Figure 7.3.

Figure 7.3.: One manufacturer and two retailers



(a) Retailer A (b) Retailer B

What is important to note is that the seller treats both retailers as a marginal buyer. That is, when negotiating with retailer A, the producer knows that he will also sell two units to retailer B. (The seller may have two representatives bargaining with the two buyers and the representative who negotiates with retailer A knows that the representative who negotiates with retailer B will agree to sell two units. Recall that an equilibrium is a common understanding of how all players will behave.) Thus, the seller's first two units of the capacity are so-to-say already booked. Producing the additional four units will thus start a bit up on the marginal cost curve. The total cost of producing the four extra units is €16 (the area below the marginal cost curve, between $q = 2$ and $q = 4$) and the bilateral surplus is €8 (the tangerine colored area between the MV and MC-curves, between $q = 2$ and $q = 4$). As a result total price for the

two units will be €20, corresponding to a unit price of €5.

Likewise, when negotiating with retailer B, the producer knows that he will sell four units to retailer A. The unit price will be €5.50.

Exercise 40. Explain why retailer B's price is €5.50.

The conclusion is thus that the retailer buying less will have to pay a higher price. The reason is that the retailer buying less so-to-say meets the manufacturer higher up on his marginal cost curve. The last few units are more expensive to produce. Buying many units means that the average cost of producing these units (the average incremental cost) will be lower.

Also note that if the two retailers would start a joint purchasing collaboration, the market would become a bilateral monopoly. Then, both would only have to pay €4.50 as demonstrated above. Achieving such additional buyer power is probably one of the important driving forces behind the evolution towards higher concentration in retailing observed in Sweden as well as in many other countries.

7.4.3. Appendix: Formal analysis of buyer power

A single manufacturer sells its product to two different retailers, called A and B. Let us assume that the two retailers are active in two different towns and that they do not compete with each other. We will study the negotiation between the manufacturer and retailer A and simply assume that it is known that the manufacturer sells q_B units to retailer B. The manufacturer's cost of producing the goods is given by $C(q_A + q_B) = \frac{c}{2} \cdot (q_A + q_B)^2$. That is the manufacturer's marginal cost is increasing and given by $MC(q_A + q_B) = c \cdot (q_A + q_B)$. To analyze the negotiation between the manufacturer and retailer A it is instructive to define the manufacturer's *incremental cost* of serving retailer A, which is

defined as

$$IC_A(q_A, q_B) = C(q_A + q_B) - C(q_B).$$

The incremental cost is simply the extra cost that the manufacturer has to incur to service retailer A over and above the cost of servicing retailer B. With the quadratic cost function, incremental cost is

$$IC_A(q_A, q_B) = \frac{c}{2} \cdot (q_A + q_B)^2 - \frac{c}{2} \cdot (q_B)^2$$

The average incremental cost is defined as

$$AIC_A(q_A, q_B) = \frac{IC_A(q_A, q_B)}{q_A}.$$

The average incremental cost is simply the incremental cost divided by the amount produced for retailer A. The marginal incremental cost is actually the same as the marginal cost, that is

$$\frac{\partial IC_A(q_A, q_B)}{\partial q_A} = \frac{\partial [C(q_A + q_B) - C(q_B)]}{\partial q_A} = \frac{\partial C(q_A + q_B)}{\partial q_A} = MC(q_A + q_B).$$

If the manufacturer would not make any agreement with retailer A, the manufacturer would earn the profit $\tilde{\pi}_M = p_B \cdot q_B - C(q_B)$. This is also the profit that the manufacturer earns during the negotiation with retailer A before they reach an agreement. Therefore it is the manufacturer's threat point in the negotiation with A.

If the manufacturer reaches an agreement with retailer A on a contract (p_A, q_A), the manufacturer's profit is given by $\pi_M(p_A, q_A, p_B, q_B) = p_A \cdot q_A + p_B \cdot q_B - C(q_A + q_B)$. By adding and subtracting the same term $C(q_B)$, the manufacturer's profit can be rewritten as

$$\pi_M(p_A, q_A, p_B, q_B) = p_A \cdot q_A - [C(q_A + q_B) - C(q_B)] + p_B \cdot q_B - C(q_B)$$

Using the definition of incremental cost and recalling that the manufacturer's profits from dealing with retailer B is equal to the man-

ufacturer's threat point, the profit can be simplified to

$$\pi_M(p_A, q_A, p_B, q_B) = p_A \cdot q_A - IC_A(q_A, q_B) + \tilde{\pi}_M.$$

Clearly the manufacturer's profit also depends on its contract with retailer B, but we suppress this dependence to simplify notation. Now the Nash product describing the negotiation between the manufacturer and retailer A can be defined:

$$N_A(p_A, q_A) = [\pi_M(p_A, q_A) - \tilde{\pi}_M] \cdot [\pi_A(p_A, q_A) - \tilde{\pi}_A].$$

Plugging in the relevant expressions:

$$N_A(p_A, q_A) = [p_A \cdot q_A - IC_A(q_A, q_B)] \cdot [R(q_A) - p_A \cdot q_A].$$

This expression is almost the same as in example 2. The only difference is that the manufacturer's cost is replaced by the manufacturer's incremental cost (i.e. extra cost) of dealing with retailer A. It is therefore easy to see that also the solution must be the same as above when cost is replaced by incremental cost.

The manufacturer and retailer A will agree on a quantity that equates marginal revenue and marginal cost (= marginal incremental cost), i.e. $MR(q_A) = MC(q_A + q_B)$. This expression can be simplified to $v = c \cdot (q_A + q_B)$ which means that the manufacturer will produce the same total amount as in the case when there is only one retailer. Consequently $q_A = v/c - q_B$, which means that the more the manufacturer sells to retailer B, the less the manufacturer will agree to sell to retailer A. The reason is the manufacturer's marginal cost is increasing in the total amount produced. That is, the more the manufacturer sells to retailer B the higher will be the cost of producing for retailer A. When production is more costly, less will be produced.

The manufacturer and retailer A will agree on a price that equates retailer A's profit to the manufacturer's extra profit of dealing with

retailer A, i.e.:

$$p_A \cdot q_A - IC_A(q_A, q_B) = R(q_A) - p_A \cdot q_A.$$

This means that the price is

$$p_A = \frac{R(q_A) + IC_A(q_A, q_B)}{2 \cdot q_A} = \frac{v + AIC_A(q_A, q_B)}{2}.$$

Similarly, in the negotiation with retailer B, the manufacturer will agree on the price

$$p_B = \frac{v + AIC_B(q_A, q_B)}{2}.$$

Now it is easy to see that retailer A will pay a lower price if and only if $q_A > q_B$. That is, the retailer buying more, gets the lower price.[9]

7.4.4. Market power at the retail level

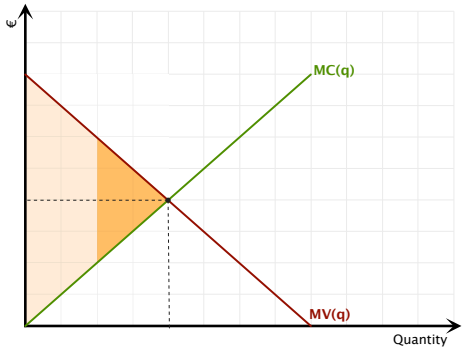
Most large retailer chains probably have some amount of market power also vis- à-vis the final consumers. If they do, their MV-curves will slope downwards, since the more they buy and resell, the lower they have to set their retail prices.

If retailers have market power over consumers, it is not certain that a retailer buying more will receive a better wholesale price than a retailer buying less from the same supplier. On the one hand, the retailer buying much will benefit from a lower average production cost, just as above. Other other hand, this retailer also has a higher

[9] To see this, note that $p_A < p_B$ if $AIC_A(q_A, q_B) < AIC_B(q_A, q_B)$ which is true if $\frac{IC_A(q_A,q_B)}{q_A} < \frac{IC_B(q_A,q_B)}{q_B}$ which is true if $\frac{\frac{c}{2}\cdot(q_A+q_B)^2 - \frac{c}{2}\cdot(q_B)^2}{q_A} < \frac{\frac{c}{2}\cdot(q_A+q_B)^2 - \frac{c}{2}\cdot(q_A)^2}{q_B}$ which is true if $\left[(q_A+q_B)^2 - (q_B)^2\right] \cdot q_B < \left[(q_A+q_B)^2 - (q_A)^2\right] \cdot q_A$ which is true if $[(q_A+q_B) - (q_B)] \cdot [(q_A+q_B) + (q_B)] \cdot q_B < [(q_A+q_B) - (q_A)] \cdot [(q_A+q_B) + (q_A)] \cdot q_A$ which is true if $[q_A + 2 \cdot q_B] < [2 \cdot q_A + q_B]$ which is true if $q_B < q_A$.

average valuation of the goods bought. It is an empirical question which of these two forces dominates the other. In the example depicted by the Figure, where (the absolute value of) the slope of

Figure 7.4.



the MV-curve is the same as the slope of the MC-curve, it is easy to see that price per unit would be the same, independent of if the retailer would buy (say) two units (implying a surplus equal to the dark-tangerine colored triangle) or (say) four units (implying a surplus equal to the tangerine colored triangle). Actually, in both cases, the price per unit would be equal to the Walrasian price (equal to €4 in this figure). This per-unit price spits the total surplus into two equal parts (one triangle above the price-line and one below).

The fact that grocery retailers buying more from the same supplier of groceries appear to pay a lower price (as reported above) may thus be taken as an indication that market power in grocery retailing is relatively small, at least in comparison to the decreasing returns to scale in production of the same goods.

7.5. Summary

In many markets there are only a few buyers and a few sellers. In the extreme, there is bilateral monopoly. It is then not possible for the seller to simply dictate a price. They buyer knows that if he would decline the offer the seller would probably be willing to make concessions, since the seller doesn't have any alternative trading partner. The trading conditions will be determined through a negotiation. Several questions arise:

1. Under what conditions will the parties come to an agreement?
2. How much will they trade?
3. How much time will they spend haggling?
4. How will the parties split the surplus from the transaction, i.e. what price will they agree upon?

To analyze this and other situations of bargaining, economists often use the so-called Rubinstein-Ståhl bargaining model. This model assumes that there are two parties who can share some amount of money (the surplus from trade), if they reach an agreement about how to split the money between themselves. The two parties make alternating offers: one party gives the first offer, and if the other party rejects the offer, he is allowed to give a counter-offer the next period. And so on, for some predetermined number of periods. Both parties are assumed to only care about how much money they receive themselves and about when they get their money. Since this is an extensive form game with perfect and complete information, we can find the subgame perfect equilibrium by backwards induction.

As it turns out there is a unique subgame perfect equilibrium. An agreement is reached immediately and it is efficient, i.e. they agree to trade the efficient quantity. The reason why agreement is immediate is that the players are impatient. They prefer a little

less money today to a little more tomorrow. Therefore, to avoid delays, whoever makes an offer will make "reasonable" claims.

The main determinant of the split is the players' relative patience. In particular, a party receives a larger share of the cake the more patient he is. In addition to the relative patience, also some details of the bargaining process appear to matter for the split. First, a party receives a larger share if he is allowed to make the first proposal. However, if the parties are very patient or if offers and counteroffers can be made very quickly, the advantage of being first is negligible. Then, it doesn't cost the first respondent much to reject an offer to get the right to make a counteroffer. Second, a party also receives a larger share if he is allowed to make the last proposal. However, if the haggling can continue for a very long time, this advantage is also negligible. Hence the parties will split the surplus equally among themselves if (i) bargaining can go on for a very long time, and (ii) the parties are very patient or if offers and counteroffers can be made very quickly, and (iii) the parties are equally patient.

The equilibrium of the extensive form bargaining game can be found by analyzing a reduced-form model instead. This reduced form model is called the Nash bargaining solution.

This bargaining theory can explain why large retailers buying a substantial share of their suppliers total output may sometimes achieve a lower price than smaller retailers buying only a small share of the suppliers output. The reason is that the supplier has increasing marginal cost of production and that the supplier treats both retailers as trading on his margin. An equal split of the surplus implies that a buyer pays the average incremental cost of producing the units he buys (and in addition an amount to split the difference between value and cost). The average incremental cost of supplying a retailer buying only a few units is higher than the average incremental cost of supplying a retailer buying a large quantity since marginal cost is increasing.

8. Auctions

8.1. Why study auctions?

Governments use auctions to sell bonds to investors and radio frequencies to mobile phone companies. Auctions can also be turned around and used for procurement. Manufacturing firms buy parts from suppliers and Governments appoint entrepreneurs for maintaining roads through auctions. Many people use auctions to buy and sell property, second-hand and antique furniture as well as fine art. Why are auctions so commonly used under so varied circumstances? Part of the answer is lack of information.

8.1.1. Government sales of radio spectrum

Sweden switched from analogue to digital TV in 2008. As a result some of the radio spectrum that was previously used for TV-transmission was freed and can now be used for other purposes. The Government decided to use this spectrum, the 800 MHz band, for high-speed mobile broadband, using 4G-technology. This technology will already in the near future allow speeds up to 150 Mbps and may later reach 1000 Mbps, to be compared with a maximum of 10 Mbps in the 3G-networks.

The Swedish telecom regulator, PTS, then had to allocate the 800-band to companies that want to provide high-speed broadband. How should PTS do that? The Swedish Telecom Act (Lagen om elektronisk kommunikation) instructs the PTS to allocate the spectrum in such a way that the social surplus is maximized. That

is, the objective is to make sure that the citizens get the broadband services that they value the most and that these services are produced in a cost efficient way. The objective is not to sell the spectrum at the highest possible price. The problem for PTS is of course that they don't know which firms would be the most efficient users of the spectrum. Which firms have the best ideas about how to market mobile broadband services to consumers? Which firms are most efficient in building up and maintaining a reliable infrastructure at a low cost? Due to this lack of knowledge, PTS cannot simply "pick the winners." Instead, PTS decided to use an auction.

The auction was held during 2011 with telecom companies, such as Telia, Telenor, Tele2 and Tre, bidding for the spectrum. Participating in such an auction is quite challenging. Even if each company knows how they would use the spectrum themselves, they do not know exactly what plans the rivals might have. They may be able to compute what the spectrum is worth for themselves, but it is much more difficult to compute what it would be worth for the competitors. It is a challenging task to decide how much to bid. A high bid increases the chance of winning the auction. The disadvantage is that value of winning is lower. The question is how to solve this tradeoff.

Given that not even the firms know which firms would be the most efficient users of the spectrum, one may wonder why PTS chose to auction the spectrum? The reason is that PTS believes that the auction process itself will actually answer exactly that question. PTS believes that it is the most efficient firms that will win the auction, despite the fact that these firms themselves do not know who they are beforehand. The question is if PTS is justified in this belief?

8.1.2. Private sales of apartments

Do you want a larger apartment? Then, the best way for you to sell your old apartment is to offer it slightly below the highest price anyone is willing to pay. The person with the highest valuation will then voluntarily come forward, ask for the apartment and pay the price.

But most sellers of apartments, fine arts and antiques do not know what other people are willing to pay for their objects. They don't even know what the highest valuation is. Then, an auction might be the solution. An auction allows all prospective buyers to tell how much they value the object. The one who claims to have the highest valuation gets the good and pays accordingly. The auction is a tool sellers can use to extract the buyers' private information.

Is auctioning a good remedy to the sellers' lack of information? Clearly, if all the buyers would really tell (bid) their true valuations, the person who values the object the most will also get it. The market would then be efficient. And if the buyer pays his own bid, the seller would appropriate the whole surplus from this efficient deal.

So, do people "tell the truth" in auctions? Do they reveal their true valuations? To answer this question we need to study the buyers' bidding behavior in detail.

8.2. Auction Designs and Types of Uncertainty

The auctioning game is shaped by the rules of the auction and the way peoples' valuations relate to each other.

Auction designs Different auctions have different rules. One common design is the "open cry" or *English* auction. The buyers are typically gathered in a room and the auction is started when

the auctioneer announces a starting price, which is lower than the expected selling price. The bidders then call out their increasingly higher bids until no one wants to go any further. The object is sold to the highest bidder who pays his bid. In many cases, people give their bids in writing and then the auctioneer announces the bids that have been made so far (or at least what the current highest bid is) and then people can increase their bids if they so wish. In a clock or *Dutch* auction, the auctioneer begins with a high price above the expected selling price and then gradually lowers it until one of the participants accepts it. The buyer pays the price he accepted. The best-known example of such an auction is the Dutch flower auctions. The English and the Dutch auctions are similar in that they have a sequential structure. The bids are publicly announced and the participants in the auction can observe what the competitors' behavior and react to that. In a sealed-bid auction, each prospective buyer hands in a written bid, simultaneously with all other buyers. The object is sold to the highest bidder. In a *first-price sealed-bid* auction, the winner pays the price he offered. In a *second-price sealed-bid* auction (also called Vickrey auction) the winner only pays the second highest bid.

The best way to bid in an auction depends on the design. For simplicity, we will only study sealed bid auctions in this course. In sealed bid auctions, the bidders are only asked to make one decision and we need not think about how they react on other peoples' bids in sequential auctions.

Types of uncertainty Oil companies bidding for the drilling rights in a certain area would all make similar profits if winning the auction and if there turns out to be oil in the area. But, before the drilling starts, the oil companies' geological experts may have different information about probability of finding any oil. That is, in some auctions all the buyers have the same or at least very similar valuations of the object, but their information about this common

value differ. Such auctions are called *common value auctions*.

In fine art auctions different people simply value the object for sale differently. Some people like naturalistic paintings, while others prefer more abstract work. Some people prefer paintings to photography. Other people disagree. Certain people are wealthier than others and are therefore willing to pay more for fine arts. Auctions with such differences in valuations are called *private value auctions*.

The best way to bid in an auction depends on how the bidders' valuations relate to each other, if valuations are private or common. This is commonly referred to as the type of uncertainty. Are the bidders uncertain about their rivals' intrinsic valuation of the good, or about what information their rivals may possess? For simplicity, we will only study private value auctions in this course.[1]

8.3. How to predict bidding behavior

Each bidder's best behavior depends on how all the other bidders behave – their decisions are interdependent. To study the buyers' bidding behavior we therefore need to use the theory of interdependent decision-making, also known as game theory. The main teaching of game theory is that people ought to analyze a situation with interdependent decisions as a whole—at the same time generating a prediction of other peoples' behavior and a recommendation for one's own behavior.

Modeling the information asymmetry Consider an auction with two bidders. Let's call them Fredrik and Mona. Fredrik knows that he values the good highly, say at v_H. And Mona knows what her own valuation is, but she doesn't know that Fredrik is willing

[1] We will also assume that all bidders have the same information about their rivals' valuations. In technical terms, we will consider the case of *independent* private values.

to pay at most v_H for the object.

We are only going to look at information asymmetry which is relatively easy to analyze. To describe Mona's uncertainty about Fredrik's valuation, we assume that there is some set of possible valuations, say v_H and v_L, and that Fredrik's valuation is high with some given probability, p, and that it is low with the remaining probability, $1 - p$. All this is common knowledge to Fredrik and Mona. How can they know all this? A possible reason is that both Fredrik and Mona know the population distribution of different valuations. By talking to their friends they have both discovered that people either seem to value the object at v_H or v_L and that the frequency of high valuations is p. It may be clear to both Fredrik and Mona that the best way to think about the rival is that the he or she is randomly drawn from this population distribution.[2]

Strategies It seems safe to assume that Fredrik's bid depends on his valuation of the good. The higher he values the good, the more he will be inclined to bid. And since Mona doesn't know Fredrik's valuation, Mona is not going to be able to predict Fredrik's bid exactly. So how should Fredrik and Mona (and we) analyze the situation?

The key is to change the notion of strategy. Since we understand that Fredrik's bid will depend on his valuation, it is convenient to also view a strategy as a rule (a function) that prescribes a bid for every possible valuation that a buyer might have. A strategy for Fredrik would thus be a list:

- If my valuation is v_L, then bid b_L.
- If my valuation is v_H, then bid b_H.

[2]In a more realistic model, the buyers may believe that people with a higher valuation are more likely to show up at the auction. Such a model would also be more complicated and will not be discussed here.

The list can also be written in a more compact form as $b_L = b(v_L)$ and $b_H = b(v_H)$, where the function symbol "b" simply stands for "my bid if my valuation is". Even more compactly the list can be written (b_L, b_H). That is Fredrik's strategy is simply two numbers. The first number tells him what to bid if his valuation is low and the second number tells him what to bid if his valuation is high.

This may, however, seem a bid odd. Even if Mona doesn't know Fredrik's valuation, Fredrik himself knows that he values the good highly, at v_H. Why should Fredrik choose a strategy that also tells him how to bid in case his valuation would have been low, v_L? The reason is that Mona doesn't know Fredrik's valuation. As far as Mona knows, Fredrik's valuation could be either v_H or v_L. In order for Mona to make a prediction of Fredrik's behavior, Mona needs to consider both cases. Only by knowing the possible bids by Fredrik, and their probabilities, Mona can compute her optimal bid, i.e. the bid that maximizes her expected utility. But Fredrik also needs to predict Mona's bid, in order to compute his own optimal bid. Clearly, in order to predict Mona's bid, Fredrik needs to consider what he himself would have bid in case he would have had a low valuation of the object. That is:

> *Even though Fredrik knows perfectly well that his valuation is v_H he actually needs to think hard about what he would have done in case his valuation had been v_L.*

Otherwise, Fredrik will not be able to predict Mona's behavior.

We may therefore think of Fredrik's choice as a two-step procedure. As a first step, Fredrik selects a strategy, $b = (b_L, b_H)$. As a second and final step, Fredrik simply selects the relevant bid from the list, i.e. b_H.

Defining a game Since we only study sealed bids auctions, in which each buyer only makes one bid and the bids are simultaneous, it is convenient to use a normal form game:

- The **players** are the buyers, Fredrik and Mona.
- A **strategy** for Fredrik is a rule specifying a bid for every possible valuation Fredrik might have. Fredrik's strategy set is the set of all such rules. Mona's strategy set is constructed in a similar way.
- Fredrik's **payoff function** tells us Fredrik's expected utility for all possible combinations of strategies that Fredrik and Mona can use. Mona's payoff is constructed in a similar way.

We will derive these payoffs below.

Equilibrium An equilibrium of the auction game is a pair of strategies, one for Fredrik ($b_{Fredrik}$) and one for Mona (b_{Mona}). Fredrik's strategy must be a best reply to Mona's strategy and vice versa.

Fredrik's strategy is a best reply to Mona's if two conditions are satisfied.

1. First, the bid $b_{Fredrik}(v_H)$ should maximize Fredrik's expected utility, in case Fredrik's valuation is v_H, given that Mona uses her strategy b_{Mona}.
2. Second, the bid $b_{Fredrik}(v_L)$ should maximize Fredrik's expected utility, in case Fredrik's valuation is v_L, given that Mona uses her strategy b_{Mona}.

Mona's strategy is a best reply if similar conditions are fulfilled.

Terminology Before we continue, it is useful to introduce a few terms. When the players of a game don't know the other players' payoff functions, we have a game of *incomplete information*. An auction is thus an example of a game of incomplete information. More precisely, sealed-bid auctions are static games of incomplete information, since all players move only once and all these moves are

simultaneous. The trick to analyze games of incomplete information is to define strategies as functions. More precisely, the strategy must instruct the player what to do, depending on his "type." In a private value auction a player's type is simply his valuation. When incomplete information is modeled as it is here, we say that we have a *Bayesian game*. A Nash equilibrium of a Bayesian game is often called a Bayesian equilibrium or a *Bayesian-Nash equilibrium*.

8.4. Second-Price Sealed-Bid Auction

Assume that there is one indivisible unit of the good and that there are two buyers, called 1 and 2. Both buyers known their own valuations of the good. Buyer 1's valuation is denoted v_1 and buyer 2's is v_2. The buyers are required to submit their bids simultaneously. The buyer submitting the highest bid wins the object and pays a price, p, equal to the second highest bid. That is, if buyer one bids b_1 while buyer 2 only bids $b_2 < b_1$, then buyer 1 gets the good and pays the price $p = b_2$.

A buyer's utility is $v_i - p$ if he wins the object and zero otherwise.

The buyers do not know their rival's valuation of the object. To model this uncertainty, we may for example assume that both buyers' valuations are uniformly distributed between zero and one. But, as you will soon see, the exact distribution doesn't matter in second-price auctions.

Equilibrium There is a unique Bayesian equilibrium in the second-price sealed-bid auction. The equilibrium requires a buyer to bid his valuation, i.e. $b_i = B_i^*(v_i) = v_i$. Expressed differently, the bidders report their own valuations truthfully in equilibrium. In fact, bidding one's true valuation is a weakly dominating strategy. That is, it does not matter what strategy the rival is using, it is always a best reply to bid one's true valuation.

To see this, consider buyer 1's decision:

- Assume that the rival's bid for some reason happens to be below player 1's valuation, i.e. $b_2 < v_1$. Then, any bid $b_1 > b_2$ maximizes buyer 1's utility. As long as buyer 1 bids more than the rival, he gets the good. And, it does not matter how much he bids, the price is given by the rival's bid. Thus, for example $b_1 = v_1$ is a maximizing bid.
- Assume that the rival's bid for some reason happens to be above player 1's valuation, i.e. $b_2 > v_1$. Then, any bid $b_1 < b_2$ maximizes buyer 1's utility. As long as buyer 1 bids less than the rival, he is not going to get the good. Thus, for example $b_1 = v_1$ is a maximizing bid.
- Assume that the rival's bid for some reason happens to be equal to player 1's valuation, i.e. $b_2 = v_1$. Then, any bid b_1 maximizes buyer 1's utility. Thus, for example $b_1 = v_1$ is a maximizing bid.

Note that in the first case it does not matter why $b_2 < v_1$. The reason could be that buyer 2 has a low valuation of the good and reports this valuation truthfully. Another reason could be that buyer 2 has a high valuation, but that he under-reports his true valuation. This means that bidding your valuation is a best reply to any bidding behavior by the opponent (and also independent of what the probability is that the opponent has different valuations). That is, bidding your valuation is a weakly dominating strategy.

In conclusion, all participants in a second-price sealed-bid auction have an incentive to bid their true valuations of the good. Intuitively, the reason is that they do not have to pay a higher price if they increase their bid.

Social welfare What does this behavior imply for social welfare? Since all players bid their true valuations and since the good goes to the highest bidder, the market is efficient. That is the participant with the highest valuation gets the good.

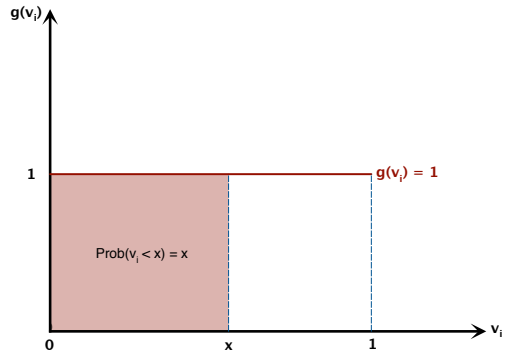
Moreover, since the price is only the second highest bid, the buyer will get a surplus from the transaction. And since sellers will only offer their goods in an auction if their own valuation is (much) lower than the expected price, also the sellers will typically earn a surplus. In fact, the buyer and the seller will share the social surplus from the transaction. All other participants are left without any gains, however.

8.5. First-Price Sealed-Bid Auction

Assume now that the winner has to pay his own bid. Then, intuitively, the bidders will not have an incentive to bid their true valuations, since bidding high involves a trade-off. A high bid not only increases the probability of winning the good, which is good, but it also means that the bidder has to pay a higher price in case he wins the auction. Therefore it seems reasonable to expect that the participants in a first-price auction will bid below their valuation of the object.

As it turns out, first-price auctions are more difficult to analyze. Part of the reason is that the exact distribution of possible valuations matter in first-price auctions. To keep the analysis simple we assume that both players' valuations are uniformly distributed between zero and one, that is to say any value between zero and one is equally likely. Or, rather, the probability that the valuation is in some interval is proportional to the length of the interval. For example, the probability that the valuation is below some number x, is x. (Examples: The probability that the valuation is below 1 is 1. We know that it is below one for sure. The probability that the valuation is below 0 is 0. We can rule out valuations below zero. The probability that the valuation is below 1/2 is 1/2. It is equally likely to be over a half.) Such a uniform distribution is described in Figure 8.1.

Figure 8.1.: Uniform distribution



The game - a sketch We will consider an auction with two participants. The two participants are thus the players of the game.

A strategy is a function prescribing a bid for every possible valuation that the player might have.[3] One possible strategy to always bid zero, independent of one's valuation. Another strategy would be to give a bid equal to one's true valuation. Yet another possibility would be to bid the square of the valuation. There are many possibilities. It would be impossible list them all. In fact, it is extremely hard to even imagine all possible strategies that a player could use in a first-price auction. We have to find a work-around.

Since the players do not know their rival's valuation of the object, they will not be able to predict the rival's exact bid. Therefore they will not be able to say with certainty if a certain bid is sufficiently high to win the auction. Given any bid they can make, they will be uncertain about the outcome. Thus we have to define the payoffs as the bidders' expected utilities.

[3]Recall: Even if Fredrik knows his own valuation he must think about what he would have bid also with other valuations. This is necessary since Fredrik must predict Mona's bid, and she doesn't know Fredrik's valuation.

Equilibrium A Bayesian equilibrium in this game is a pair of strategies, one for each bidder, such that both players maximize their expected utility, given the strategy used by that other bidder. Since it is hard to even imagine all strategies that the bidders could use, finding an equilibrium could be difficult. But luckily someone else has already proved that there exists an equilibrium in this game. And, as it turns out, once you know this equilibrium, it is relatively easy to prove that it indeed is an equilibrium.

This equilibrium requires both bidders to bid half of their valuation of the good. That is, if bidder 1 uses the strategy $b_1 = B_1(v_1) = \frac{1}{2} \cdot v_1$ then it is a best reply for bidder 2 to follow the strategy $b_2 = B_2(v_2) = \frac{1}{2} \cdot v_2$, and vice versa. Actually this is the only equilibrium of the game. But proving uniqueness (i.e. that no other equilibria exist) is more difficult, so we will leave that issue.

Why bidding half of one's valuation is an equilibrium To prove this claim, we start out (somewhat more generally) by assuming that player 1 uses some linear strategy $B_1(v_1) = z \cdot v_1$ where z could be any number, e.g. a half. Second, we note that if player 2 has valuation v_2 and bids b_2, then player 2's expected utility is

$$EU_2 = (v_2 - b_2) \cdot prob\{b_2 > b_1\} + 0 \cdot prob\{b_2 < b_1\},$$

where $v_2 - b_2$ is player 2's utility in case he wins the auction and $prob\{b_2 > b_1\}$ is his probability of winning. (For simplicity, we don't specify what the expected utility would be in case of equal bids. This, will happen with probability zero, anyway.)

The next issue is to compute the probability of winning the auction. Clearly the probability of winning is higher the higher one bids. So, what is the probability that $b_2 > b_1$? Actually we can compute this number, given that we know both player 1's bidding strategy and the probability distribution over player 1's valuations. Given that player 1 uses his linear strategy, player 2 wins the auc-

tion if $b_2 > z \cdot v_1$, or expressed differently if $v_1 < b_2/z$. That is, player 2's bid is high if player 1 happens to have a sufficiently low valuation of the good. More precisely the probability that player 1's valuation is below b_2/z is simply

$$prob\left\{v_1 < \frac{b_2}{z}\right\} = \frac{b_2}{z},$$

since v_1 is uniformly distributed on the unit interval. That is, we can rewrite 2's expected utility as[4]

$$EU_2 = (v_2 - b_2) \cdot \frac{b_2}{z}.$$

To figure out what player 2 bids we need to look at his first-order condition, i.e.

$$\frac{\partial EU_2}{\partial b_2} = (v_2 - b_2) \cdot \frac{1}{z} - \frac{b_2}{z} = 0.$$

This first-order condition describes the trade-off a bidder faces when contemplating increasing his bid slightly. The first term reflects the fact that a higher bid increases the chance of winning, in which case player 2 receives a surplus $(v_2 - b_2)$. The second term reflects the fact that higher bid (by €1) leads to higher price (by €1) in case

[4]There exists an alternative, but equivalent, way to compute the probability that $b_2 > b_1$. In particular, we may derive the probability distribution over player 1's bids, b_1. It is easy to see that if player 1 happens to value the good at 0, $v_1 = 0$, then he will also bid 0, $b_1 = 0$. If player 1 happens to value the good at 1, $v_1 = 1$, then he will bid $b_1 = z$. Thus we know what the lowest and the highest possible bids are. It is also easy to see that since all valuations between 0 and 1 are equally likely, then all bids between 0 and z must be equally likely. That is, b_1 is uniformly distributed over the interval $[0, z]$. Moreover, since the sum of all probabilities must equal one, the probability density function must be given by $f(b_1) = 1/z$. We can then compute the probability

$$prob\{b_1 < b_2\} = \int_0^{b_2} f(b_1) \cdot db_1 = \int_0^{b_2} \frac{1}{z} \cdot db_1 = \left[\frac{1}{z} \cdot b_1\right]_0^{b_2} = \frac{1}{z} \cdot b_2$$

which is the same answer as we derived above.

of winning (which happens with probability b_2/z).

The first-order condition is a single equation in one unknown, namely the bid b_2. The optimal bid is found by solving the first order condition and the solution is

$$b_2 = \frac{1}{2} \cdot v_2.$$

That is, if player 1 uses some linear strategy, it is a best reply for player 2 to bid half of his valuation.

Note that if player 1 uses a linear strategy, player 2's best reply is to also use a linear strategy. And since the two players are in a symmetric situation, we can conclude that it is then optimal also for player 1 to bid half of his valuation.

Social welfare Since both players use identical bidding strategies and since both bids are increasing in the player's valuation, the outcome of the first-price auction will be that the bidder with the highest valuation always wins the good. That is, the first-price auction is efficient since it allocates the good to the participant with the highest valuation.

The winner earns a surplus since the bid is below his willingness to pay. You can also say that the players "tell the truth" in a certain sense. Even if they give bids that are below their valuations, it is easy to infer what their true valuations are. If one of the players bid 0.4, his valuation must be 0.8. The bidders reveal their valuations.

8.6. Comparison of Auction Designs

For anyone thinking about organizing an auction to sell something it is important to understand that different designs lead to different outcomes. If it is the Government that designs the auction the primary criterion might be to find a design that leads to the efficient outcome. This is for example the case when the Swedish Govern-

ment sells radio spectrum to mobile phone networks. Then, we can conclude that the Government can choose either a second-price or a first-price auction. It doesn't matter. The spectrum will go to the firm with the highest valuation of the spectrum.[5]

Other people contemplating setting up an auction may be more interested in finding a design that maximizes their revenues. Should they then use a first-price or a second-price auction?

Revenue Equivalence With a second-price auction, the price will be equal to the second highest valuation. With a first-price auction, the price will be equal to a half of the highest valuation. The question is which is the highest? Since the seller doesn't know the potential buyers' valuations it is not clear if the second highest valuation is lower or higher than half of the highest valuation. Both prices are random variables. The seller can only compare the prices in expectations.

As it happens, one can show that the expected price will actually be the same in the two auctions. In fact, this can be shown to be true independent of how the valuations are distributed. (That is, the valuations need not be uniformly distributed over the unit interval.) This result is known as the "Revenue Equivalence Theorem." The intuition is the following.

Assume first that you are bidding in a second-price auction. Then you know that you will only win the object if you happen to have the highest valuation. You bid your own valuation, but your expected price is the expected second-highest valuation, conditional on your own valuation.

Assume second that you are bidding in a first-price auction. Then you know that you will only win the object if you have the highest

[5]But a problem is that it might actually not be socially efficient to sell the spectrum to the firm with the highest valuation. Intuitively since the monopoly profit is higher than the duopoly profit, a monopolist may have a higher willingness to pay for some new spectrum, thereby remaining a monopolist, than what a potential entrant is willing to pay for the new spectrum.

valuation. To ensure that the person with the second highest valuation does not have an incentive to bid over you bid, you should give a bid equal to the second highest valuation (plus a tiny extra amount). Bud since you don't know the second highest valuation exactly, you bid the expected second-highest valuation, conditional on your own valuation being the highest value.[6] That is, also in a first-price auction, the winner's expected price is the expected second-highest valuation, conditional on his own valuation.

Actually, it can also be shown that both the English and the Dutch auction will give rise to the same expected price.

One may also think more broadly, than just in terms of auctions. Another possibility would be for the seller to bargain with all the potential buyers. Or there may be other, more exotic ways to organize the market. But - and this is really surprising - one may actually prove that there doesn't exist any other way to organize the market which gives the seller a higher expected price for the good.[7]

[6]When valuations are uniformly distributed over the unit interval, the expected second-highest valuation is half of the highest valuation. That is if you know that your own valuation v is the highest, then you know that the second highest valuation is uniformly distributed on the interval [0, v]. The expected value is the mid point, i.e. v/2.

[7]This result follows from a theorem called the "Revelation Principle".

9. Quality Uncertainty and the Market Mechanism

9.1. Why study quality uncertainty?

In most labor markets, employers will not be certain about the productive capabilities of the individuals applying for jobs. And, when you consider buying a used car, you will be aware that there is a risk that you will be buying a lemon. (Lemon is american slang for a car found to be deficient after it was bought.) Banks cannot easily assess the quality of all the projects proposed by various entrepreneurs asking for credit. (Credit derives from a Latin word which means he/she/it believes.) Insurance companies do not know the risk that a certain individual will fall ill.

It should come as no surprise that when there is uncertainty about the quality of the goods to be traded, the market mechanism does not work as efficiently as otherwise. In fact, the market mechanism may break down completely as a result of uncertainty. What is surprising, however, is that it is not uncertainty *per se* that creates the worst problems. Closer analysis reveals that the worst problems arise as a result of *asymmetric information*. While the buyers are unable to observe the quality of the goods for sale, the sellers are usually better informed.

The people applying for jobs probably know quite well themselves how productive they are. Or, at least they know more than the potential employer. The seller of a used car or a house often

knows exactly what the problems might be. The entrepreneur is typically much better informed than the bank both about the market he intends to serve and the technology he will employ. And people seeking health insurance probably know more about their own health conditions than the insurance companies. At least the individual knows how much he exercises and smokes, how good his sleep is and how well he copes with the everyday stress.

Symmetric information Clearly if both the buyer and the seller know the quality of the good for sale, they will trade if and only if the buyer values the good higher than the seller. They will agree on a price somewhere in-between their valuations and share the surplus from trading. Both will be happy.

Also if both parties would be equally *un*informed about the quality of (say) a used car, the market works well. If you inherit your uncle's car you may be equally uninformed about the car's expected remaining lifetime as the potential buyer. You may both understand that the car will last for another one or two years. You may both understand that both outcomes are equally likely. And if the buyer values such a risky car higher than you, you will sell it to him. You may value the car at €1000 if it lasts one year and €2000 if it lasts two. The buyer may value the car at €4000 if it lasts one year and at €8000 if it lasts two. To you the expected value of the car is €1500 and for the buyer it is €6000. You may then agree on the price €3750, which is the average of your expected valuations, and you will both be happy. In this particular case nobody would even regret their decision, independent of what the true quality turns out to be.

In other cases, the buyer will regret his decision if the car turns out to be a lemon, while the seller would regret the decision if the car turns out to be a "peach." Say that the buyers valuation is only €2000 if the car lasts only one year and €4000 if it lasts two years. Then the natural trading price would be €2250, in which case the

buyer indeed will regret his decision in case the car turns out to be a lemon. But, the point is, that even if the buyer is aware of this possibility in advance, he will agree to trade. And clearly trading the car is efficient. Independent of the quality of the car, the buyer values it more than the seller.

The market mechanism thus works well even when buyers and sellers lack information about quality, as long as they have access to the same information.

Asymmetric information & adverse selection If the current owner of the car knows the quality, the situation is radically different. Then the current owner will try to take advantage of his superior information, only selling the car if the quality is low enough. Even if the pool of cars contains mainly cars of good quality, the ones for sale are the ones with hidden faults. That is:

Being for sale is in itself bad news about quality!

For a prospective buyer to make a good decision, it is important to base the decision on this insight that the cars for sale is a consciously selected subset of the stock of cars. It is the cars with inferior quality that have been selected. This phenomenon is called *adverse selection* and it arises as a result of asymmetric information.

The immediate consequence of adverse selection is that the prospective buyer cannot use information about how common different qualities are in the pool of all cars. The buyer must base his decision on how common different qualities are among the cars for sale. Expressed differently, the buyer must understand that the seller's decision to sell his car contains information about its quality.

Let's say that there are two types of cars, lemons and peaches, and that they are equally common. Let's say that the sellers value lemons at €0 and peaches at €2000, while buyers value them at €0 and €3000. Every seller knows the quality of his car, but the buyers cannot tell the difference between lemons and peaches. Since the

buyers cannot see any difference, the price of the two types of cars must be the same. (Expressed differently, even if there would be one market place for lemons and one for peaches, all sellers would bring their car to the peach market.)

Clearly it would be efficient if all cars are sold. And if both sides would be uninformed about quality, a price of €1250 would lead to exactly that outcome, since the buyers' expected valuation is €1500 and the sellers' expected valuation is only €1000.

But when the sellers know the quality of their own car, the market breaks down. If the market price would be €1250 only the people with lemons would offer their cars for sale. That is, even if there are equally many lemons and peaches in the pool of all cars, there are only lemons among the cars for sale. While the buyers' expected value of a car in the pool is €1500, the expected value is €0 for the cars for sale. Thus, no prospective buyer would actually accept to trade at the suggested price. In fact, any price below €2000 would imply that sellers only offer lemons and, consequently, that buyers would not accept their offers. Any price above €2000 would imply that all cars will be put out for sale. But, since the buyers expected valuation then would be €1500, there would not be any trade. That is, there does not exist any price that would make the buyers and the sellers to trade, despite the fact that trading would have been efficient. The market mechanism may thus break down as a result of adverse selection. In this case, the would not even exist any market for used cars.

Asymmetric information and adverse selection probably affect most markets, even if the consequences do not need to be as drastic in as in the the example above. The purpose of the rest of this chapter is to analyze also milder forms of adverse selection to better understand the main determinants of how sever the problems are in different markets. To do so, we also need to consider a somewhat more complicated formal model. As an extra benefit, the study of the more complicated model will make it easier for you

to understand research involving adverse selection in the future.

9.2. Bayes's Rule

Most situations are characterized by asymmetric information. Firms may not know each other's costs. And they may not know individual consumers' willingness to pay. Consumer may not know the quality of goods for sale. Municipalities may not know the quality of the services offered by private entrepreneurs. Employers may not know the productivity of job applicants. Banks may not know the bankruptcy risk of entrepreneurs. Insurance company may not know risk that a person falls ill.

Often, however, people disclose some of their private information when they act. Expressed differently, people may learn about other peoples' private information by observing their behavior. A person willing so sell his car at a certain price reveals that his valuation of the car must be below that price. Prospective buyers may therefore learn something about the quality of the car by observing the seller's decision. Similarly, a firm may learn about a competitor's cost by observing the competitor's price. Employers may learn something about an applicant's productivity by observing how many years the applicant has spent in school. A bank may learn something about an entrepreneur's bankruptcy risk by observing how much of his own money the entrepreneur is willing to risk.

This chapter describes what people should infer from observing other peoples' behavior in more detail. That is, this chapter is about learning, which is neatly summarized by the so-called Bayes's rule. Bayes's rule describes how we should update our beliefs when we receive new information.

9.2.1. Example 1: Banks and entrepreneurs

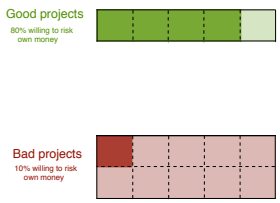
Consider a bank offering loans to small entrepreneurial firms. All the entrepreneurs who come to the bank have some, but not enough, money to finance their projects. Even if they do not have full information, the entrepreneurs know relatively well if their project will succeed or fail. The bank, in contrast, doesn't know if a particular firm will succeed. If the project succeeds the entrepreneur will be able to pay the loan. But if the project fails, the firm will go bankrupt and the band will lose its money.

A key question for the bank is thus if there is a way for it to learn about the entrepreneurs' private information? Obviously, it cannot ask the entrepreneur as any entrepreneur independent of the quality of his project would have an incentive to exaggerate his likelihood of success. What the bank can do, however, is to ask if the entrepreneur is willing to risk his own money in the project. If the entrepreneur believes the project will succeed, he is willing to risk his own money. Otherwise not.

Let's continue with a numeric example. Let's say that there are two types of entrepreneurs and that only a third of the projects are good. To fix ideas, lets say that there are five entrepreneurs with good projects and ten with bad projects as described in Figure 9.1. If the bank could not assess the quality of the projects it would loose a lot of money. However, among the entrepreneurs with good projects, 80 percent believe the project is good and are willing to risk their own wealth. Among entrepreneurs with bad projects only 10 percent believe that the project is good and are willing to risk their own wealth. Before reading any further, please solve the following exercise.

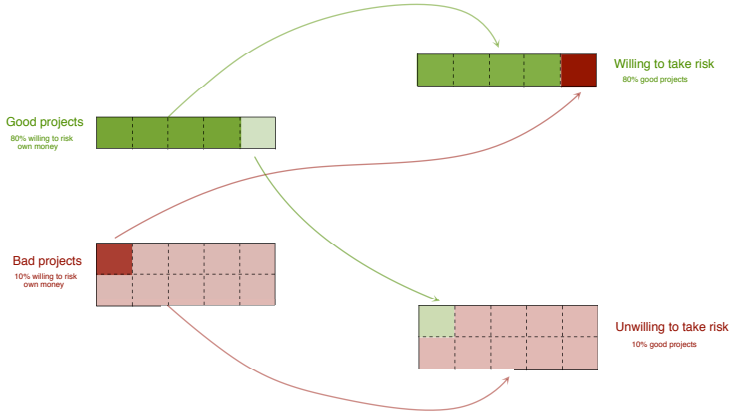
Exercise 41. What is the probability that a random entrepreneur has good project in the population? What is the probability that a random entrepreneur has good project among those with some own funding? Among those without own funding?

Figure 9.1.: Projects of different qualities



Answer: In the population as a whole a third, i.e. 33 percent, of all projects are good. Among the entrepreneurs with good projects four out of five are willing to risk their own money and among the entrepreneurs with bad projects one out of ten are willing to do so. Thus, there are in total five entrepreneurs who are willing to risk their own money. And, among those four have good projects. This means that 80 percent of those who are willing to take the risk have good projects. Among those who are not willing to take the risk, only ten percent have a good project, as revealed by the Figure.

Figure 9.2.: Projects of different qualities



◀

Thus, by observing the entrepreneurs' behavior (how much of their own money they are willing to risk) a bank may learn something about their private information (probability of success). In this particular example, the bank can reduce the risk of bankruptcy from 67 percent to 20 percent.

9.2.2. Example 2: Job market

An employer in a small town advertises to hire more people to an attractive job. A major problem for the employer is that he cannot observe the productivity of individual job applicants. But from experience the employer knows that half of the population have high productivity and half have low. Let us say that there are in total 1000 people living in the town and that they all apply for the job. There are thus 500 applicants with high productivity and 500 with low productivity, but the employer doesn't know who is who.

However, the employer also knows from experience that among people with high productivity 90 percent invest in a master, while only 10 percent of people with low productivity invest in a master.

Exercise 42. What is the probability that a job applicant with a master has high productivity? What is the probability that a job applicant without a master has high productivity?

Answer: The number of high-productive who invest in a master is 450 (= 0.9 * 500). The number of low-productive who invest in master is 50 (= 0.1*500). Thus the total number of people with a master is 500 (= 450 + 50). The share of people with a master who are high-productive is 0.9 (= 450/500). Similarly, the share of people without master who are high-productive is 0.10 (= 50/500).

◀

Thus, in this example education contains information about productivity. While the share of high productive in the population as

a whole is 50 percent, the share of high productive among people with a master is 90 percent. The share of high productive among people without master is only 10 percent.

9.2.3. The general rule

It is now time to establish a general rule for how the employer should reason. Recall that the employer is assumed to know two things:

1. The employer knows the population shares of different productivity levels in the population. We denote the share of high productive people in the population by $P(H)$ and the share of low productive people in the population by $P(L)$. Obviously $P(H) + P(L) = 1$.

2. The employer also knows something about peoples' behavior. We denote the share of people with high productivity who invest in a master by $P(M \mid H)$ and the share of people with low productivity who invest in a master by $P(M \mid L)$.

The question is how the employer could combine these to pieces of information. In particular, what does a masters' degree say about productivity?

Exercise 43. Compute what share of people with a master degree that are highly productive, i.e. $P(H \mid M)$.

Answer: Start out by computing the share of people who are both highly productive and have a master. This is given by the share of people with high productivity multiplied by the share of people with high productivity who get a master, i.e. $P(H\&M) = P(H) \cdot P(M \mid H)$. Then compute the share of people who have a master. This is given by the sum of high productive people with a master and low productive people with a master, i.e. $P(M) =$

$P(L) \cdot P(M \mid L) + P(H) \cdot P(M \mid H)$. Finally, the share of people with a master degree that are highly productive is simply

$$P(H \mid M) = \frac{P(H\&M)}{P(M)} = \frac{P(H) \cdot P(M \mid H)}{P(L) \cdot P(M \mid L) + P(H) \cdot P(M \mid H)}.$$

In the numerical example above, this share is given by

$$P(H \mid M) = \frac{\frac{1}{2} \cdot \frac{9}{10}}{\frac{1}{2} \cdot \frac{1}{10} + \frac{1}{2} \cdot \frac{9}{10}} = \frac{9}{1+9}.$$

◀

The main point here is that people can learn something about other peoples' private information by observing their behavior. How much can one learn?

Exercise 44. What happens if high and low productive people behave the same, i.e. $P(M \mid L) = P(M \mid H)$? What happens if they behave very differently, e.g. $P(M \mid L) = 0$ and $P(M \mid H) = 1$?

Answer: If $P(M \mid L) = P(M \mid H)$, then

$$P(H \mid M) = \frac{P(H)}{P(L) + P(H)} = P(H).$$

That is, the employer does not learn anything. If $P(M \mid L) = 0$ and $P(M \mid H) = 1$

$$P(H \mid M) = \frac{P(H) \cdot 1}{P(L) \cdot 0 + P(H) \cdot 1} = 1$$

and

$$P(H \mid L) = \frac{P(L) \cdot 0}{P(L) \cdot 0 + P(H) \cdot 1} = 0.$$

That is, the employer will know a persons productivity with certainty by observing the person's behavior.◀

To sum up: Education can be a *signal* of productivity. If people

with different productivity invest in a certain degree with different probability, then observing that a person has this degree carries information about the person's productivity. Employers who cannot observe a peoples' productivity directly can, instead, base hiring decisions and wage offers on an applicant's level of education. To do this, the employer should use Bayes's rule.

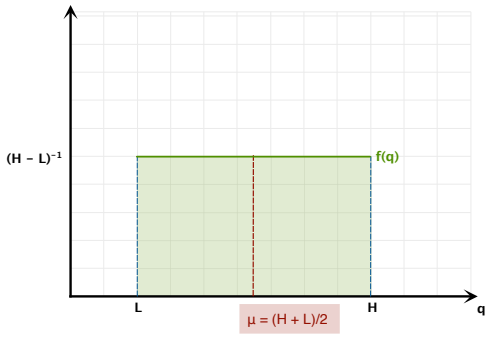
9.3. Market for Lemons

The basic point of the present chapter is that asymmetric information about quality may disrupt a market as a result of adverse selection. The intuition is straightforward. Assume that buyers don't observe the quality of (say) used cars. If the market price would be €1000, then the buyers would conclude that only cars with a quality below €1000 will be supplied. Thus, the average value of cars actually supplied is low, say €500. Then, the buyers would only be willing to pay €500. But if that price would prevail in the market, the buyers would only be prepared to pay an even lower price. But, luckily, if buyers and sellers have sufficiently different valuations of quality, the information problem may be at least partly overcome.

To show that asymmetric information about quality can disrupt a market, we will consider a model that only differs from the model of perfect competition in this one respect, that buyers cannot observe the quality of goods. We will thus assume that there are infinitely many (potential) sellers as well as infinitely many (potential) buyers. We will also assume that the price is set to clear the market, i.e. to equalize supply and demand, without going into the details of who it is that actually sets this price. Notice that there can only be one price for used cars since all the cars look the same to the buyers.

Sellers Consider a local market for used cares. There is a continuum (mass 1) of potential sellers with one car each. The quality of these cares is uniformly distributed over $[L, H]$. This means that the lowest quality is L and that the highest quality is $H > L$ and that all qualities in-between are equally common. This distribution is described in Figure 8.1.[1] It is easy to see that the average of such

Figure 9.3.: Uniform distribution



a uniform distribution must simply be the midpoint between the two endpoints, that is $\mu = (H + L)/2$, which is represented by the red vertical line.[2]

A current owner's utility depends on whether he sells the car or not. A current owner who keeps his car gets utility $\Theta_S \cdot q + m$ and utility $m + p$ if the car is sold. All current owners know the quality of their own car, which means that an owner will sell if and only if $\Theta_S \cdot q \leq p$. That is a current owner will sell if his car is of sufficiently low quality $q \leq p/\Theta_S$.

Buyers There is also a continuum (mass 1) of buyers and they are all identical. A buyer's utility depends on whether he owns a car

[1] The green line represents the density function which is constant and given by $f(q) = (H - L)^{-1}$ for all qualities $q \in [L, H]$.

[2] The average quality is given by $\mu = \int_L^H q \cdot f(q) \cdot dq$ that is $\mu = (H - L)^{-1} \cdot \int_L^H q \cdot dq$ or $\mu = (H - L)^{-1} \cdot \left[\frac{1}{2} \cdot q^2\right]_L^H$ and hence $\mu = (H - L)^{-1} \cdot \left(\frac{1}{2} \cdot H^2 - \frac{1}{2} \cdot L^2\right) = (H + L)/2$.

or not. Without a car, the utility is m which is simply the income. With a car, the utility is $m+\Theta_B \cdot q-p$ where q is the quality of the car, p is the price paid for the car and Θ_B is the buyer's valuation of quality. The buyers do not care about the risk associated with buying a second-hand car.

We assume that the buyers know the average average quality of cars for sale. We denote the average quality of cars for sale by μ^*. The reason may be that the buyers are very sophisticated, that they know the distribution of quality in the pool and that they use Bayes's rule to take into account the sellers' behavior. Alternatively, the reason may simply be that they have observed the quality of many second hand cars that have been traded in the past.

All buyers prefer to buy a second-hand car if the expected utility of owning a car, i.e. $E\left\{m+\Theta_B \cdot q-p\right\}=m+\Theta_B \cdot \mu^*-p$, is greater than the utility of not owning a car, i.e. m. Expressed differently, the buyers prefer to trade if $\Theta_B \cdot \mu^*>p$. Then, demand is 1, since the mass of potential buyers is 1. They all prefer not to trade if the inequality is reversed. Then, demand is 0. In case the expected utility of owning is exactly equal to the utility of not owning, then all the buyers are indifferent. In this case, demand can be said to be any number between 0 and 1. Thus,

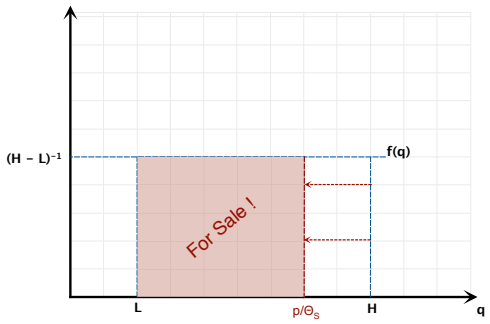
$$D=\begin{cases} 1 & p<\Theta_B \cdot \mu^* \\ [0,1] & p=\Theta_B \cdot \mu^* \\ 0 & p>\Theta_B \cdot \mu^* \end{cases}$$

Thus, demand is falling in price as usual.

Efficiency We assume that buyers' valuation of a used car of a certain quality is higher than the sellers' valuation, i.e. $\Theta_B>\Theta_S$. Thus, all cars should be sold from an efficiency point of view.

Adverse selection There is, however, adverse selection. A current

Figure 9.4.: Adverse selection



owner will only sell if his car is of sufficiently low quality $q \leq p/\Theta_S$. That is the best cars will be kept and the worst cars will be put out for sale. The lower the price of used cars would be, the fewer cars would be for sale.

The advantage of assuming that the distribution of cars is uniform is that also the distribution of used cars is uniform. In particular, the lowest quality for sale is L and the highest quality for sale is p/Θ_S and all the qualities in-between are equally common. This means that we can compute the average quality of cars for sale without explicit use of Bayes's rule.

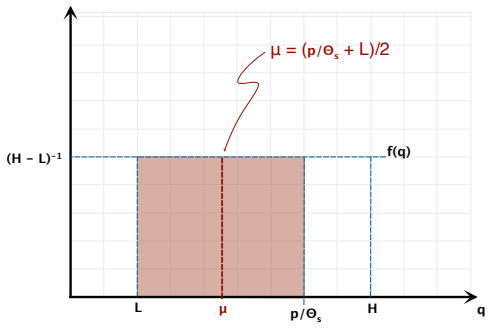
The average quality of the cars for sale is simply the midpoint between the highest quality for sale and the lowest quality for sale, i.e.

$$\mu^* = \frac{L + \frac{p}{\Theta_S}}{2}. \tag{9.1}$$

The average quality is illustrated in Figure 9.5. And, recall that we assume that the buyers know this quality.

Equilibrium An equilibrium is a price that clears the market and the quantity of cars traded at that price. But, since there is an easy one-to-one relationship between the quantity traded and the average quality, it turns out to be more convenient to focus on the

Figure 9.5.: Average quality of cars for sale



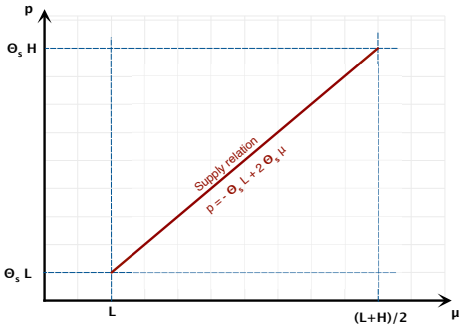
price and the average quality for sale. For example, if all cars are sold, the average quality is $(H + L)/2$. If fewer cars are sold, the average quality will be lower. In the extreme, if only the cars of the lowest quality is sold, then the average quality is L.

To describe the supply of cars we use the fact that the average quality for sale is given by equation 9.1 which can be rewritten as

$$p = -\Theta_S \cdot L + 2 \cdot \Theta_S \cdot \mu^*.$$

This equation is described in Figure 9.6. It shows that the higher

Figure 9.6.: Supply



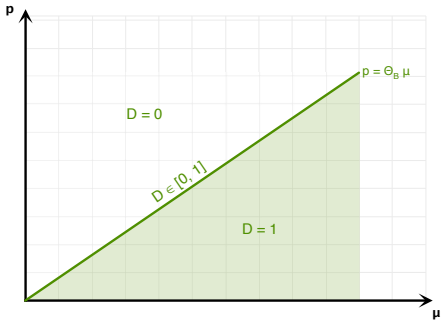
the price is, the higher the average quality of cars for sale will be.

The demand for cars is described by the price which makes the buyers indifferent between buying and not, i.e.

$$p = \Theta_B \cdot \mu^*.$$

The demand relationship is described by the green line in Figure 9.7. If the price is on the line, the price is equal to the value of the

Figure 9.7.: Demand



expected quality. If the price is above the line, all consumers prefer not to buy (white area). If it is lower, all consumers prefer to buy (green area).

Consider now a combination of average quality and price (μ^*, p) on the supply relationship, but below the green line. An example is the blue dot in Figure 9.8. Clearly at such a point, there is excess demand. All the potential buyers wish to trade, so demand is equal to 1. But since the average quality is below the maximum, not all sellers wish to trade, so supply is below 1. In such a case the price must be increased to equate supply and demand.

Consider next a combination of average quality and price (μ^*, p) on the supply relationship, but above the green line. An example is the blue dot in Figure 9.9. Clearly at such a point, there is excess supply. None of the potential buyers wishes to trade, so demand is equal to 0. But since the average quality is above the maximum,

Figure 9.8.: Excess demand

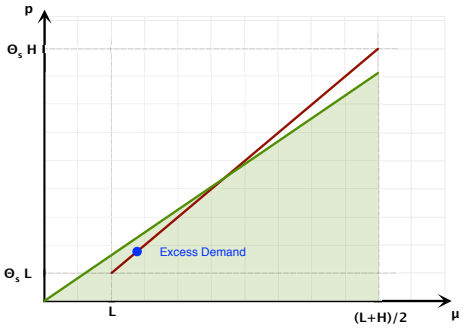
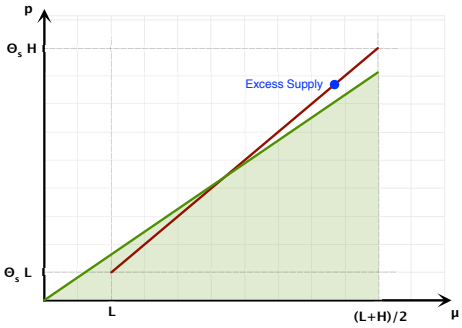


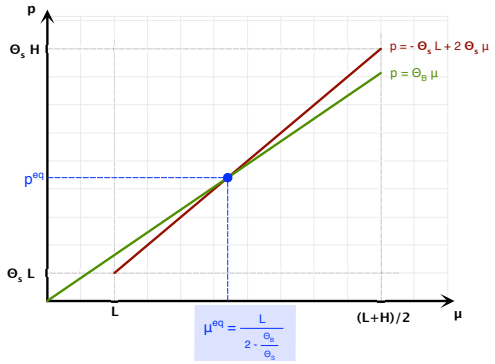
Figure 9.9.: Excess supply



some sellers wish to trade, so supply is above 0. In such a case the price must be lowered to equate supply and demand.

Consider finally the combination of average quality and price (μ^{eq}, p^{eq}) at the intersection of the supply and demand relations, as described by the blue dot in Figure 9.10. At this point, supply equals demand. Some of the potential sellers wish to trade and since all the buyers are indifferent between trading and not, all agents can realize their plans at the same time. (This presupposes that the buyers somehow coordinate, selecting the ones who should trade. But we simply assume that they succeed to arrange this somehow. For example, it may be that those who happen come

Figure 9.10.: Equilibrium



first are served.)

The main conclusion is that some, but not all the used cars will be sold. In particular, the average quality of cars sold in equilibrium is given by

$$\mu^{eq} = \frac{L}{2 - \frac{\Theta_B}{\Theta_S}}. \tag{9.2}$$

This outcome is inefficient since some current owners keep their cars despite the fact that there are potential buyers who value them more.

Exercise 45. Show that the average quality of cars sold in equilibrium is indeed given by equation 9.2.

Hint: Use the two equations describing the supply and demand relations above. Together they can be solved for the equilibrium price and the equilibrium average quality. ◀

As Figure 9.10 is drawn, the supply and demand relations do intersect. But in order for this to happen it is necessary that the buyers' valuations of quality is not too much larger than than the sellers' valuation of quality. More precisely, it is required that $\Theta_B < 2 \cdot \Theta_S$. We will simply assume that this inequality is satisfied. (In contrast, if $\Theta_B > 2 \cdot \Theta_S$, the supply relation would be below the demand relation everywhere. In that case all the cars would be

sold. We will not study this possibility in more detail, however.)

A simple indicator of the efficiency of the market is to compute the proportion of cars sold in equilibrium. This proportion is given by $e = \frac{h-L}{H-L}$, where h is the highest quality sold in equilibrium. After a few steps of algebra, market efficiency can be written[3]

$$e = \frac{\Theta_B - \Theta_S}{2 \cdot \Theta_S - \Theta_B} \cdot \frac{2 \cdot L}{H - L}.$$

Notice that if the buyers and the sellers would have almost the same valuation of quality, i.e. $\Theta_B \approx \Theta_S$, then almost no cars would be sold in equilibrium. Likewise, if the lowest quality is very low, i.e. $L = 0$, then also no cars would be sold. The latter case is illustrated in Figure 9.11. In contrast, the market mechanism is

Figure 9.11.: Market disappears

relatively efficient if the difference in quality between different cars is relatively small. In that case, the sellers information advantage is smaller. The market is also relatively efficient if the buyers' valuation of quality is much higher than the sellers' valuation.

[3]The highest quality sold can be found by using the supply relation. It is $h = p^{eq}/\Theta_S$. The equilibrium price is given by $p^{eq} = \Theta_B \cdot L / \left(2 - \frac{\Theta_B}{\Theta_S}\right)$. Thus $h = \frac{\Theta_B}{\Theta_S} \cdot \frac{L}{2 - \frac{\Theta_B}{\Theta_S}}$.

Incomplete but symmetric information If also the sellers would be uninformed about the quality of their cars, they would all be willing to sell if $\Theta_S \cdot \mu \leq p$. The buyers would then all be willing to buy if $\Theta_B \cdot \mu \geq p$. Then, if $\Theta_B \geq \Theta_S$ there exists an equilibrium where all cars are sold, at uniform price, e.g. $p = \mu \cdot (\Theta_B + \Theta_S)/2$.

9.4. Applications

Insurance markets People who know that they have a high risk of falling ill have a strong incentive to buy a health insurance. Insurance companies must then charge high fees to cover their costs. As a result of the high fees, low-risk individuals will not buy insurance. But when the low-risk people leave the market, the insurance companies must charge even higher prices from those who still buy insurance, which force even more people with relatively low risk of fall ill out of the market. A possible solution to this problem is mandatory insurance, i.e. that the government requires all individuals to be insured. Another solution is that large groups of people buy insurance together. For example if all people who belong to a labor union can buy insurance, the insurance company can be certain that only a limited amount of adverse selection will reduce its profits.

Labor markets If both people with low and high productivity apply for new jobs, the employers would have to set relatively low wages. But, then, high-productivity workers would prefer to stay at their old jobs. A possible solution to this problem is that large firms can have internal labor markets and thereby use the information they have about their employees. Another solution is that high productive people signal their productivity by investing in education. Education is less costly for people with high productivity. Therefore they may be willing to invest in education to get a wage

premium. The same wage premium is not sufficient to motivate people with low productivity to invest in the same education, since their cost of getting through school is higher.

Part III.

Limits to Price Competition

10. Sources of market power

A study of the car market in Europe reveals that car producers earn substantial markups, varying both across brands and across countries. Some representative Lerner indexes are presented in Table .

Table 10.1.: Lerner index in the car industry

Model	Belgium	France	Germany	Italy	U.K.
Fiat Uno	7.6	8.7	9.8	21.7	8.7
Ford Escort	8.5	9.5	8.9	8.9	11.5
Peugeot	9.9	13.4	10.2	9.9	11.6
Mercedes	14.3	14.4	17.2	15.6	12.3

Exercise 46. What does the Lerner index measure? How is it defined?

The table thus shows that the car manufacturers charge prices exceeding their marginal costs, despite the competition that exists between different brands. Comparing different brands, especially Mercedes vs. all the others, suggest that luxury cars command particularly high markups. Comparing different countries suggest that car producers engage in third degree price discrimination. For example, cars appear more expensive in their home countries.

This is not what we should observe according to the Bertrand model. The Bertrand model suggests that oligopoly prices are competed down all the way to marginal cost, even with only two firms

in the market. Clearly the Bertrand model must miss something important about oligopolistic competition.

The reason for why firms in many markets are not forced to set price equal to (marginal) cost is that, unlike the firms in the Bertrand model, they would not loose all their customers, even if a competitor would charge a (slightly) lower price. There are many reasons for why a high-price firm may keep some of its customers.

If products are *differentiated*, and if some customers prefer one firm's product to the other, they will continue to buy that firm's product even at a (slightly) higher price. As a consequence, a firm that attempts to poach its rivals' customers must reduce price quite a lot. Such a reduction in price would be less profitable. Hence, when products are differentiated, firms have less incentive to lower their prices.

If firms have *capacity constraints*, a single firm will not be able to sell to all customers. If a firm cannot supply more customers, it will not have an incentive to lower price to win other the competitors' customers over. If firms have different costs, a low cost firm will be able to charge a price above its own cost without losing all customers. As long as the price is below the competitors' costs, they will not have an incentive to undercut.

If *customers are not informed* of all firms' prices, they may continue to buy from their regular firm even if other firms would lower their prices. In such a case, a firm would not be able to attract many new customers by lowering its price. Hence, the incentive to lower price is reduced.

If it is *costly for consumers to switch* suppliers, firms will be able to charge higher prices without losing customers. An example of a switching cost is that consumers may have to learn how to use the competitor's product.

Firms may also have the ability to form a *cartel*, and thereby to set price above cost. That is, the individual firm does not loose its customers by increasing its price, if all firms in the market increase

their prices in a coordinated fashion, e.g. by the same amount and at the same time.

This list is not complete. There are many factors that influence how much market power a firm has, in addition to the number of firms. Additional dimensions are how price sensitive customers are, and how easy entry is to the market. Yet another important factor for what prices firms charge is how frequently and when firms revise their prices, and exactly what information they have about their competitors and the customers. In this part we will look closer at some of these sources of market power. We will start with capacity constraints.

11. Product differentiation

Olive oil Many food stores stock a dozen or more different olive oils. And apparently people continue buying also those that are several times more expensive than the cheapest available ones. The leading daily newspaper in Gothenburg, GP, asked a panel of nine people to blind-test and compare eight different oils with prices ranging between €5 and €26 per liter. The group, which included both a cook and non-professionals, focused mainly on taste but also scent, texture and color was given some interest. The panel discovered first and foremost that olive oils are different. Some oils are fruity, others are nutty, yet others reminded the panelists of pepper. Some oils are neutral, others have a more pronounced taste. The panel also found some oils oilier than other oils.

All the panelists individually marked all the oils, using the newspaper's lighthouse-scale, with one lighthouse meaning "fail" and five lighthouses meaning "brilliant." There turned out to be some agreement. Most of the panelists disliked one of the oils, which was considered too bitter. That oil only received one lighthouse. On the other extreme, Florens Grappolini extra vergine, must have pleased most of the panelists as it received four lighthouses in the overall assessment. But there was also some disagreement. One of the oils received all possible marks (1-5) from different members of the panel. The disagreement appears to some extent be caused by differences in preferences, but also differences in intended use (together with cold dishes or with salads) may have played a role.

As one would expect, the most expensive oil was also the oil most people liked the most. But, apart from Flores Grappolino, the

relation between price and the number of lighthouses was rather weak. What could possibly explain that?

What does the olive oil example teach us? The olive oil market illustrates several related points that are actually common to most consumer markets. Consumer goods firms produce different versions rather than exactly the same good. This way, there is more variety available to consumers. Every consumer can choose the version closest to his own preferences, or different versions depending on what the product is used for. Product differentiation clearly contributes to consumer welfare.

But product differentiation also reduces price competition. The huge variation in price among different olive oils is clear evidence. With homogenous goods, any firm charging a price slightly below the competitors will capture the whole market. As a result, all prices are competed down to marginal cost, which must be the same for all surviving firms. With differentiated goods, some consumers will not switch to the firm with the lowest price, since they like the other firms' variants better and are willing to pay a premium for those variants. To put it in technical terms: the cross-elasticity of demand is not infinite at equal prices. Consequently firms will be less eager to reduce their prices than if their products were perfect substitutes.

The primary purpose of this lecture is to study how prices are determined in a market with differentiated goods. In particular, to what extent will product differentiation create market power? Recall that market power can be defined as a firm's ability to set price above marginal cost without loosing all its customers. We will consider both the case when people agree that one product is better than the other, so-called vertical differentiation, as exemplified by the Flores Grappolino olive oil, as well as the case when people disagree in their rankings, so-called horizontal differentiation.

In a later chapter, we will also touch upon a second issue, namely

product choice. When firms design their products, they will have to consider both roles of the design. The design should appeal to the consumers' tastes and make them willing to pay a high price for the good. But the firms also want to differentiate their products from their competitors' products to avoid price competition.

Goods as bundles of characteristics A computer may be defined in terms of the size of the hard disc, the memory, the speed of the processor, the number of USB ports, its portability and so on. Laptops score high on portability but are typically less powerful than desktops. But the exact specification also depends on the brand. Similarly cars may be defined in terms of their number of horsepowers and fuel efficiency as well as differences in design. Restaurants differ in terms of the quality of the food they prepare, the length of their wine lists, the level of service and ambiance.

It is often convenient to think of goods as bundles of characteristics and to think of consumers as having preferences over characteristics, rather than over the goods per se. Commuters may need their computers to be highly portable while graphic designers may consider power more important. An old couple may be satisfied with a slow, stationary computer as long as it is reliable.

But, of course, knowing the consumers' preferences over characteristics, one may then derive their preferences for different goods bundling characteristics in different proportions. In other words, we may derive demand for goods based on their characteristics.

There are several advantages with this so-called characteristics approach. A first advantage is that this approach provides much more intuition for how product differentiation affects pricing. A second advantage is that by describing the space of all possible products, it is also possible to analyze how firms choose to position their products within this space. That is, we can start discussing how companies like Lenovo and Apple think when deciding how to equip their computers? A third advantage of the characteristics

approach arises when estimating demand functions. In order to estimate demand for ten differentiated products an econometrician has to estimate ten different own-price elasticities and 90 cross-price elasticities. But if demand is instead derived from demand for only a few different characteristics, the econometricians may economize on the number of parameters to estimate, and thereby reduce the need for data.

11.1. Horizontal Differentiation

Some people like fruity olive oils better than nutty olive oils while others prefer the nutty ones. People also have different preferences for different convenience stores, depending on how close they are to home. Some people rather drink Coke than Pepsi while others prefer Pepsi to Coke. When goods or services differ and people rank the available alternatives differently, we talk about horizontal differentiation.

This section presents the so-called Hoteling model, which is one of the most commonly used workhorse models of oligopolistic interaction when goods are horizontally differentiated.

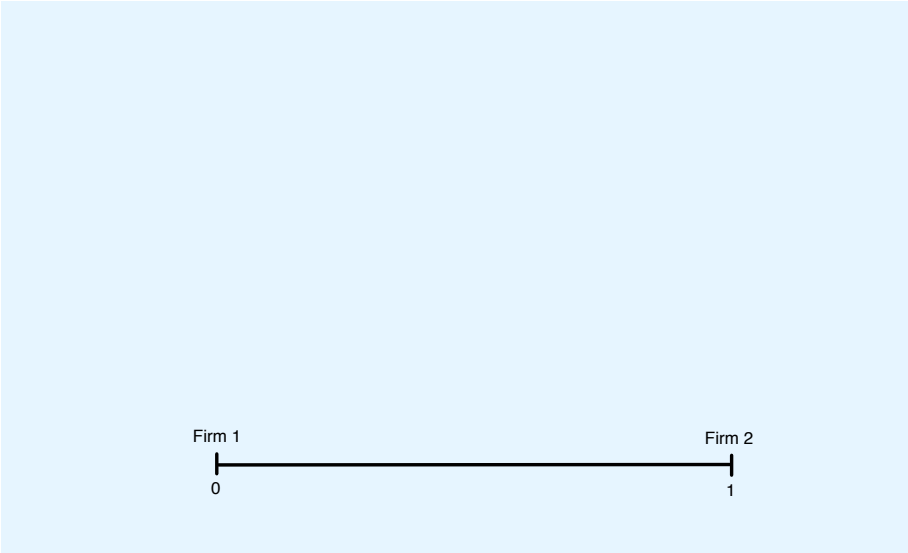
11.1.1. The Hoteling model

The product room Consider a one-kilometer long street with people living and firms locating their convenience stores at various points on this street. Even if the stores sell exactly the same products, they will still be different in the eyes of the consumers, as traveling times will differ. Different consumers will have different preferences for the convenience stores depending on where they live and where the store is located.

Firms There are two firms with one store each. One store is located at one end of the street (at point 0); the other is located at

the other end (at point 1). The stores sell exactly the same good

Figure 11.1.

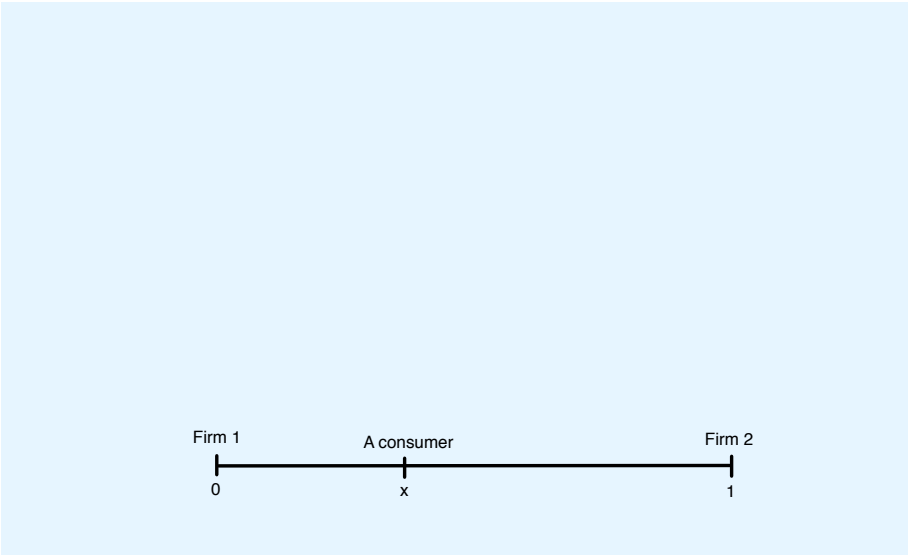


and they both have the same constant marginal cost, c, of supplying their goods.

Consumers All consumers live on the same street, at various locations between the two stores. Going to a store is costly and the cost is proportional to the distance traveled. Going from one end of the street to the other costs €t and going from the middle of the street to visit one of the stores at either end costs €$t/2$. A consumer who lives at point x is at the distance x from store 1 and at the distance 1–x from store 2. It would thus cost this consumer $t \cdot x$ to visit store 1 and $t \cdot (1 - x)$ to visit store 2.

There are very many consumers living on the street and we assume that they are evenly spread along the full length of the street. That is, the number of people living on a segment of the street is proportional to the length of the segment. The total population is m. The number of people living to the left of point x is $x \cdot m$ and the number of people living to the right of point x is $(1 - x) \cdot m$.

Figure 11.2.



For example if the total population is 2000 then 600 people live to the left and 1400 people live to the right of point $x = 0.3$.[1]

Timing First the two firms simultaneously decide on their prices and make them public and then the consumers decide whether they want to buy a unit of the good and from which firm to buy it.

Working backwards, we start out solving the second stage of the game, i.e. with deriving demand. How much will each firm be able to sell for every possible combination of prices that they might set in the first stage?

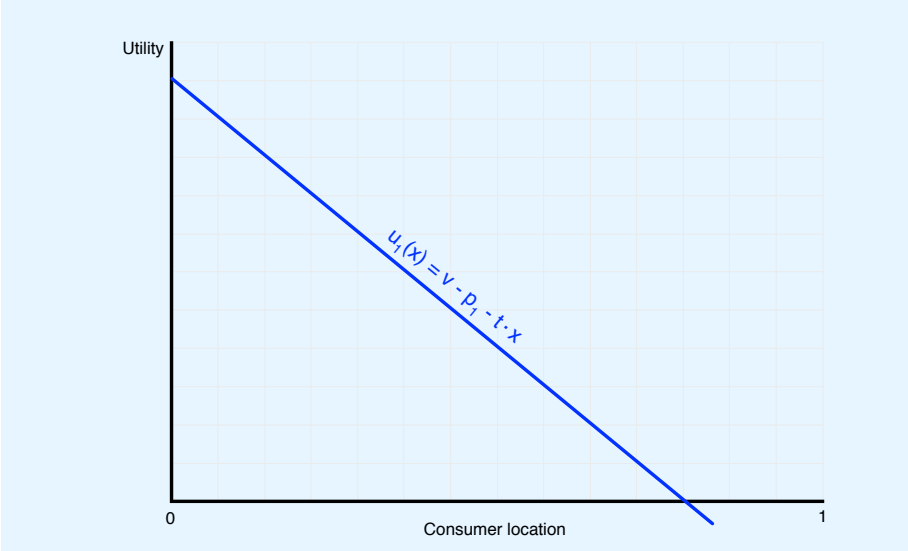
11.1.2. Deriving the firms' residual demand

All consumers have unit demand and are willing to pay v for one unit of the good. In particular, the price plus the travel cost must

[1] More formally we assume that there are infinitely many consumers. In fact, there is a continuum of consumers with mass m. One advantage of this assumption is that we do not need to bother with assuring that the number of consumers living on a street segment is a natural number (as opposed to a rational number). Another advantage is that the people living at a certain point of the street are negligible.

not exceed v. Otherwise a consumer prefers not to buy the good. More formally, a consumer living at point x buying the good from firm 1 will get utility $u_1(x) = v - p_1 - t \cdot x$ where p_1 is the price charged by firm 1. The same consumer would get utility $u_2(x) = v - p_2 - t \cdot (1 - x)$ if buying from firm 2 instead. It the consumer does not buy from either firm, the utility is zero. Note that the utility of buying from firm 1 depends on where the consumer lives. The utility decreases linearly with distance, as described by the downward sloping line in the figure below: The further away a

Figure 11.3.



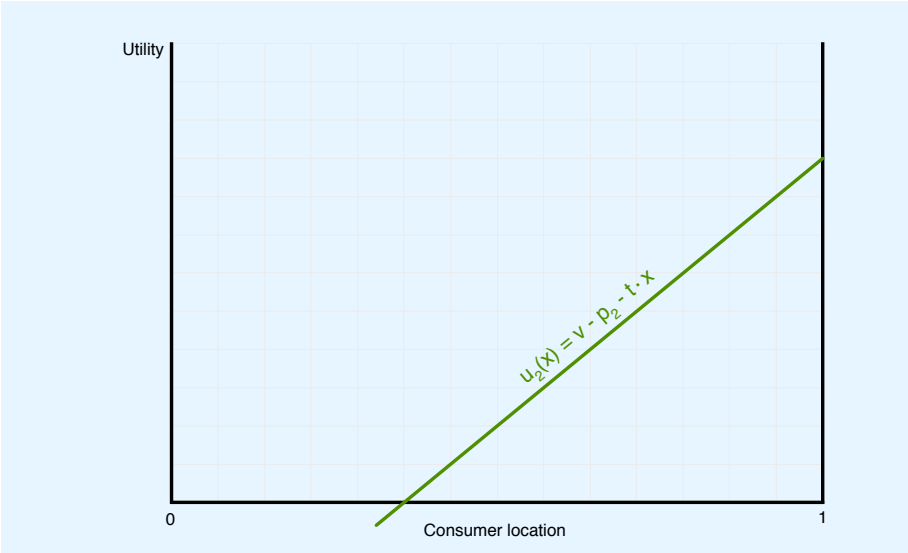
consumer lives from the store, the lower is the utility of buying the good from that store. Consumers living very far away prefer not to buy the good at all to buying it from firm 1.

The "utility line" is higher the larger is the value of the good and the lower is the price. The curve will be steeper the higher is the transportation cost. That is, utility falls quicker with distance, the larger is the transportation cost. Also note that more people prefer to buy the good from firm 1 to not buying at all the higher is the value, v, the lower is the price, p_1, and the lower is the

transportation cost, t.

The utility of buying from firm 2 is larger the further to the right the consumer lives:

Figure 11.4.

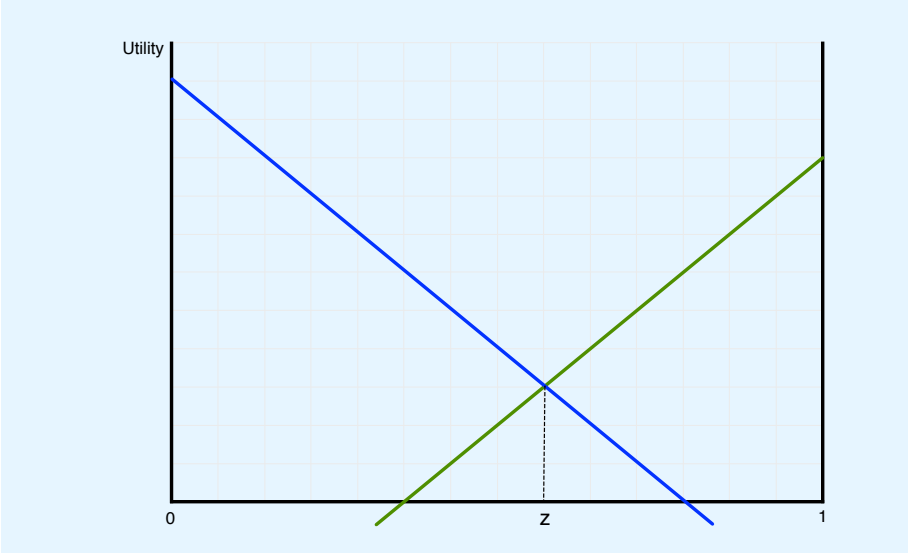


The consumers have the choice to buy from either firm. If the prices are not too high compared to the consumers' valuation, v, all consumers will buy a unit of the good. We say that the market is fully covered. The consumers living at some point z will have the same utility of trading with both firms, $u_1(z) = u_2(z)$ and will thus be indifferent between buying from store 1 and store 2.

Clearly, all consumers living to the left of z will prefer store 1 to store 2 and all consumers living to the right of z will prefer store 2. This means that firm 1 will sell $D_1 = z \cdot m$ units and firm 2 will sell $D_2 = (1 - z) \cdot m$. To simplify notation, we will assume that $m = 1$ from now on.

Comparative statics The firms' demands depend on the prices they charge. If firm 1 reduces its price from p_1 to p_1' more people will prefer firm 1 to firm 2: Firm 1's demand increases from z to

Figure 11.5.

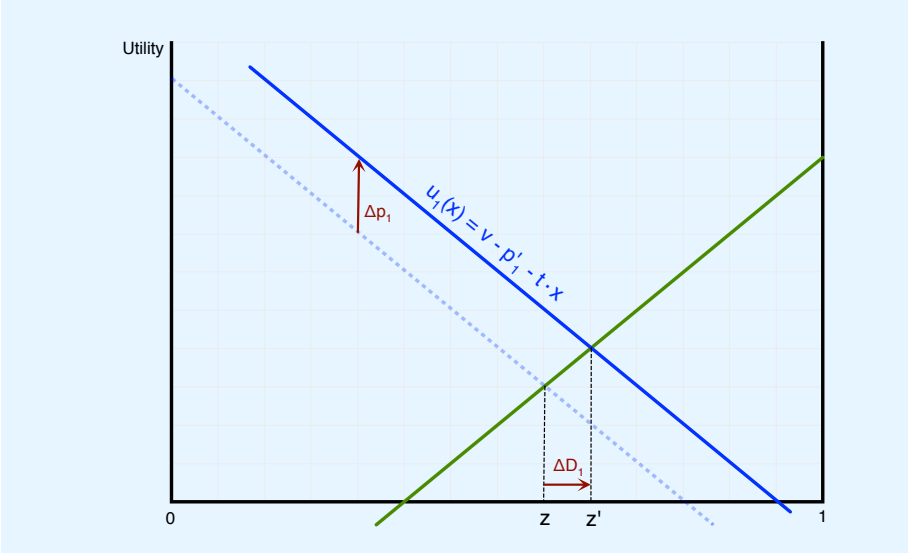


z'. At the same time firm 2's demand is reduced to $1 - z'$.

Note that, unlike the Bertrand model with homogenous goods, a small reduction in price will not lead to a massive expansion of sales. Even if one firm charges a somewhat lower price than the rival, some consumers living close to the high price store will still prefer not to take the travel cost to save on the price. If the transportation cost is small, the utility-curves are flat, meaning that the net utility does not drop very fast with the distance to the store. In this case, a reduction in price will give a more substantial increase in residual demand: That is, residual demand is more sensitive to price changes if the transportation cost is low. If the transportation cost is very low, even a small reduction in price below the competitor's price will make it worth wile also for people living far away to come to the cheaper store.

The algebra To derive demand algebraically, we assume that the firms charge prices that are sufficiently low for the market to be covered. That is, the consumers who are indifferent between the two stores have a positive utility and will buy the good from one

Figure 11.6.



of the stores. We also assume that the difference between the two prices is sufficiently small for both firms to sell a positive quantity.[2]

All consumers living at a location x such that $u_1(x) > u_2(x)$ will buy from firm 1. This condition can be rewritten as $v - p_1 - t \cdot x > v - p_2 - t \cdot (1 - x)$. Solving the inequality for x:

$$x < \frac{1}{2} + \frac{p_2 - p_1}{2 \cdot t} \equiv z.$$

All consumers who live at a location x below z will buy from firm 1. Firm 1's demand is given by

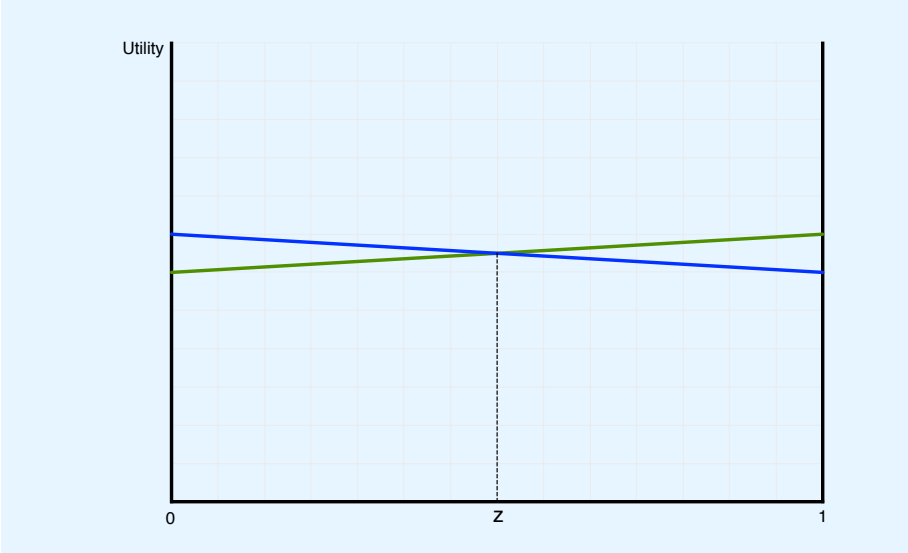
$$D_1(p_1, p_2) = \frac{1}{2} + \frac{p_2 - p_1}{2 \cdot t}.$$

Firm 2's demand is given by

$$D_2(p_1, p_2) = \frac{1}{2} + \frac{p_1 - p_2}{2 \cdot t}.$$

[2]It is easy to verify that the market is fully covered if $p_1 + p_2 < 2 \cdot v - t$ and that both firms sell a positive quantity if the difference in prices is less than t. Later it will be possible to verify that the firms indeed will charge prices that satisfy these conditions under certain circumstances.

Figure 11.7.



Firm 1's demand curve is thus linear at prices that are not too far from the rival's price: As already seen, demand is more price-sensitive if the transportation cost is low. The linear segment of the demand curve pivots around the point $\left(\frac{1}{2}, p_2\right)$ when t is changed. That is, if firm 1 charges the same price as the rival, firm 1 will always supply half the market. If the transportation cost is reduced a price increase will lower demand more and a price reduction will increase demand more. That is to say, the lower the transportation cost is, the more elastic is residual demand.

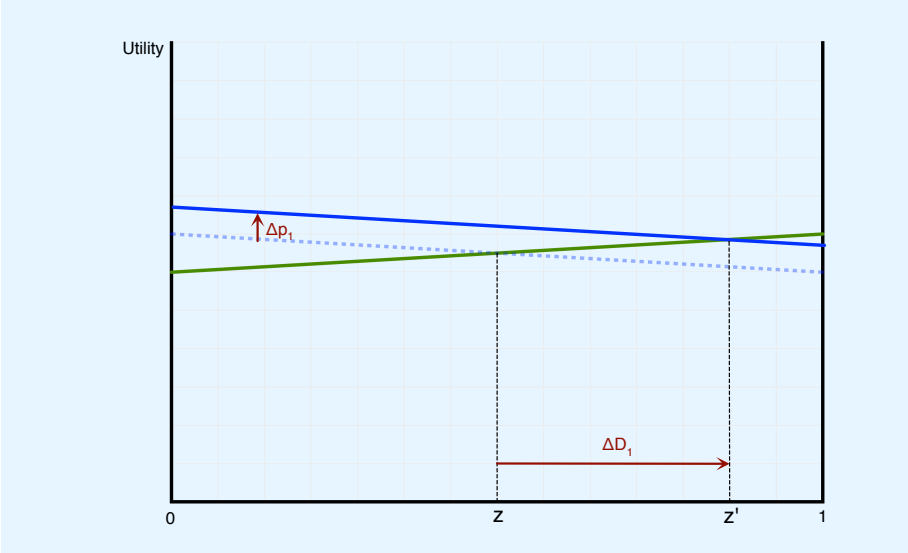
Exercise 47. What happens to residual demand if the value of the good, v, is increased? Explain!

11.1.3. Price-Setting Game

Firm 1's profit is given by the markup times quantity sold, i.e. $\pi_1(p_1, p_2) = (p_1 - c) \cdot D_1(p_1, p_2)$. That is

$$\pi_1(p_1, p_2) = (p_1 - c) \cdot \left[\frac{1}{2} + \frac{p_2 - p_1}{2 \cdot t}\right]$$

Figure 11.8.



where the second factor on the right hand side is firm1's residual demand.

Best reply functions Firm 1's first-order condition is

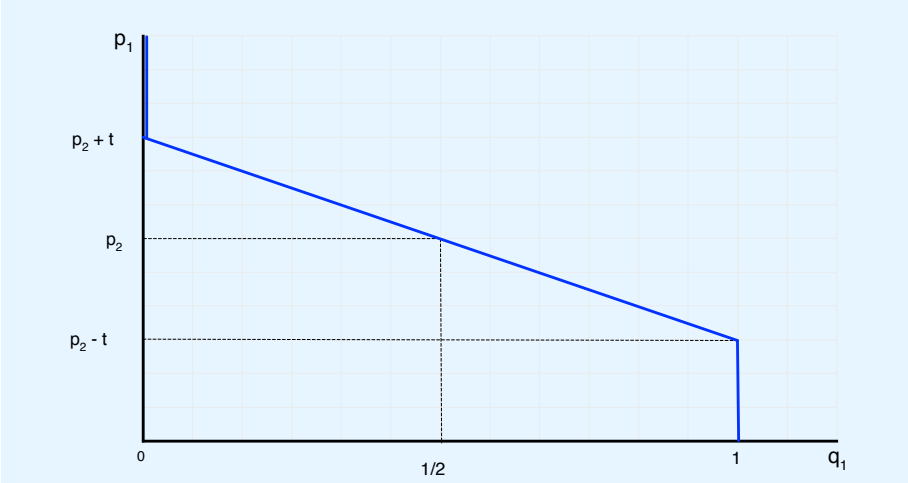
$$\frac{\partial \pi_1 (p_1, p_2)}{\partial p_1} = \left[\frac{1}{2} + \frac{p_2 - p_1}{2 \cdot t}\right] - (p_1 - c) \cdot \frac{1}{2 \cdot t} = 0$$

where the first term is the benefit of increasing the price by one unit (i.e. the number of goods sold) and the second term is the cost of increasing price (i.e. the profit margin times the reduction in demand). Solving the first-order condition for the own price gives us firm 1's best-reply function:

$$p_1 = \frac{c + t}{2} + \frac{1}{2} \cdot p_2.$$

The best reply function is linear in the rival's price and is displayed in the following figure: Clearly, the higher the price charged by the rival, the higher price firm 1 prefers to set. Firm 1 will also prefer to charge a higher price, independent of the price charged

Figure 11.9.



by the rival, if the cost of production goes up or if the consumers' transportation cost goes up.

Exercise 48. Derive firm 2's best-reply function.

Firm 2's best-reply function is displayed in the figure.

Equilibrium For the two prices p_1 and p_2 to be an equilibrium, it is required that firm 1's price is a best reply to firm 2's price and that firm 2's price is a best reply to firm 1's price. The equilibrium is described by the following figure:

The algebra The equilibrium can also be computed algebraically, using the two best reply functions:

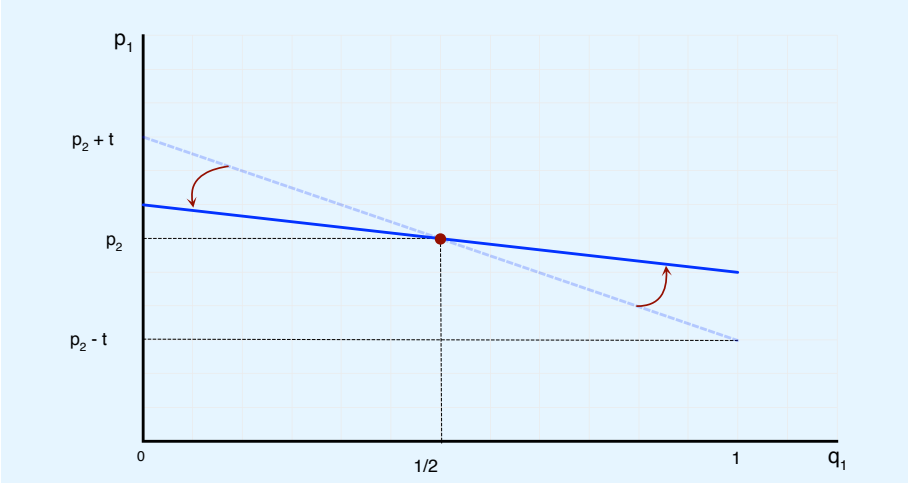
$$p_1 = \frac{c+t}{2} + \frac{1}{2} \cdot p_2.$$

and

$$p_2 = \frac{c+t}{2} + \frac{1}{2} \cdot p_1.$$

This is a system of two linear equations in two unknowns, the prices. The easiest way to solve this equation system is to substitute one

Figure 11.10.



best-reply function into the other:

$$p_1 = \frac{c+t}{2} + \frac{1}{2} \cdot \left[\frac{c+t}{2} + \frac{1}{2} \cdot p_1 \right].$$

We now have one linear equation in one unknown. The solution is easy to compute: $p_1 = c + t$. Finally, substituting firm 1's equilibrium price into firm 2's best-reply function shows that also firm 2 will charge the same price. Hence, the equilibrium is symmetric and given by

$$p_1 = p_2 = p = c + t.$$

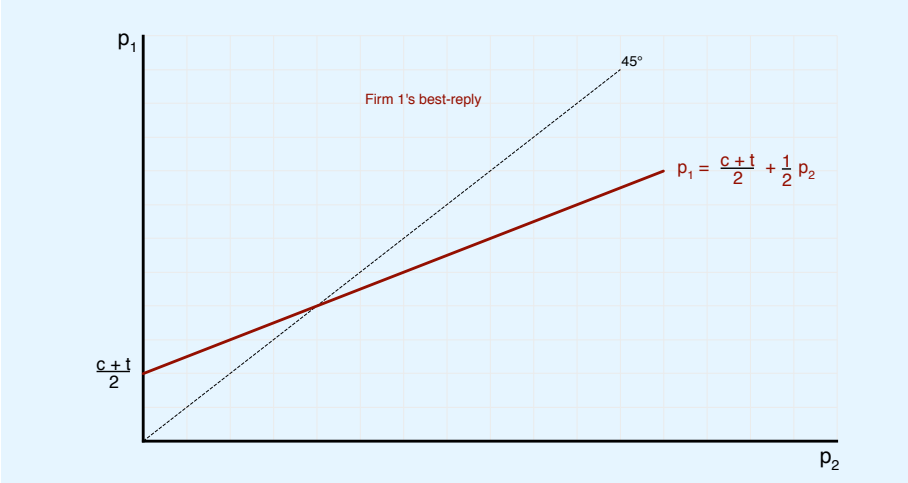
In equilibrium, firm profits will be determined by the transportation cost,

$$\pi_i = (p - c) \cdot \frac{1}{2} = t/2.$$

11.1.4. Comparative Statics

Exactly as the Bertrand model, with homogenous goods, the equilibrium price is increasing in the cost of production. The pass-on is unity: one Euro more in cost leads to a one Euro higher price. Also as in the Bertrand model, the equilibrium price is independent

Figure 11.11.



of the consumers' valuation of the good, v. The value is so-to-say competed away completely.

The difference to the Bertrand model is the transportation cost, t. If the consumers have to pay a positive transportation cost, the firms will be able to charge a price above their marginal cost in equilibrium. That is, the firms will have some market power. The reason why the firms have market power is that the transportation cost makes the two stores different in the eyes of the consumers. It matters to which one you go. Some people live closer to firm 1's store and will prefer that store to firm 2's store, even if firm 1 would charge a somewhat higher price. Firm 1 exploits this advantage that it has over Firm 2 in catering for the people living nearby.

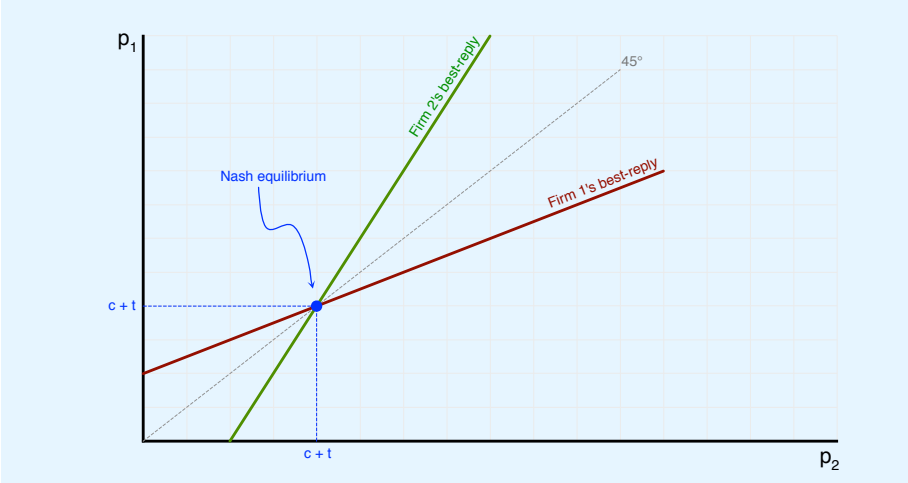
Exercise 49. Consider the same model as above, but assume that the two stores have different marginal costs. Solve the model and show that the the high cost firm will charge a higher price even though the two firms can be considered to offer the same quality. (Quality is not easily defined, but (i) the value v is the same and (ii) there are equally many people living close to store 1 as there are living close to store 2.)

Figure 11.12.

11.1.5. Interpretation of the transportation cost

The Hoteling model can be interpreted many different ways. The most obvious interpretation is that the two firms have different geographical locations and that the consumers live in-between the stores. But the Hoteling-line may also represent different product characteristics, e.g. the sweetness in soft drinks. One firm's drink may be sweeter than the other firm's drink and different people have different preferences for sweetness. In this case the line may measure the amount of sugar in the drink. A drink without sugar is located at point zero and a drink with one unit of sugar is located at point one. A consumer's location on the line indicates that consumer's ideal soft drink. If your ideal drink would contain x units of sugar you "live" at point x on the line. Those who are located close to the second firm prefer a rather sweet drink and will tend to buy the second firm's product. The drink is not perfect, according to their tastes, but it is certainly better than the only other available alternative. The transportation cost describes how important the consumers consider sweetness in drinks. If the transportation cost is close to zero, they do not care much. Moving away from the

Figure 11.13.



ideal drink doesn't reduce the value by much and the two drinks are rather good substitutes. If the transportation cost is large, the consumers have strong preferences and those who like sweet drinks are prepared to pay a high premium for their preferred alternative. In this case, the Hoteling model suggests that the prices of soft drinks will be much higher than the production cost.

11.2. Vertical Differentiation

BMWs are better cars than Skodas. End of discussion. When people agree that one product is better than the other, we say that the two goods are vertically differentiated. Clearly what a person buys also depends on the price difference and how important the consumer think the quality difference is in comparison to the price difference.

11.2.1. "Vertical Hoteling"

This section presents a model of oligopolistic interaction when goods are vertically differentiated, which is similar to the Hoteling model

discussed in the previous section.

Consumers Consider a good that may be produced in different qualities. Each consumer either buys one or no unit of the good. The utility if not buying is zero and the utility of buying is given by

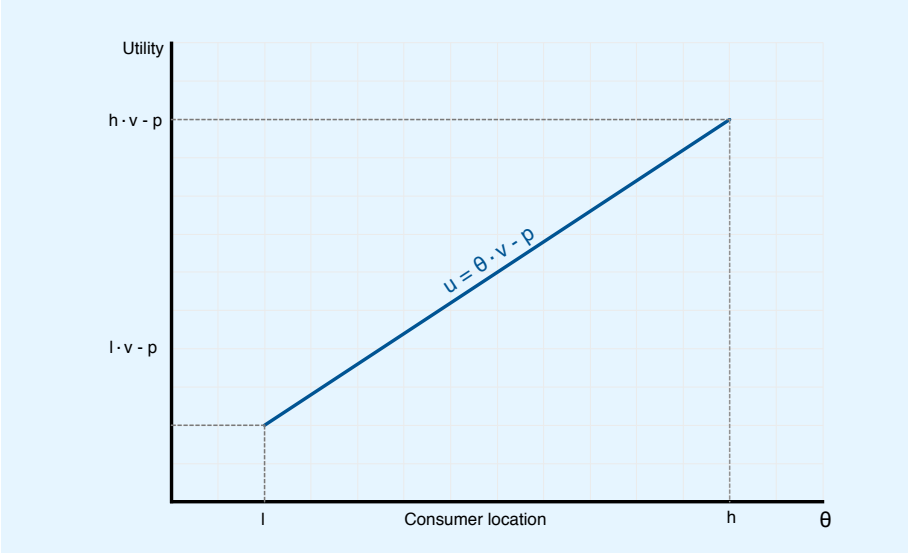
$$u = \theta \cdot v - p$$

where v is the quality of the good and p is the price. The quality, v, is an aspect that all consumers agree to be valuable, e.g. safety in a car. All consumers agree that a good with a high v (a car with high safety) is better than a good with a lower v (a car with low safety).

The parameter θ describes the consumer's taste for quality, i.e. how important the consumer considers quality to be relative to price. If the quality is increased by one unit, the consumer is willing to pay θ more for the good. A consumer with a high θ considers quality to be important. Different consumers have different tastes for quality. We may think of the consumers as living on a street between l and $h > l$ and that their location on the street exactly matches their taste for quality: Perhaps the left end of the street is a poor neighborhood while the right end of the street is wealthy. Anyway, a person living at l would get net utility $l \cdot v - p$ from the car and a person living at h would get net utility $h \cdot v - p$. For simplicity we assume that the consumers are uniformly distributed between the two end-points. To measure consumer heterogeneity we define $\Delta\theta = h - l$. If $\Delta\theta$ is large, there is a large difference in the taste for quality between different consumers.

Clearly, an increase in price leads to a lower net utility for all consumers, independent of their taste for quality: An increase in quality increases the net utility for all consumers: Note that increased quality increases the utility more for those who have a higher taste for quality.

Figure 11.14.



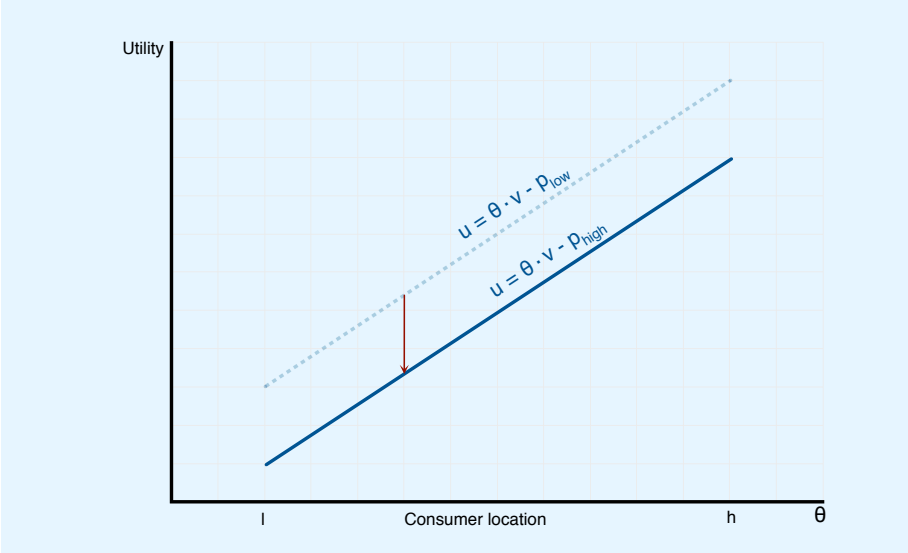
Firms There are two firms called L and H, producing goods of different qualities. In particular, it is assumed that H produces a higher quality than L, i.e. $v_H > v_L$. The quality difference is $\Delta v = v_H - v_L > 0$. Despite the difference in quality, the two firms have the same marginal cost of producing their goods, c.

Market Rules The rules of the game is that first the two firms simultaneously decide on their prices p_H and p_L and make them public and then the consumers decide whether they want to buy a unit of the good and from which firm to buy it. Working backwards, we start out solving the second stage of the game, i.e. with deriving demand.

11.2.2. Deriving the firms' residual demand

To derive demand we need to determine which consumers buy from L and which buy from H. The utility of buying the low-quality good is depends on the consumer's taste for quality $u_L(\theta) = \theta \cdot v_L - p_L$ and is described by the graph below. Also the utility of buying

Figure 11.15.

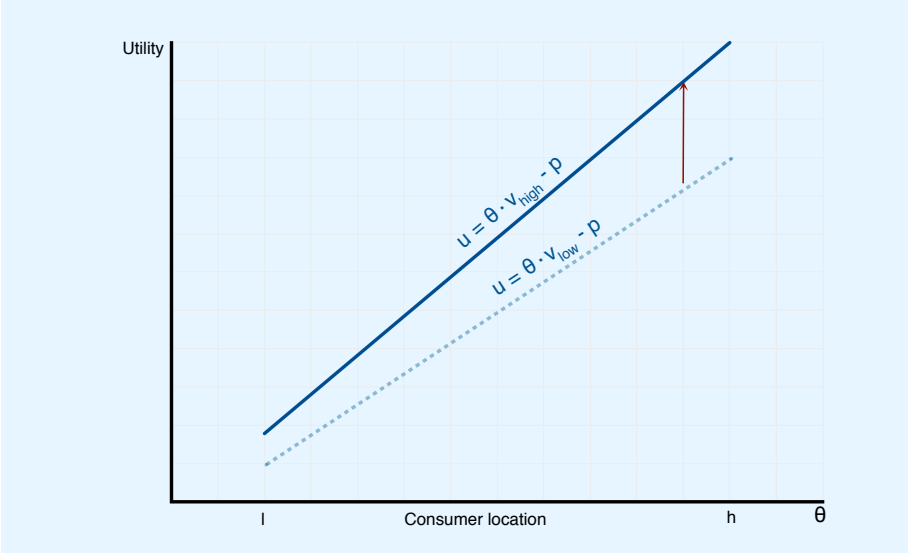


the low-quality good is depends on the consumer's taste for quality $u_H(\theta) = \theta \cdot v_H - p_H$ and is described by the graph below. Note that it is assumed that the price of the high-quality good is somewhat higher than the price of the low-quality good. The utility of buying the high-quality good may be lower for people with little taste for quality, due to the higher price. However, for people with more taste for quality the higher price does not matter as much and they prefer the high-quality good to the low-quality good.

Somewhere on the street, there is a "marginal consumer" who is indifferent between the two qualities. This person has taste for quality $\theta = z$. Clearly, all people living to the left of z strictly prefer the low-quality good and all people living to the right of z prefer the high-quality good.

That is, people living on the segment between z and l buy from L and people living on the segment between h and z buy from H. Since the population is uniformly distributed over the street, the number of people living on a certain segment of the street is proportional to the length of that segment. With a total population of m, the

Figure 11.16.



demand for the two goods is therefore

$$D_H = \frac{h-z}{h-l} \cdot m = \frac{h-z}{\Delta\theta} \cdot m$$

and

$$D_L = \frac{z-l}{h-l} \cdot m = \frac{z-l}{\Delta\theta} \cdot m.$$

To say more, we need to find the exact location of the "marginal consumer," i.e. compute z. A consumer with taste θ is indifferent between the two goods if $u_L(\theta) = u_H(\theta)$. The indifferent consumer's taste parameter z must therefore be determined by:

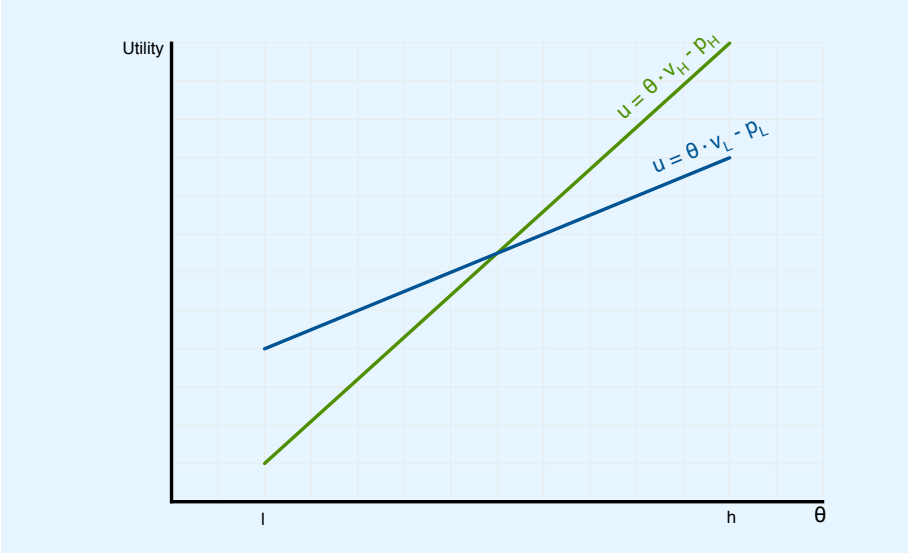
$$u_H(z) = z \cdot v_H - p_H = z \cdot v_L - p_L = u_L(z).$$

Solving the equation for z gives:

$$z = \frac{p_H - p_L}{v_H - v_L} = \frac{p_H - p_L}{\Delta v}.$$

Substituting z into demand, assuming $m = 1$, and simplifying gives:

Figure 11.17.



$$D_H(p_H, p_L) = \frac{h \cdot \Delta v - (p_H - p_L)}{\Delta\theta \cdot \Delta v}$$

and

$$D_L(p_H, p_L) = \frac{(p_H - p_L) - l \cdot \Delta v}{\Delta\theta \cdot \Delta v}.$$

Note that the demand functions are linear in the prices.

11.2.3. Price-Setting Game

Firm H's profit is given by

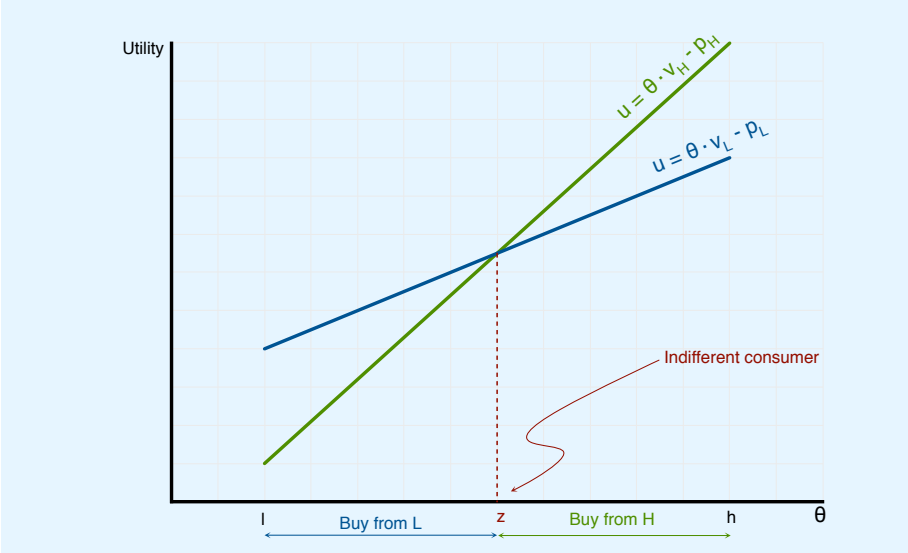
$$\pi_H(p_H, p_L) = (p_H - c) \cdot \frac{h \cdot \Delta v - (p_H - p_L)}{\Delta\theta \cdot \Delta v}.$$

Firm L's profit is given by a similar expression.

Best-reply functions Firm H's first-order condition is

$$\frac{\partial \pi_H(p_H, p_L)}{\partial p_H} = \frac{h \cdot \Delta v - (p_H - p_L)}{\Delta\theta \cdot \Delta v} - (p_H - c) \cdot \frac{1}{\Delta\theta \cdot \Delta v} = 0.$$

Figure 11.18.



Solving the first-order condition for p_H, shows that firm H's best-reply function is given by

$$p_H = \frac{c + h \cdot \Delta v}{2} + \frac{1}{2} \cdot p_L.$$

Similarly, firm L's best-reply function is given by

$$p_L = \frac{c - l \cdot \Delta v}{2} + \frac{1}{2} \cdot p_H.$$

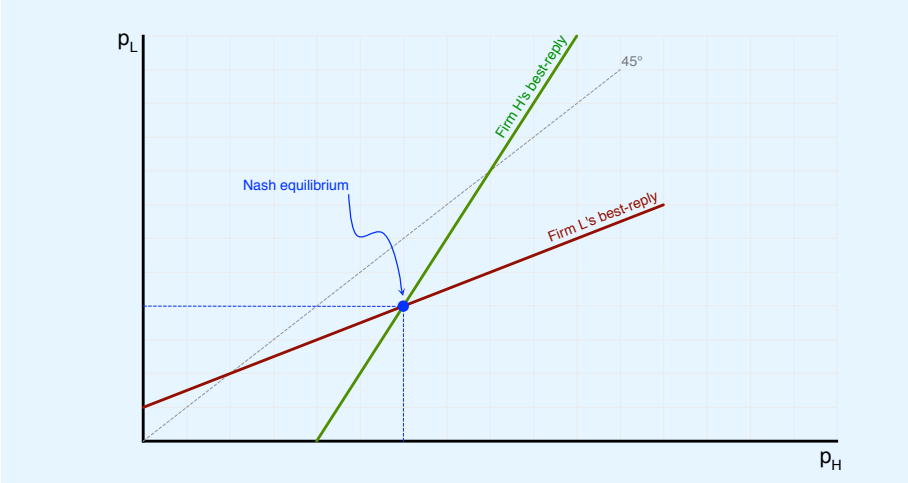
Exercise 50. Derive L's best-reply function.

The two best-reply functions are linear and described in the following figure:

Equilibrium For the two prices p_H and p_L to be an equilibrium, it is required that firm H's price is a best reply to firm L's price and that firm L's price is a best reply to firm H's price. The equilibrium is described by the intersection of the two best-reply functions in the figure above.

The equilibrium can also be computed algebraically, using the

Figure 11.19.



two best reply functions. They constitute a system of two linear equations in two unknowns, the prices. The easiest way to solve this equation system is to substitute one best-reply function into the other:

$$p_H = \frac{c + h \cdot \Delta v}{2} + \frac{1}{2} \cdot \left[\frac{c - l \cdot \Delta v}{2} + \frac{1}{2} \cdot p_H \right].$$

We now have one linear equation in one unknown. The solution is easy to compute:

$$p_H = c + \frac{1}{3} \cdot (\Delta\theta \cdot \Delta v + h \cdot \Delta v).$$

Finally, substituting firm H's equilibrium price into firm L's best-reply function shows that

$$p_L = c + \frac{1}{3} \cdot (\Delta\theta \cdot \Delta v - l \cdot \Delta v).$$

It is easy to see that the high-quality firm charges a higher price than the low-quality firm in equilibrium.

Comparative Statics If the two firms would produce the same quality, $\Delta v = 0$, the equilibrium price is equal to the marginal cost, exactly as in the Bertrand model. If all consumers would have approximately the same taste for quality, $\Delta\theta \approx 0$, the low-quality firm would not be viable. Instead the high quality firm would sell to all the consumers.

If there is a quality difference, $\Delta v > 0$, and if there is sufficient consumer heterogeneity, $\Delta\theta > l$, both firms will supply some customers and both prices are above marginal cost. Both firms have market power. In fact, in this case, the two firms focus on different consumers. The high-quality firm sells to high-value consumers and the low-quality firm sells to low-value consumers. Even though the high-quality firm's price is higher than the low-quality firm's price, some consumers continue to buy from H, thanks to the higher quality. It would be unprofitable for H to lower its price. Even if it would poach customers from L, it would also have to reduce the price to its "captive" customers. One may also show that the high-quality firm has a higher market share and a higher profit.

The most surprising result of this model is that the low-quality firm, would actually not want to increase its quality. Higher quality would clearly make all consumers willing to pay more for L's product, but the higher quality would also reduce the difference between the two firms, which would trigger a more aggressive pricing from the rival.

Consumer price information

To be included.

12. Capacity Constraints

To provide mobile broadband services the telecom companies have to build a network of base stations. In rural areas, the density of base stations is made just sufficient to provide uninterrupted connection in the whole geographical area they wish to cover. This minimal capacity will never be fully utilized. In more populated areas, the companies have to increase capacity beyond the minimal level, primarily by installing more radio frequencies in the base stations but also by setting up a denser network of stations. Through their choices of capacities the companies decide how many mobile broadband subscriptions (of a certain speed) that they will be able to sell to their customers.

The mobile broadband services provided by different networks are relatively similar. There are differences in network coverage, especially in rural areas. But for most urban customers, these differences are probably of minor importance. It seems likely, therefore, that a firm should be able to attract a large share of all consumers by offering the lowest prices on the market. Once the necessary network capacity is built, the marginal cost of providing broadband services is close to zero. With Bertrand competition we should therefore expect prices close to zero. It would then be impossible for multiple networks to cover their investment cost. Only one network would be commercially viable.

But this is not what we observe. Instead, prices are substantially higher than marginal cost and several networks are able to operate side by side. One important reason is that the capacity constraints have a profound effect on the pricing of mobile broadband services.

Even if a small price cut could attract many new customers, there is no point in cutting price when these customers cannot be served anyway. Capacity constraints reduce price competition.

But referring to capacity constraints is only a partial explanation. We also need to explain why the firms don't build more capacity, to be able to poach customers from their rivals. We will discuss also this issue here. To prepare for studying the role of capacity constraints in oligopoly, we start out with the simpler study of a monopoly with capacity constraints.

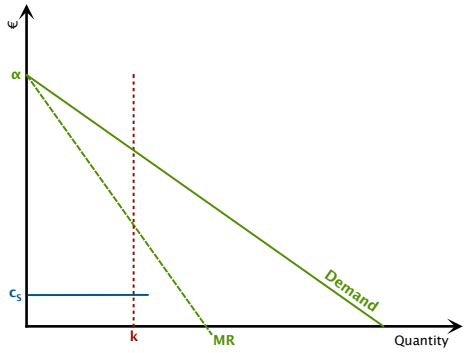
12.1. Monopoly with capacity constraints

Consider a monopoly firm with linear (inverse) demand, given by $p = P(q) = \alpha - \beta \cdot q$. The monopolist can produce its product at constant marginal cost of production, c_S, but only up to a capacity constraint, k. The marginal cost may for example represent hours worked and consumption of materials. The capacity constraint may represent a limitation in the installed equipment, for example a machine that can only produce a certain maximal volume per day. The situation is summarized in Figure 12.1. In this case the monopolist cannot produce the quantity equalizing marginal revenue to the marginal cost c_S; the capacity is simply not sufficient. (Actually marginal cost is not even well defined for quantities beyond the capacity constraint; it is simply impossible to produce an additional unit *at any cost*.)

12.1.1. Optimal price and quantity

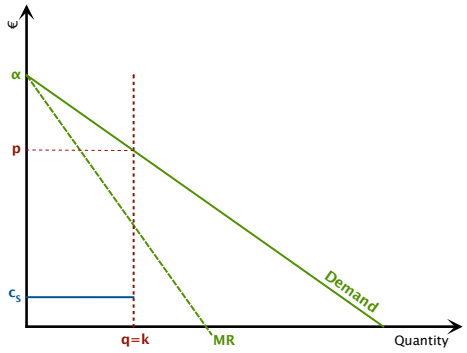
When the capacity constraint is binding, it is easy to see that the monopolist maximizes profits by producing as much as it possibly can, that is $q = k$, and to set the market clearing price, that is

Figure 12.1.: Monopoly with capacity constraint



$p = P(k)$, as described in Figure 12.2.

Figure 12.2.: Capacity constrained monopoly.



Exercise 51. Assume that demand is given by q = 10 − p and that the monopolist has a capacity to produce at most 2 units of output. What price would the monopolist set?

As a result, the monopolist charges an even higher price than it would have set absent the capacity constraint. Note that we here simply assume that the monopolist's capacity is too small to produce the output equalizing marginal revenue to marginal cost.

That is, we simply assume that the capacity constraint is binding. We will return to this issue below, after discussing how capacity is set.

12.1.2. Choice of capacity

The question is then why the monopolist has chosen to install such a small capacity? To answer this question we have to consider a two-period model: in the first period the monopolist installs capacity and in the second it sets price. As usual it is convenient to solve the model backwards.

This means that we first determine what price the monopolist will set for every possible capacity it may have chosen in the first period. And this is exactly what we did above. Thus, in the second period, the monopolist chooses quantity $q = k$ and price $p = P(k)$, where k is the capacity chosen in period one. The monopolist's second-period profit is thus given by $\pi(k) = [P(k) - c_S] \cdot k$.

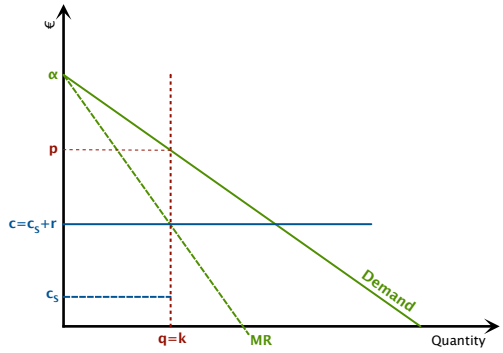
In period one, the monopolist installs the capacity. Capacity is any resources such as machines that are fixed in the second period. The price of a unit of capacity is denoted r and the monopolist thus has to select k to maximize

$$\Pi = \pi(k) - r \cdot k = [P(k) - c_S - r] \cdot k = [P(k) - c] \cdot k.$$

Note that this maximization problem looks very similar to the usual monopoly quantity setting problem. The only difference is that the marginal cost here, $c = c_S + r$, is the long-run marginal cost. The long-run marginal cost includes both the short-term costs such as working hours and materials, c_S , and the cost of installing capital, r.

The solution to this maximization problem is to install the capacity equalizing long-run marginal cost to the marginal revenue, as described by Figure 12.3. Note that since the long-run marginal

Figure 12.3.: Monopolists choice of capacity



cost is higher than the short-run marginal cost (when defined), the capacity constraint will always be binding when the monopolist chooses output. Another way to say this is that the monopolist will not install more capacity in period 1 than it knows it will use in period 2.

12.2. Oligopoly with capacity constraints

Consider the same market as above with linear (inverse) demand, given by $p = P(q) = \alpha - \beta \cdot q$. But now there are $n \geq 2$ firms active in the market selling the same homogenous good. All firms can produce at the same constant marginal cost, c_S, but only up to their respective capacity constraints, k_i, which may be different.

12.2.1. Price setting

The firms set their prices simultaneously. Then consumers decide which firm to buy from. The firms are, however, only able to supply consumers up to their capacity constraints. If there is excess demand, some people will not be served.

Even if this model looks simple, the analysis turns out to be quite complicated. But luckily the equilibrium outcome is easy to describe. In equilibrium, all firms decide to charge the price that clears the market, given the available capacities, that is

$$p_i = P\left(\sum_{j=1}^{n} k_j\right).$$

Since all firms charge the same price, consumers are willing to buy from them all. That is, all firms will be able to produce and sell according to their capacity constraints, that is

$$q_i = k_i.$$

At least this is the unique equilibrium under certain conditions. The main requirement is that all firms have relatively tight capacity constraints, which will indeed be the case if the cost of installing capital is relatively high. To keep things simple, we assume that this condition is fulfilled and skip the proof.

12.2.2. Capacities

The next step in the analysis is then to study the firms' choices of capacities. We assume that firms choose capacities simultaneously. But since we already know that capacities and production levels will be the same in the end, $q_i = k_i$, there is not really much point in making a distinction between the two. For simplicity we will therefore say that firms choose quantities. But it should be remembered that this is just shorthand for capacity. It should also be remembered that the marginal cost should be thought of as including the cost of installing capacity.

As it turns out, the model we have just defined is equivalent to the Cournot model, which is a commonly used model of oligopolistic interaction. This model has already been discussed in section 6.4.

Exercises

Sources of market power

Give some examples of why oligopoly firms may charge prices above marginal cost despite competition from rival firms. Which of these factors do you believe are particularly relevant for the car market?

Part IV.

Measuring competition

One of my first consulting assignments was to write a report for a government commission, set up to reform the rules of merger control in the Swedish competition act. The commission wanted to know if there is any threshold level for concentration above which the firms in a market stop competing with each other. If such a threshold exists, merger policy would be greatly simplified. The competition authority would still need to define the relevant market and to predict how concentration would be affected by the merger, e.g. by summing the market shares of the merging firms. But then the decision would be straightforward: allow a deal to proceed if, and only if, concentration remains below the threshold. Such a simple rule would also benefit the firms, making it relatively easy to predict the authority's decision and thereby reducing uncertainty.

As it turns out there exists a large number of empirical studies that address this and related questions. I am referring to the so-called Structure-Conduct- Performance (SCP) literature, which was the dominating line of research in industrial economics during the 1950:ies and 1960:ies. The basic idea was to compare different industries (say cars, trucks, sugar, ...) to see if *e.g.* firm profits are much higher in industries with high concentration. We will review this literature in the first chapter of this part of the book.

Most of this part, however, is devoted to the more recent empirical methods that are used to measure the strength of competition. This line of research, often referred to as the New Empirical Industrial Organization (NEIO), advocates a strict focus on individual markets. Looking at (say) the sugar market, is it possible to tell whether the firms are price takers or if they are all colluding in a cartel?

13. Structure-Performance-Conduct

The purpose of the Structure-Conduct-Performance methodology was to establish the relationship between market structure, *e.g.* market concentration and entry barriers, and market performance, *e.g.* measured by profits or price cost margins.[1] The idea is that market structure (S) determines firm conduct (C), *e.g.* cartel membership, which in turn determines market performance (P). That is, causality runs from structure to performance. Since firm conduct is difficult to observe, this intermediate link is typically left out. Instead, it is investigated how performance depends on structure directly.

Another important idea of the SCP-methodology was to establish relationships between market structure and performance that are robust and the same across different industries. Again, knowing the robust relationship between market structure and market performance (assuming it exists) would *e.g.* be useful for the design of merger policy. Suppose that one would discover that firms have market power (as revealed by high profits) when concentration is above some critical level but not below this level, then it might make sense to block all mergers increasing concentration beyond the critical level and to allow those leaving market concentration

[1]This section is based on Schmalensee, R. (1989). "Inter-Industry Studies of Structure and Performance." *Handbook of Industrial Organization.* R. Schmalensee and R. Willig (eds.). Amsterdam: North-Holland. Church, J. & Ware, R. (2000). *Industrial Organization: A Strategic Approach.* McGraw-Hill.

below the critical level.

13.1. Methodology

There are numerous SCP-studies and while they differ in some respects, they all shared the same basic methodology. The key idea was to first collect data on concentration and profitability in various industries and then to compare these industries to see if there is a relationship between concentration and profitability: Do firms in more concentrated markets earn higher profits?

To this end the researchers estimated some equation, such as

$$Performance_i = \alpha + \beta \cdot Concentration_i + \gamma \cdot Barrier_i$$

where $Performance_i$, $Concentration_i$ and $Barrier_i$ are some measure of market performance, market concentration and entry barriers, the index i stands for different observations (*i.e.* different industries) and α , β and γ are the parameters to be estimated.

The basic assumptions behind this formulation are that the structural variables are exogenous and that they have the same effect on performance in all industries. This particular formulation also assumes that the relation between structural variables and performance can be approximated by a linear function.

Market performance was measured in different ways in different studies. Some common measures are

- Accounting profits
- Rate of return on investments = (revenues – variable cost – economic depreciation)/Book value of assets
- Lerner index, approximated by (revenues – variable cost)/revenues
- Tobin's q = market value of firm/replacement value of assets

Concentration was sometimes measured by the Herfindahl index (the sum of squared market shares of all firms) and sometimes by the four-firm concentration-ratio "CR4" (the sum of the four largest firms' market shares).

Entry barriers include economies of scale, which can be measured by the so- called minimum efficient scale in relation to industry sales, where the minimum efficient scale is an assessment of how much a firm must produce in order to produce at the lowest possible cost. Other entry barriers are product differentiation and brand loyalty, which often where measured by advertising and R&D expenditures.

It was also common to include control variables in the estimation, to take out the effect of other factors that might influence profitability. Examples include the degree of unionization and the presence of buyer power.

13.2. Results

The first hypothesis to test is that market power increases with concentration, that $\beta > 0$ in the equation above. Most studies found a positive relation between concentration and performance, but the economic effect of concentration was often small. One representative study found that a 50% increase in CR4 would increase profit margins by 6%. In addition, the effect typically has a low statistical significance.

The second hypothesis to test is that market power increases with entry barriers, that $\gamma > 0$ in the equation above. These results are more robust and significant. There appear to be a positive relationship between profitability and scale economies, advertising (in consumer goods industries) and R&D.

13.3. Critique

Markets differ The SCP-methodology has later fallen out of fashion as a result of a number of important shortcomings. The first problems are due to basic methodology of comparing different industries. Some important determinants of profitability such as the elasticity of demand differ greatly between industries. Since, demand elasticities are not readily observable, they are typically not included in the studies, giving rise to omitted variable biases.

A related problem is that the effect of concentration on profitability probably is different in different industries. In some industries it might be possible for the firms to cartelize the market if it becomes highly concentrated. In other industries, cartelization is probably much more difficult, *e.g.* as a result of demand variability. Still, the SCP-methodology presumes that the effect of concentration is the same in all industries.

Measurement problems Some measurement problems arise when defining markets (= observations). The correct definition of a market is based on the cross elasticities of demand. But for practical reasons, the econometric studies typically use the standard industrial classification (SIC). This implies, for example, that airplanes and helicopters are considered to be parts of the same market and that all markets are thought of as national markets.

Another problem is that plants producing several different types of products are classified based on their primary activities.

Other measurement problems pertain to specific variables. To measure profitability one needs to measure the depreciation of assets, which is the reduction in the value of the assets. But for practical reasons most studies have to rely on the historic costs of the assets and some arbitrary rules for allocating this cost to different time periods. Moreover, some intangible assets such as advertising and training are not treated as investments but as costs

during the year, due to accounting conventions.

Endogeneity problems The SCP-methodology also suffers from an endogeneity problem. It is assumed that market concentration can affect market performance, *i.e.* that more firms can lead to tougher competition. In fact, one of the main goals is to quantify this relationship. The problem is that the methodology presumes that market concentration can be treated as an exogenous variable, which is not affected by market performance. But this is not correct. Concentration is determined through a process of entry and exit, which is affected by market performance i.e. profitability (see Chapter 16). If competition for a given number of firms is lax, more firms will tend to enter the market. So what are the likely effects of this endogeneity problem?

- On the one hand, market concentration is determined by entry barriers. In markets with low entry costs there will be many firms and prices will be low. Comparing two markets with different entry costs will therefore suggest a positive relation between concentration and market power. Again, it is this relationship that the SCP-studies wishes to quantify.

- On the other hand, market concentration is also determined by product differentiation. If products are good substitutes prices will be low and few firms will enter the market. Comparing two markets with different degrees of product differentiation will therefore suggest a negative relation between concentration and market power.

Most SCP-studies probably include markets that differ both in terms of entry barriers and in terms of the degree of product differentiation. Therefore we should not expect any clear positive relation between concentration and market power. We should however expect a positive relation between entry barriers and market power. Both these conclusions where indeed borne out by the SCP-studies.

Implications A final problem with the SCP-approach concerns what conclusions we should draw from a study suggesting a positive relation between concentration and profits. The SCP-advocates presumed that the relationship is primarily due to market power being higher in more concentrated industries. If this presumption is correct, it may make sense to block mergers leading to increased concentration. An alternative explanation is that economies of scale are more important in certain industries. In such industries there will only be room for a few firms and these firms will earn high profits primarily due to efficient operations and low costs. If this alternative explanation is correct, a restrictive merger policy would result in less efficient production — a social loss.

14. New Empirical Industrial Organization

14.1. Basic Ideas of NEIO

Modern empirical work in industrial economics has taken a giant leap forward from the SCP-methodology. Today it is emphasized that industries differ greatly, making it rather difficult to extract any meaningful information from a comparison of say the car market and the sugar market. Therefore studies nowadays focus on single markets. Competitive conditions may be better understood by comparing *e.g.* the car market in different countries or the sugar market at different points in time.

Another important difference is that modern studies acknowledge that price-cost margins most often cannot be observed. Marginal costs cannot be found in the firms' annual reports. Instead the marginal costs are inferred from the firms' behaviors: since high cost firms behave differently from low cost firms, we may draw conclusions about a firm's cost by observing its behavior. But, to do so, we need a theory describing the link between unobserved cost and observed behavior. Modern empirical industrial economics is therefore based on the theories of market interaction developed during the "game-theoretic revolution" that started in the late 1970:ies. This often means that the firms' pricing behavior is estimated based on some specific oligopoly model, such as Cournot or Bertrand. Such a model will also require us to estimate the demand functions

and to make specific assumptions concerning the shape of the cost curves. Given such a model we may be able to simulate to what extent the entry of new firms reduces prices, i.e. to understand the link between structure and performance. One may also test which of the oligopoly models that seem to fit the data best. Do the firms compete a la Cournot or Bertrand? Or, do they seem to be involved in a cartel?

The drawback with the new methodology is that it often requires much more data. In particular, in markets with product differentiation it is necessary to collect firm level data on *e.g.* prices.

14.1.1. What is conduct?

Market power is the ability to charge a price above marginal cost without loosing all customers. A firm's market power can be measured by the Lerner index, defined as the firm's mark-up as share of the price, *i.e.*

$$L = \frac{p - c}{p}$$

where p is price and c is marginal cost. Market power is the opposite of competition. How much market power does a firm have? There are, of course, many factors that influence a firm's markup. Here we will focus on one that may be called conduct. To illustrate what it is, we will consider some well-known examples.

Example 1: Monopolies & Cartels One may think of monopoly as the complete absence of competition. But market power varies also across monopoly markets. According to the inverse-elasticity rule, a monopoly firm's market power is determined by how price sensitive the consumers are, that is

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

where ε is the price elasticity of market demand. (Note that here the elasticity is defined to be negative, i.e. $\varepsilon < 0$.) In case it is easy for the consumers to stop consuming the product sold by a certain monopolist, to start consuming other products instead, if the monopolist increases its price, then the monopolist does not have much market power.[1]

Of particular interest here is that the inverse elasticity rule also describes the firms' market power in a "perfect" cartel. That is, if all the firms in the market succeed to collaborate in their sales and pricing decisions and act as an integrated firm - as a monopolist - their market power will of course be described by the inverse elasticity rule.

Example 2: Cournot oligopoly One may show that the average market power of the firms in a Cournot market, with homogenous goods, is given by

$$L = \frac{p - c}{p} = -\frac{1}{\varepsilon} \cdot HHI$$

where

$$HHI = \sum_{i=1}^{n} s_i^2$$

is the Herfindahl concentration index. Recall that $HHI = 1/n$ if all the firms are of the same size. In particular, $HHI = 1$ in case of monopoly, and $HHI \approx 0$ if there is a very large number of firms in the market. In a Cournot model, the firms' market power is determined not only by the elasticity of substitution, but also by market concentration.

[1] The distinction between monopoly and competition is a bit blurred. One could argue that the monopolist competes with all the other products that the consumers switch to in case of a price increase. But there is a difference. Unlike an oligopolist, the monopolist doesn't loose customers to any particular other firm; the monopolist's consumers switch to numerous other firms. The link between the monopolist's product and every single other product is therefore weak.

Example 3: Bertrand oligopoly Bertrand firms set price equal to marginal cost and cannot earn any markup. In other words, Bertrand firms do not have any market power at all, that is

$$L = \frac{p - c}{p} = -\frac{1}{\varepsilon} \cdot 0$$

The same formula also describes the market power of firms in a market with perfect competition.

Conduct - A synthesis It is actually possible to summarize all these results in one single equation. That is, we can use a single equation to describe the firms' market power independent of whether they form a cartel or compete a la Bertrand or a la Cournot. To do so, we need define the notion of conduct to describe how intensely the firms in a market compete with each other. Conduct will vary between fierce competition in a Bertrand market and the complete absence of competition in a cartel. In particular it is common to define a conduct parameter, θ, to measure how much the firms in a market compete with each other. Then, the following "conduct equation"

$$L = \frac{p - c}{p} = -\frac{1}{\varepsilon} \cdot \theta$$

nests all the equations above. Expressed differently, each equation above is a special case of this more general equation. In particular, the monopoly equation is the special case when the conduct parameter is equal to one ($\theta = 1$), the Bertrand equation is the special case when the conduct parameter is equal to zero ($\theta = 0$) and the Cournot equation is the special case when the conduct parameter is equal to the Herfindahl concentration index ($\theta = HHI$).

How can we use the conduct equation? The answer is that it may help us draw inferences about the intensity of competition in a market. Consider the following simple, but admittedly rather

unrealistic, example as a first illustration of this point. Assume that we can observe both price and cost in a duopoly market, enabling us to compute the Lerner index, and that we also know the elasticity of demand. Then, if it turns out that $L \cdot \varepsilon = -1$, we can infer that the conduct parameter is equal to one, $\theta = 1$, consistent with the idea that the two firms are engaged in a cartel. If it turns out that $L \cdot \varepsilon = -1/2$, we can infer that the conduct parameter is equal to the Herfindahl index ($\theta = HHI$) of a symmetric Cournot duopoly. Finally, if it turns out that $L \cdot \varepsilon = 0$ we may infer that the conduct parameter is equal to zero, $\theta = 0$, consistent with the duopoly firms competing a la Bertrand. That is, we may use the conduct equation as a basis for drawing inferences about the intensity of competition in a market.

In actual empirical work, the conduct parameter will be estimated simultaneously with demand and cost (more about this below). And, typically, one will find that the estimated conduct parameter does not coincide exactly with the value corresponding to one of the standard game-theoretic models. Still, it will sometimes be possible to confirm (or, rather, not reject) the hypothesis that firm conduct corresponds to one these models. Alternatively, one may investigate which model provides the best "fit" to the data in relative terms.

14.1.2. Why do we want to measure conduct?

The conduct equation can also be rewritten to show that the market price

$$p = \frac{c}{1 + \frac{\theta}{\varepsilon}}$$

is determined by three factors, namely the marginal cost of production, the consumers' price sensitivity and firm conduct. Again, conduct is short for the intensity of competition between the firms.

It is of great interest to be able to tell if a price is high due to high

production costs, low consumer price sensitivity or bad conduct among the firms (*i.e.* non-competitive behavior). Depending on the reason for a high price, it may be possible to improve the situation by various policy interventions:

- If prices are high because consumers are price insensitive, it may sometimes be possible to improve market performance. One example is the telecom market. The national regulatory authority in Swedish, the Post and Telecommunications Board (PTS), maintain a website to assist consumers to compare the price plans offered by different telecom operators. And if you find out that you could lower your costs by switching to another carrier, you are by law entitled to keep your phone number ("number portability"), substantially reducing the hustles of a switch.

- If the reason is bad conduct, it might be possible to improve the situation by fighting cartels, blocking mergers or imposing price regulations.

- If the reason is a high cost, there may be little room for public intervention in the market place.

14.1.3. Appendix: Conjectural variations

Sometimes the conduct equation is motivated in a slightly different way, in terms of so-called conjectural variations. The earliest theories of oligopoly assumed that when the firms set prices or quantities, they form expectations about how their rivals will react to their decisions. (The term "conjectural variations" means expected reactions.) One may view the game-theoretic models as special cases of such conjectural variations models. The Cournot model, for example, is the case when firms set quantities and assume that their rivals will keep their quantities fixed, no matter

what quantity they choose themselves.[2]

More formally, consider a duopoly market and let firm i's conjecture,

$$\nu_{ij} = \frac{\partial q_j}{\partial q_i}$$

take on any value, possibly different from zero. Then, if firm i's profit is given by

$$\pi_i = P\left(q_i + q_j\right) \cdot q_i - C\left(q_i\right)$$

firm i's first-order condition becomes

$$\frac{\partial \pi_i}{\partial q_i} = P\left(q_i + q_j\right) + P'\left(q_i + q_j\right) \cdot q_i \cdot \left[1 + \nu_{ij}\right] - C'\left(q_i\right) = 0.$$

If both firms have the same conjectures,$\nu_{ij} = \nu$, and the equilibrium is symmetric, *i.e.* $q_i = Q/2$ and $HHI = 1/2$, we can rewrite the first-order condition as

$$\frac{P - C'}{P} = -\frac{P'\left(Q\right) \cdot Q}{P\left(Q\right)} \cdot \frac{q_i}{Q} \cdot \left[1 + \nu\right],$$

or as the following "conjectural variations equation":

$$\frac{P - C'}{P} = -\frac{1}{\varepsilon} \cdot HHI \cdot \left[1 + \nu\right].$$

Clearly, if the firms assume their rivals quantities to be fixed, $\nu = 0$, we are back in the normal Cournot model. Another possibility is that $\nu = 1$, *i.e.* that the firms assume their rivals to change their output by the same amount. Then, the market price will be the same as the cartel price. In contrast, if $\nu = -1$, each firm assumes that any increase in its output is exactly offset by a decrease of its rivals' output. Then, the market price remains unchanged (price taking) and the outcome is the same as Bertrand.

[2]Note that the Cournot firms' conjectures are indeed true given the assumption that the firms choose their quantities simultaneously.

Today, most people find the conjectural variations models problematic. If the firms really can react to each other's decisions, these possibilities should be included in the model. Moreover, the theory should explain why the firms hold the expectations that they do. The main point here, however, is that the conjectural variations equation nests all the standard game-theoretic oligopoly models, exactly as the conduct equation above. Moreover, the conjectural variations parameter, ν, may be estimated in the same way as the conduct parameter, θ. In fact, the two parameters are simple transformations of each other. One may also test hypothesis that the conjectural variations parameter coincides with the value corresponding to one of the standard oligopoly models.

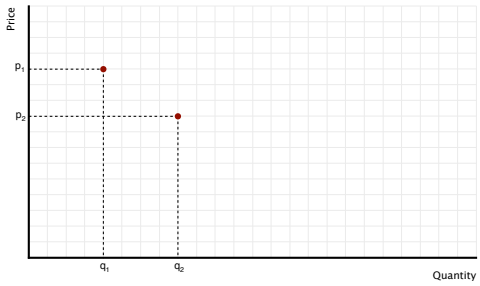
14.2. Inferring conduct from data on prices and quantities

It is of great interest to estimate how much market power firms have and also to estimate to what extent the firms' market power is due to inelastic demand and to what extent it is due to "bad conduct." The same high level of market power may be the result of normal oligopolistic competition (e.g. Cournot) in combination with a low elasticity of demand or, alternatively, a cartel operating under a demand curve with average elasticity. There are empirical techniques to estimate demand and, in particular, the price elasticity of demand. The question, therefore, is if it is also possible to separate the cost and conduct factors from each other? That is, suppose that we know the elasticity of demand. How can we tell whether a high price is due to high costs or bad conduct?

14.2.1. Identification of demand – A reminder

Assume that we wish to estimate demand in a monopoly market. Assume that we observe price and quantity during two time periods, say two quarters. Let's say that we observe a fairly high price and low quantity (p_1, q_1) in the first quarter and a fairly low price and large quantity (p_1, q_1) in the second quarter. Our two data points can then be plotted in a standard price/quantity diagram:

Figure 14.1.: Data points



The question is if we can estimate demand from these data points? The answer is no. To do so, we need to know that the two point lie on the same demand curve. Expressed differently, we need to know what caused the change in price and quantity. If we are informed that the *only* change between the two time periods is a reduction in the cost of production, then we know that the two points indeed are located on the same demand curve. The interpretation of the data is then described by the Figure 14.2.

The problem is that price may also vary due to shifts in the demand curve. If our data would include also such variations, the estimated demand function would be biased. It may for example be that the reduction in price was partially caused by a fall in demand. Then, we will not succeed to estimate the true demand curve(s), as illustrated in Figure 14.3. In particular, the estimated demand curve is less elastic than the true demand curve in this example.

Figure 14.2.: Correct estimation of demand

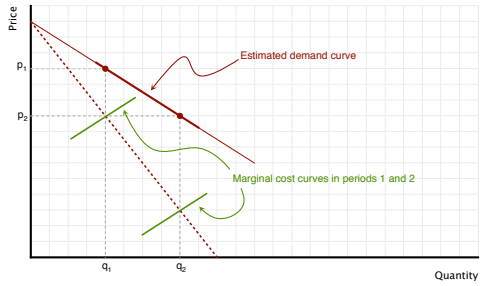
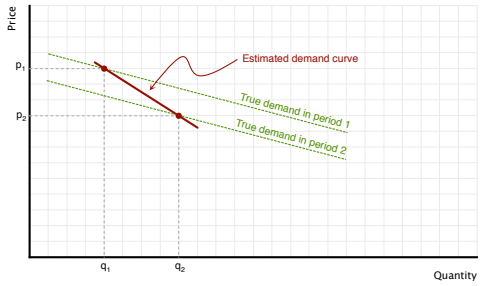


Figure 14.3.: Biased estimate of demand



As a result, we will not be able to tell why the price is as high or low as it is.

In practice, this need to know the causes of the observed variations in price and quantity implies that we need to observe changes in cost to be able to estimate demand. Normally it is not possible to observe the firms' cost functions, but it may be possible to observe (at least some) changes in the cost level. It may *e.g.* be possible to observe the price of some important input, or a change in a tax or the exchange rate. For example, if we want to estimate the demand for sugar, we may use the price of raw sugar, which clearly is an essential input in the production of refined sugar. Lets say, for simplicity, that we know that the price of sugar is changed by one Euro when the price of raw sugar is changed by one Euro. Then,

we can estimate demand by studying how the quantity of refined sugar varies (in the geographical market that we are interested in) as the (world market) price of raw sugar varies over time. Another possibility may be to observe changes in the price of the same good in another geographical market. Some of the price changes in the other market are due to changes in the production cost and some are due to demand changes in the other market. If the production costs are the same in the two markets, we can view the price in the other market as a signal of production cost in "our" market. That is by measuring the price in the other market, we observe changes in the cost of production plus an error term. The error term is simply the changes in the price in the other market that are caused by demand changes in the other market. If demands in the two markets are uncorrelated, the error term is well-behaved.

To conclude: Market demand describes how much the consumers want to buy for every possible price that the firms may charge. It is possible to estimate such a demand relationship, by observing how the quantity sold in a market varies with price. But we need to make sure that we only include price variations caused by variations in cost and not any price variations caused by variations in demand.

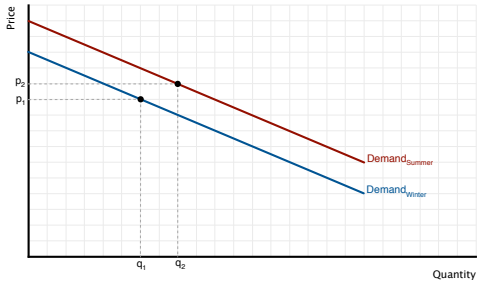
14.2.2. Identifying cost and conduct

The above section shows that we can identify the demand curve if we observe how price and quantity move together as supply (marginal cost) is shifted. Similarly, we can identify supply (marginal cost and conduct) if we observe how price and quantity move together when demand shifts. The main question is how to proceed to separate cost and conduct? How can we tell if a high price is due to high costs or "bad" conduct?

Assume that we have estimated demand and that it turns out that demand is higher during summers than during winters. Assume also that we observe two equilibrium points, one price-quantity combi-

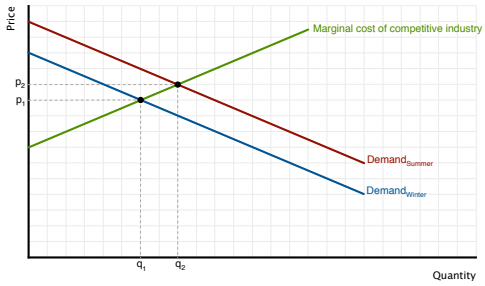
nation for the last winter (p_1, q_1) and one price quantity combination for the last summer (p_2, q_2). This is illustrated in Figure 14.4. The problem is that this data can be interpreted in several different

Figure 14.4.: Data



ways. One interpretation is that the firms are price takers and that their marginal costs are increasing, as described by Figure 14.5. Another possibility is that the firms have formed a cartel (behav-

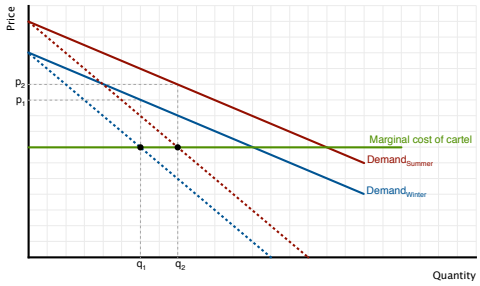
Figure 14.5.: Interpretation: competition



ing as a monopolist) and that they have constant marginal costs. This possibility is described in Figure 14.6. In addition there are intermediate interpretations that may explain the observed data, with intermediate market power and moderately rising marginal cost.

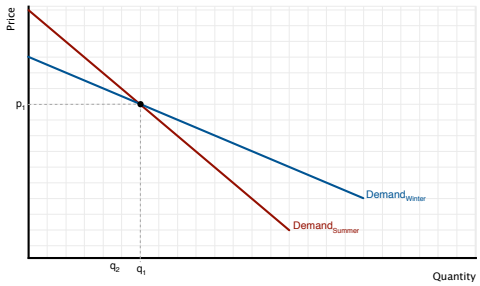
The problem to separate cost and conduct in this example arises because we have assumed that the only difference between the two

Figure 14.6.: Interpretation: cartel



demand functions is that the intercept is higher during the summer. If also the slope of demand changes, conduct can be identified.[3] The easiest way to see the logic of this idea is to assume that demand rotates around the old equilibrium point (p_1, q_1), as described by Figure 14.7. If the firms do not have any market power (they

Figure 14.7.: Rotation



are price takers) such a rotation will not cause any changes in the equilibrium. This possibility is described by the Figure 14.8. In contrast, if the firms operate a cartel, the equilibrium point moves as a result of the demand shift, as described by Figure 14.9. That is, if we can observe how the equilibrium changes when both the intercept and the slope of the demand curve changes, we will be

[3] Timothy F. Bresnahan. The Oligopoly Solution Concept is identified. *Economics Letters* 10 (1982), pp. 87-92.

Figure 14.8.: Competitive industry

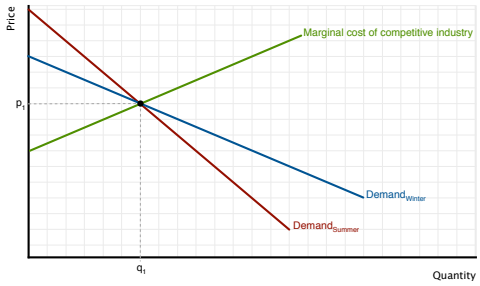
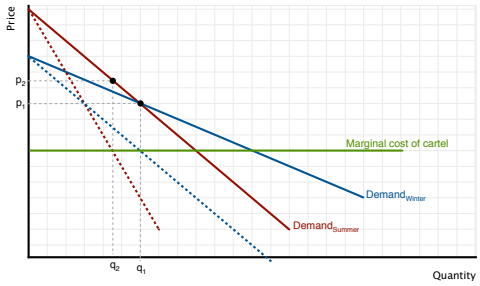


Figure 14.9.: Cartel



able to tell the difference between a competitive industry and an industry with market power. More generally, we will be able to estimate the conduct parameter ϑ and to test if it coincides with the values corresponding to the various oligopoly models.

An interesting example of this methodology is a study of the American sugar industry around 1900.

14.2.3. Example: Does extremely high market shares imply lack of competition?

By most indications, the American sugar industry should have been a very competitive industry around 1900. Refined sugar is a homogenous product (100% sucrose) and the producers shipped it in

barrels to grocers, who in turn packaged the sugar for final consumers without any identification of the manufacturer. There were no price differences. All producers also used the same, quite simple, technology buying raw sugar at the same price. Any cost differences were small. The minimum efficient scale was small in relation to market demand and there were no important entry barriers in the form of advertising and R&D.

Still the federal government was worried about the competitive conditions in the industry. One reason was that one firm, the American Sugar Refining Company (ASRC) controlled between 80 and 95 percent of the market. The government filed a lawsuit, charging monopolization and restraint of trade, seeking the dissolution of ASRC. As a result, the ASRC initiated a partial, "voluntary," dissolution. The question then arises, whether the government was right in its worries about the concentration in the American sugar market. Should an admittedly extreme market share be takes as evidence for anticompetitive conduct, despite all other indications of the opposite?

To investigate this issue, Genesove and Mullin[4] used the method outlined above. They estimated demand and separated cost and conduct using the "rotation argument." Their conclusion is that the firms may have had some market power, but that the industry was close to being competitive ($\vartheta = 0.04$). It may thus appear that the antitrust activities in this case were unnecessary.

The study by Genesove and Mullin was done many years later. How about using this kind of methods today, to inform current antitrust activities? Should competition authorities today try to undertake this type of econometric studies before taking action? The answer, of course, depends on how much faith we have in the power of these methods deliver a correct answer. How much confidence

[4]David Genesove and Wallace P. Mullin: Testing Static Oligopoly Models: Conduct and Cost in the Sugar Industry, 1890-1914. The RAND Journal of Economics, Vol. 29, No. 2. (Summer, 1998), pp. 355-377.

should we have in for example our ability to choose the functional form of demand right, or our ability to find good instruments, or in the rotation argument, and so on? Genesove's and Mullin's study also addresses this issue. Due to the simplicity of the technology and since they had detailed information about input requirements and input prices, partly from testimonies by company officials in connection to the government investigation, they can construct a measure of marginal cost separate from the measure of marginal cost that arises from the econometric study. This means that they can investigate if the econometric study succeeds in separating production cost and conduct correctly. As it turns out, they confirm that the econometric method did work well in their case. They probably underestimated market power somewhat, but this error was not sufficient to overturn their conclusions.

14.2.4. Results

Unfortunately there is no modern survey of the results in the NEIO literature. A survey by Bresnahan from 1989[5] concluded that there is a great deal of market power in some industries. The following table provides some examples of estimates of market power: The table does not present separate estimates of the elasticity of demand and firm conduct, but Bresnahan argues that an important cause of market power is that firms behave anti-competitively. Bresnahan also concludes that we do not know much about the importance of concentration since most studies are concerned with concentrated markets.

[5]Timothy F. Bresnahan. Empirical studies of industries with market power, in: R. Schmalensee & R. Willig (ed.), Handbook of Industrial Organization, edition 1, volume 2, chapter 17, pages 1011-1057 Elsevier, 1989.

Table 14.1.: Some examples of estimates of market power

INDUSTRY	L	
Food processing	0.50	
Coffee roasting	0.06/0.02	Large vs. small firms
Rubber	0.05	
Textile	0.07	
Electrical machinery	0.20	
Tobacco	0.65	
Railroads	0.40	
Retail gasoline	0.10	
Automobiles	0.10/0.34	Standard vs. luxury
Aluminum	0.59	
Banks, before entry deregulation	0.88/0.21	Large vs. small firms
Banks, after entry deregulation	0.40/0.16	Large vs. small firms

14.3. Inferring conduct from data on concentration and market size

Sometimes it is not only impossible to observe firms' costs and profits, but often also data on prices and quantities may be out of reach for the researcher. The question is if we can, nevertheless, draw some inferences about firm conduct?

Here, we will focus on one issue only, namely the most fundamental issue in industrial economics, the relation between market concentration and conduct. Do firms compete more, the less concentrated a market is? All the theoretical models of oligopoly such as Cournot, Bertrand and Hoteling, as well as repeated game models of cartels, suggest that they do. Can we test this prediction empirically? Can we quantify the strength of this relationship?

It seems obvious that we cannot say anything about how much market power firms have in an industry, and how market power is affected by concentration, if we do not have any data on prices, quantities and costs. But, in fact, we can.[6]

[6]Timothy F. Bresnahan and Peter C. Reiss. Entry and Competition in Con-

14.3.1. Basic idea

Theoretical models of entry suggest that there should be more firms the larger a market is. We should, for example, observe more pharmacies in large towns than in small towns. It is not clear, however, that there will be twice as many pharmacies in a town with twice as many inhabitants. It may take (say) 10 000 people before a first (monopoly) pharmacy enters a town and 30 000 people before a second pharmacy enters. The reason is that the entry of a second pharmacy in the same geographical market implies that the two firms have to compete with each other and, therefore, that they earn lower profits per inhabitant. As a result each duopoly pharmacist needs to sell to more customers, here 15 000 people, than a monopoly pharmacist in order to cover the fixed costs. That is to say, the number of firms in a market does not only depend on the size of that market, but also on the intensity of competition.

Entry theory thus suggests that market concentration is determined by market size and the intensity of competition. Since market concentration and market size may be more readily observed than prices and quantities, such data may be used to infer how strong competition is. In particular, by comparing the number of customers per firm in towns with different levels of concentration, we get an idea about how much more competitive the market becomes as concentration is reduced. At the extreme, if all firms join a cartel independent of concentration, firms will have the same number of customers independent of concentration. Therefore, according to this logic, if we observe that the number of firms is roughly proportional to the number of inhabitants when we compare different towns, we should suspect cartelization. At the other extreme, if firms compete a la Bertrand with very little differentiation, a second firm would only enter in extremely large cities,

centrated Markets, Journal of Political Economy, vol. 99(5) pp. 977-1009, October 1991.

implying that the number of customers per firm is much larger in duopolies than in monopolies. Therefore, according to this logic, if we observe monopolies in all towns, except for the very largest, we should assume that competition is fierce (whenever it exists). This logic can be made a little more precise using a simple entry model.

14.3.2. What entry thresholds say

Consider a market for some product ("widgets"), which is segmented into several isolated geographical markets ("towns"). People in one town cannot shop for widgets in other towns as they are too far away. Concentration differs between towns: in some towns there are no firms, while in others there are one or two widget firms. There are typically more firms in larger towns and it is often assumed that, as a result, prices are lower in larger towns. Nobody knows for sure, however. Assume that we can observe the number of widget firms as well as the number of inhabitants in every town, but nothing more. How should we organize this data to say something about competition in the widget market? The answer is that we should define entry thresholds.

Assume that consumer tastes are the same in small and large towns, implying that the demand for widgets is twice as large in a twice as large town. Let p be the price, $d(p)$ the demand of a representative consumer and S the number of consumers. Then the market demand for widgets is given by $d(p) \cdot S$ in a town with S inhabitants.[7] For simplicity we also assume that there are constant returns to scale, represented by a fixed unit cost c. The cost of entry is denoted F.

Monopoly entry threshold In order for a first firm to enter the widget market in a certain town, the population must be sufficiently

[7] In a more general model one may account for differences in demand that depend on observable factors such as population demographics such as age and sex as well as income and education.

large for a monopolist to break even and cover the entry cost. In particular, let p_1 be the optimal price that a monopoly firm would charge. (It is easy to see that the monopoly price is independent of market size, given the assumptions of the model.) Let S_1 be the number of inhabitants required for the monopoly firm to break even. We will refer to this number as the monopoly entry-threshold. This entry threshold is given by the zero-profit condition:

$$\pi_1 = [p_1 - c] \cdot d(p_1) \cdot S_1 - F = 0.$$

Solving for the entry threshold,

$$S_1 = \frac{F}{[p_1 - c] \cdot d(p_1)}$$

shows that the population needs to be larger if the entry cost, F, is higher or the profit per inhabitant, $[p_1 - c] \cdot d(p_1)$, is smaller.

Duopoly entry threshold In order for a second firm to enter a town, it is required that

$$\pi_2 = [p_2 - c] \cdot d(p_2) \cdot \frac{S_2}{2} - F = 0$$

where p_2 is the duopoly price in equilibrium. Expressed differently, S_2 is the minimum number of inhabitants required for the zero-profit condition to be satisfied—the duopoly entry threshold. For simplicity it is assumed that firms produce homogenous goods implying that they share the consumers equally. Thus, $S_2/2$ is the number of customers per firm in a duopoly market where the firms break even. Solving for the duopoly entry threshold,

$$\frac{S_2}{2} = \frac{F}{[p_2 - c] \cdot d(p_2)}$$

shows that it is determined by the same variables as the monopoly threshold.

Entry threshold ratio Notice that the difference in the (per firm) entry thresholds is explained by the difference between the monopoly price, p_1, and the equilibrium duopoly price, p_2. The difference between the monopoly and duopoly prices, in turn, is simply the strength of duopolistic competition.

Now, let's form the ratio between the *per-firm* entry thresholds. It is given by:

$$\frac{S_2/2}{S_1} = \frac{[p_1 - c] \cdot d(p_1)}{[p_2 - c] \cdot d(p_2)}.$$

Notice that this statistic, $(S_2/2)/S_1$, measures how profit per customer changes with concentration. And, within a certain product market, the profit per customer is itself an indicator of conduct.

To be concrete: If entry does not change conduct, $p_2 = p_1$. Then, $[p_1 - c] \cdot d(p_1) = [p_2 - c] \cdot d(p_2)$ and hence

$$\frac{S_2/2}{S_1} = 1.$$

But, if entry leads to competition, $p_2 < p_1$. Then the profit per customer must be smaller, *i.e.* $[p_2 - c] \cdot d(p_2) < [p_1 - c] \cdot d(p_1)$ and hence

$$\frac{S_2/2}{S_1} > 1.$$

If we know the entry thresholds, we can deduce how intense competition is.

A numerical example Assume that we observe that a first firm enters towns of 10 000 inhabitants and that a second firm enters towns of 30 000 inhabitants, then clearly a duopoly firm needs 50% more customers than a monopoly firm to break even, i.e.

$$\frac{S_2/2}{S_1} = 1.5$$

which means that the profit per customer is 50% larger in monopoly markets than in duopoly markets. Reversely, the earnings per cus-

tomer are only about 70% in duopoly compared to monopoly:

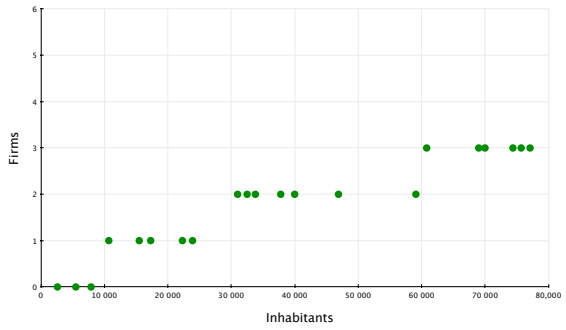
$$\frac{[p_2 - c] \cdot d(p_2)}{[p_1 - c] \cdot d(p_1)} = \frac{2}{3} \approx 0.7.$$

If we are willing to assume that consumer preferences are the same in small and large towns and that there are constant returns to scale, we may conclude that the margins must be much lower in duopoly markets than in monopoly markets.

14.3.3. How to find the entry thresholds

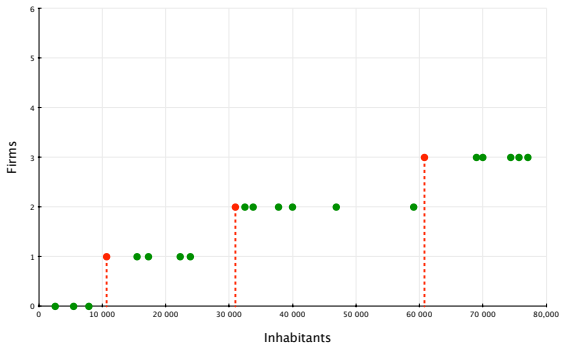
Assume that we have collected the hypothetical data on the number of firms and the number of inhabitants in 21 towns, as described in Figure 14.10. Then, with this highly orderly data, it is straightfor-

Figure 14.10.: Data



ward to find the entry thresholds. Figure 14.11 describes how. It thus appears that a first widget firm enters towns with at least 10 000 inhabitants and that a second firm enters in towns with at least 30 000 inhabitants. As already argued this means that each firm needs 50% more customers to survive in duopoly than in monopoly.

Figure 14.11.: Data



14.3.4. Entry thresholds - How to find them with "fuzzy data"

The hypothetical data set above displays a simple monotone relation between market size and the number of firms: there are more firms in larger towns. In reality, however, the relation is much more blurred. Sometimes smaller towns have more firms than larger towns. Then, estimating entry thresholds is more complicated.

To account for the fact that towns of the same size may have different number of firms we need to acknowledge other differences between towns than size. For instance, the entry costs may differ across towns. Clearly, a town with high entry costs will have fewer firms than a town of the same size but with lower entry costs. Other important differences may be the demographic structure, eg the share of elderly, and the per capita income. All these factors influence demand and therefore the profitability of entering the market.

The question here is how to work with a data set where the number of firms varies between towns of the same size, by interpreting these differences in terms of entry cost differences.

A simple model To show how this can be done it is convenient to write a firm's profit (excluding entry costs) as

$$\Pi = \pi(n) \cdot m,$$

where n is the number of firms and m is the number of inhabitants in the town. Moreover, $\pi(n)$ is the (per firm) profit per inhabitant in a town with n competitors.

Let us look at some examples. Assume that (inverse) market demand is linear and given by $p = \alpha - Q$ and that all firms have the same constant marginal cost c. Then a monopoly firm would earn profit

$$\Pi = \left(\frac{\alpha - c}{2}\right)^2 \cdot m.$$

Expressed differently

$$\pi(1) = \left(\frac{\alpha - c}{2}\right)^2.$$

If entry doesn't trigger any price competition, i.e. if the firms succeed to enforce a cartel, independent of the number of rivals, the firms will simply share the monopoly profit, i.e.

$$\pi^{Cartel}(n) = \pi(1) \cdot \frac{1}{n}.$$

On the other extreme, if entry leads to Bertrand competition, all firms earn zero profit. Thus,

$$\pi^{Bertrand}(n) = 0$$

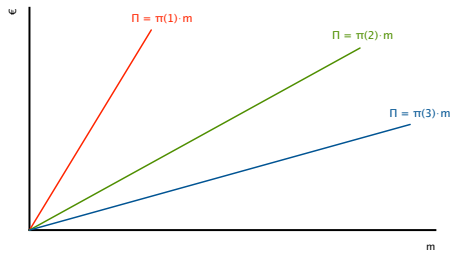
whenever $n \geq 2$. In the intermediate case, if entry leads to Cournot competition,

$$\pi^{Cournot}(n) = \left(\frac{\alpha - c}{n+1}\right)^2.$$

The following figure shows how a (say) Cournot firm's equilibrium

profit increases with the size of the market for three different levels of concentration, namely monopoly, duopoly and triopoly. Clearly,

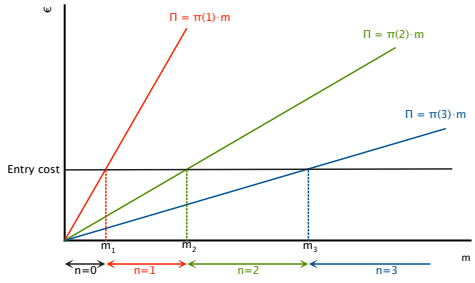
Figure 14.12.: How profit depends on the size of the market, for different levels of concentration.



profits increase faster in market size the fewer firms there are. The reason is of course that the profit per customer is higher the less competition there is.

Uniform entry cost In the next figure we also display the entry cost, which for the moment is assumed to be the same in all markets. Given the assumption that the entry cost is the same in all

Figure 14.13.: Entry thresholds

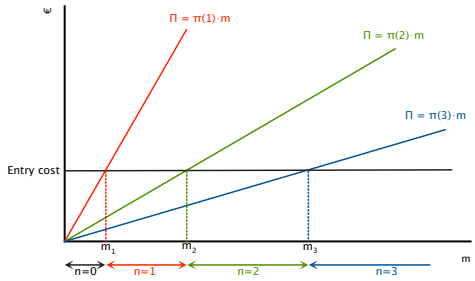


markets, we can compute the entry thresholds. Thus, m_1 indicates the number of inhabitants necessary for a monopoly firm to enter a market, m_2 is the entry threshold for duopoly, and so on. The

arrows thus indicate the range of market sizes for which we will observe no, one, two or three firms.

The next figure illustrates how the entry threshold for duopoly varies with conduct. If the widget market is prone to cartelization,

Figure 14.14.: How duopoly entry threshold varies with conduct.



the duopoly-profit-line is close to the monopoly-profit-line. In this case m_2 is close to m_1. Observing m_2 close to m_1, we would also conclude that duopoly firms do not compete much.

If there is Cournot competition in widget markets, the duopoly-profit-line is further away from the monopoly-profit-line. In this case m_2 is much higher than m_1. Observing that m_2 is much higher than m_1, we would also conclude that duopoly firms do compete with each other.

If there is Bertrand competition in widget markets, the duopoly-profit-line coincides with the x-axis. No market size is sufficient to support two firms.

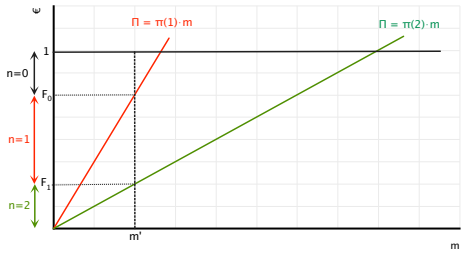
Allowing for product differentiation and other ways of competition, the general insight is that more competition means that the duopoly-profit line is flatter, further away from the monopoly-profit-line. Moreover, if we can observe the duopoly entry-threshold, we may infer how far away the duopoly-profit-line is and, thus, how much competition there is in the widget market.

This is the same logic as before.

Varying entry costs Assume now that we observe different number of firms in markets of the same size. How can we then infer conduct from data on market size and concentration? As usual, we will here consider the simplest possible model to make sense of the data. Let's say that we know that entry costs vary between different towns. All firms pay the same entry cost in the same town, but it may be more expensive to enter certain towns than certain other towns. For simplicity, let us assume that entry costs vary between 0 and 1 and that they are uniformly distributed. This means for example that 50% of all towns have an entry cost larger than a half.

Also, for simplicity, let us focus on towns that all have the same size m', as displayed in the following figure. Note that there will

Figure 14.15.



be no entry in any town of size m' if the entry cost in that town is larger than F_0. At least one firm will enter if the entry cost is lower than this threshold. And there will be two firms if the entry cost is below F_1. That is,

- For all towns of size m' with entry costs in the interval $(F_0, 1]$, there will be no entry, *i.e.* $n = 0$.
- For all towns of size m' with entry costs in the interval $(F_1, F_0]$, one firm will enter the market, *i.e.* there will be monopoly, $n = 1$.

- For all towns of size m' with entry costs in the interval $[0, F_1]$, two firms will enter the market, *i.e.* there all be duopoly, $n = 2$.

Expressed differently, the probability that a town of size m' has no firm is $1-F_0$ since the entry costs are uniformly distributed between zero and one. The probability that a town of size m' has one firm is $F_0 - F_1$ and the probability that it has two firms is $F_1 - 0$.

Inference Assume now that we have observed the number of widget firms in a very large number of towns and that they all have m' inhabitants. Then, we are able to infer the probabilities of having no, one or two firms in towns of this size. For example, if we observe that there are no widget firms in a quarter of all towns, we can infer that $F_0 = \frac{3}{4}$. If, in addition, half of the towns have one firm, we can infer that $F_1 = \frac{1}{4}$.

Having inferred the entry-cost-thresholds F_0 and F_1 we are able to locate the profit lines in the figure above. We know that the red line has to go through the point (m', F_0) and that the green line has to go through the point (m', F_1). That is, we are able to infer how the duopoly profit relates to the monopoly profit and therefore how much competition there is when a second firm enters the market.

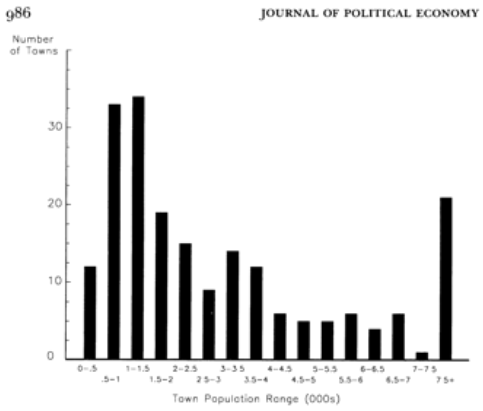
The intuition is actually quite simple. If we observe that there is at least one firm in most towns of size m' we may infer that the widget market is profitable, *e.g.* because consumers like widgets or because production costs are low. If in addition many of the towns with at least one firm also have a second firm, we may infer that the second firm doesn't reduce profits by much, *i.e.* that widget firms do not compete much. On the other hand, if only a small proportion of towns with at least one firm also has a second firm, duopoly competition must be quite intense. We may thus infer the intensity of duopolistic competition from the proportion of duopoly towns relative to the proportion of monopoly towns.

14.3.5. Example: How much competition is there in the service sector?

Bresnahan and Reiss employed the entry model to estimate how much competition there is in a set of narrowly defined service markets, namely doctors, dentists, druggists, plumbers and tire dealers. They were able to identify all firms in these markets, simply by using the telephone directory.

Data The study included 202 towns (county seats in the western part of the US). Only isolated towns were included, meaning that there were no other town nearby.The sample included a wide range of market sizes as described by the following figure. It is ev-

Figure 14.16.



ident that there are more firms in larger towns. The figures below show this for dentists. But it is also clear that there is quite some variation, some larger towns have fewer dentists than some smaller towns.

Entry thresholds Bresnahan and Reiss estimated the following entry thresholds: That is, for a first doctor to open a practice, a town has to have about 900 inhabitants. For a second doctor to

Figure 14.17.

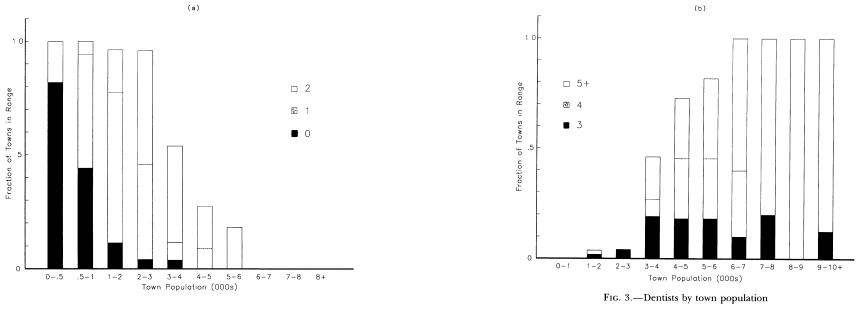


FIG. 3.—Dentists by town population

(a) (b)

Table 14.2.: Entry thresholds (000's)

Profession	S_1	S_2	S_3	S_4	S_5
Doctors	0.88	3.49	5.78	7.72	9.14
Dentists	0.71	2.54	4.18	5.43	6.41
Druggists	0.53	2.12	5.04	7.67	9.39
Plumbers	1.43	3.02	4.53	6.20	7.47
Tire dealers	0.49	1.78	3.41	4.74	6.10

enter, the town has to have about 3 500 inhabitants. Tire dealers require fewer consumers to enter a town.

The next table displays the entry thresholds per firm. This table

Table 14.3.: Entry thresholds per firm (000's)

Profession	s_1	s_2	s_3	s_4	s_5
Doctors	0.88	1.74	1.93	1.93	1.83
Dentists	0.71	1.27	1.39	1.36	1.28
Druggists	0.53	1.06	1.68	1.92	1.88
Plumbers	1.43	1.51	1.51	1.55	1.49
Tire dealers	0.49	0.89	1.14	1.18	1.22

reveals that in order for two doctors to survive, they need to each have 1 740 potential patients, substantially more than a monopoly-doctor requires (880 potential patients).

To facilitate the comparisons, the next table displays the entry

threshold ratios. Those in red suggest a substantial competitive effect. This table reveals that the entry of a second firm creates

Table 14.4.: Entry thresholds ratios

Profession	s_2/s_1	s_3/s_2	s_4/s_3	s_5/s_4
Doctors	1.98	1.10	1.00	.95
Dentists	1.78	.79	.97	.94
Druggists	1.99	1.58	1.14	.98
Plumbers	1.06	1.00	1.02	.96
Tire dealers	1.81	1.28	1.04	1.03

substantial competition in all services except plumbing. Duopoly firms need almost twice a many customers as monopoly firms to break even. However, a third firm doesn't create much additional competition, except for druggists and perhaps druggists.

Conclusion Bresnahan and Reiss study this suggests that competition is important, at least up to a point. If this would be a general pattern, public policies should primarily be focused on avoiding monopoly. Protecting or promoting competition in less concentrated markets would be less important.

Part V.

Non-Price Competition

15. Product choice

15.1. Monopoly

Swedes often complain about train prices. When Ryan Air can fly us to southern Europe for 50 Euros, how can a train ticket between Stockholm and Gothenburg cost 100 Euros? But during winters, people complain even more about all the delays. During the winter 2010 a third of all trains where late as a result of snow and ice. In the public debate the delays are often interpreted as incompetence on the part of the SJ management. In a country this far north, shouldn't they have learned about snowy winters and planned accordingly? But, maybe it is not a question of incompetence. Maybe SJ offers us the quality maximizing their profits. Could it be that a profit-maximizing monopolist will offer a quality that is lower than the quality their consumers would be willing to pay for?

The basic quality trade-off for a monopolist like SJ is that higher quality increases the consumers' willingness to pay for the good but that the good will also be more costly to produce. More trains will for example be in time if the company would maintain a larger reserve of train sets to fall back on. To maximize its profits, the monopolist must balance the cost increases against the consumers' increased willingness to pay.

A problem is that different consumers value quality differently. The natural solution would then be to design the quality which the "average consumer" in mind. Expressed, differently, it seems natural to consider the average increase in the consumers' willingness

to pay. But this is (unfortunately) not what firms do.

15.1.1. The marginal consumer rules

A monopoly firm designs its product with the so-called marginal consumer in mind. The marginal consumer is the consumer who is indifferent between buying and not buying the good. The reason why the marginal consumer is decisive is that the whole purpose of increasing quality is to increase the price of the good and it is the fear of losing the marginal consumer that prevents a firm from raising its price. A simple example can illustrate this point.

Consider a good sold to two types of consumers, the marginal and the infra-marginal consumers. The infra-marginal consumers buy the good with a positive consumer surplus at the current price and given the current quality. Lets say that the infra-marginal consumers value the good at €100 and that the price is €50, leaving them with a surplus of €50. By definition, the marginal consumers are indifferent between buying and not buying the good at the current price and given the current quality. Their valuation of the good at the current quality is thus €50, leaving them with no consumer surplus. Assume now that the monopolist is contemplating improving the quality of the good. Producing the higher quality would increase unit cost by C. Producing the higher quality would also increase the marginal consumer's valuation by W_M and the infra-marginal consumers' valuation by W_I. The question is whether or not the monopolist will supply the higher quality?

Assuming that the monopolist wants to continue to supply both groups, the quality improvement will enable the monopolist to increase the price of the good by W_M. The marginal consumers will then, by definition, continue to buy the good and receive no consumer surplus. Also the infra-marginal consumers will continue to buy the good and earn consumer surplus $100-50+W_I-W_M$ which is positive as long as $W_M - W_I < 50$, which is assumed. The mo-

nopolist will then improve the quality if and only if the unit cost is increased by less than price, i.e. if and only if $C \leq W_M$. The infra-marginal consumers' valuation of quality does not matter for the monopolist's decision. It doesn't matter if $W_I > C$ or if $W_I < C$. The monopolist's decision is solely based on the marginal consumers' valuation of the improvement (as long as $W_M - W_I < 50$).

15.1.2. Social welfare

From a social welfare point of view, the average consumer should be decisive. A quality improvement increases the total surplus in the market if the sum of valuations exceed the sum of costs, $W_I + W_M \geq 2 \cdot C$. Expressed differently a quality improvement increases total welfare if the average valuation of the quality improvement exceeds the cost:

$$\frac{W_I + W_M}{2} \geq C.$$

We should thus typically expect monopoly firms to choose an inefficient level of quality. It seems unlikely that the marginal consumer has the same willingness to pay for quality improvements as the average willingness to pay among all consumers.

Under-provision In the example above, the monopolist does not provide the higher quality if $C > W_M$. This decision reduces social welfare if $C < \frac{W_I + W_M}{2}$. In this case, we say that there is under-provision of quality. Under-provision can only happen if the marginal consumers have a lower valuation of quality than the infra-marginal consumers. To see this note that both inequalities can be fulfilled only if $W_M < \frac{W_I + W_M}{2}$, i.e. only if $W_M < W_I$.

Is under-provision likely to be a common problem? To discuss this issue, we need to make some additional assumptions. Let us assume that the only reason why people have different willingness to pay for the good is that they have different incomes. There are no taste differences. Let us also assume that the basic good is a so-called

normal good, which means that people with a higher income have a higher willingness to pay for the good. Then, the infra-marginal consumers are people with high income and the marginal consumers are people with low income. If quality is also a normal good, then, by definition, people with low income have a lower willingness to pay for quality increases. When the good is normal, this means that the marginal consumers have a lower willingness to pay for quality than the infra-marginal consumers. Thus, a monopolist may provide a too low quality of a normal good, if quality is a normal good.

A crucial assumption is, of course, that the firm can only provide one variant of the good. Otherwise, an even better strategy would be to provide one variant for every consumer type.

A possible example is the train operator SJ. Some of their customers are business travelers or people commuting to work. They probably have a high willingness to pay for traveling to their destination (the basic service) and they probably have a high valuation of arriving in time (quality). Other customers are retired or students traveling to meet their families. They probably have a lower willingness to pay for both the trip and for arriving according to schedule. Our analysis then suggests that SJ will aim for the punctuality that retired people and students are willing to pay for, disregarding the business travelers and commuters' higher willingness to pay for arriving at their destinations in time. SJ is likely to be less punctual than it should.

As most goods are probably normal and most qualities are probably normal too, we may expect under-provision to be a typical phenomenon in monopolistic markets.

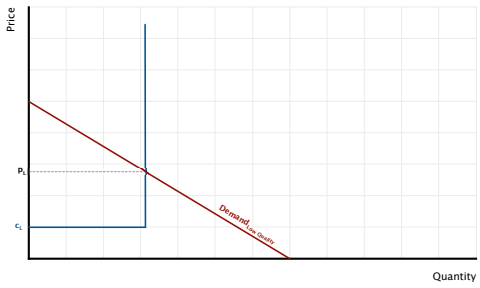
Over-provision In the example above, the monopolist provides the higher quality if $C \leq W_M$. This decision reduces social welfare if $C > \frac{W_I+W_M}{2}$. This would be an example of over-provision of quality. Over-provision can only happen if the marginal consumers have a

higher valuation of quality than the infra-marginal consumers. To see this note that both inequalities can be fulfilled only if $W_M > \frac{W_I+W_M}{2}$, i.e. only if $W_M > W_I$. It is much more difficult to come up with likely examples of over-provision. If the basic good is normal, quality must be a so-called inferior good to make the monopolist produce a higher quality than is socially optimal. An example could be speakers included in stationary computers. Poor people may be willing to pay more for these speakers, as rich people may already have bought really good external speakers. Please let me know if come to think about any better examples.

15.1.3. An example with more consumer types

Consider a market for a normal good and assume that people only differ in income. Every consumer either buys one or no unit of the good. If peoples willingness to pay for this a unit of the good is uniformly distributed, demand is linear. To simplify the analysis we assume that the monopolist has a relatively low marginal cost up to some capacity constraint. (The capacity constraint ensures that the marginal consumer stays the same throughout the analysis.) This situation may also be described in Figure 15.1.

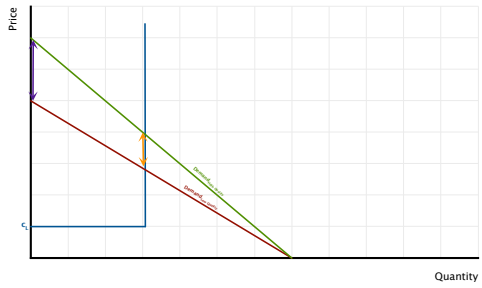
Figure 15.1.



Improved quality increases the consumers' willingness to pay for the good. Since we are considering unit demand for a normal good,

the rich consumers are located to the left on the demand curve, close to the y-axis. With quality a normal good, the rich consumers' willingness to pay increases more than the poor consumers' willingness to pay. In Figure 15.2, the increase in the richest consumers' will-

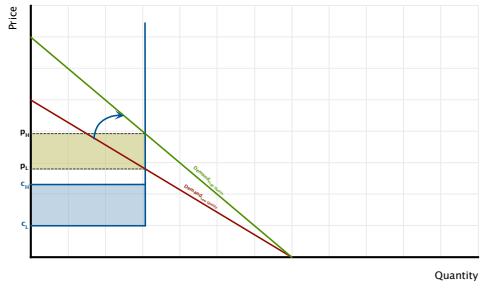
Figure 15.2.



ingness to pay is described by the purple arrow and the increase in the poorest consumers' willingness to pay is described by the orange arrow.

The monopolist's gain from improving quality is that it can increase the price of the good. The revenue increase is described by the green rectangle in Figure 15.3. The disadvantage is the cost

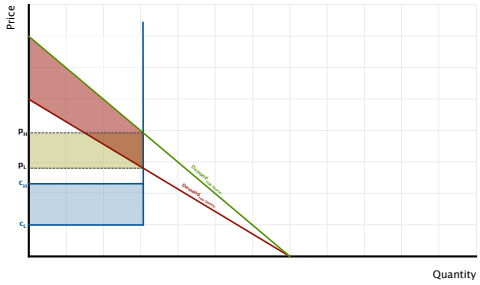
Figure 15.3.



increase, described by blue. The social value of improved quality is the total willingness to pay for the improvement as described by the

difference in the willingness to pay for all consumers. This amount is described by the red area in Figure 15.4. Clearly the social value

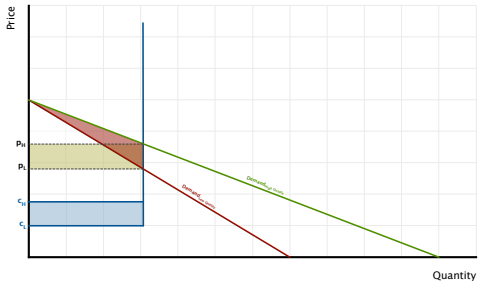
Figure 15.4.



of the quality improvement is larger than the private value of the quality improvement to the firm. And if the cost of providing the quality improvement happens to be in-between the social and the private value, too little quality will be provided.

Consider now a quality improvement that itself is an inferior "good," meaning that rich people are willing to pay less for the quality improvement than poor people. This situation may is described in Figure 15.5. In this case, it is the marginal consumers'

Figure 15.5.

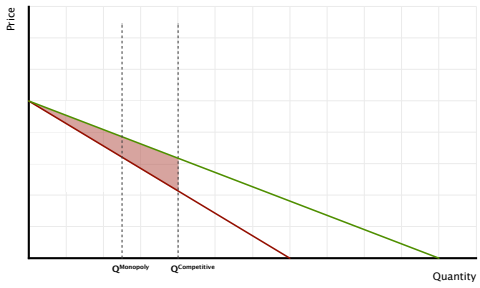


willingness to pay for the improvement that is the highest. The monopolist's gain is the possibility to increase price and earn higher

revenues (the green rectangle). The disadvantage is the increase in production cost (the blue rectangle). The social value of the improvement is described by the red area. In this case the social value is lower than the firm's private value, implying that there may be too much quality in equilibrium.

A caveat We have considered normal goods and assumed that the main difference between consumers is differences in income. The main conclusion is that a private firm will provide too little of qualities that are themselves normal goods and too much of qualities that are themselves inferior goods. In this analysis we have only questioned the firm's quality choice, but not the monopolist's choice of price and quantity. However, in markets with imperfect competition, the socially optimal quantity is larger than the actual quantity. As a result, the marginal consumer under monopoly may actually be close to the average consumer under competition, as described by in Figure 15.6. Hence, the quality chosen by the monopolist may actually be the same as the quality chose by a social planner who could also implement the competitive quantity. Still, assuming that we have to accept the monopoly output, there will be under- or over-provision as described above.

Figure 15.6.



15.1.4. Conclusion

Monopoly firms always charge too high prices. But when it comes to quality, the picture is more complex. A monopolist may set its quality both too high and too low, depending on the circumstances. There may be both over-provision and under-provision of quality. For example, if the main reason for differences in willingness to pay is differences in income, then monopolists may provide too low quality of a normal goods, if quality itself is a normal good. As most goods are probably normal and most qualities are probably normal too, we may expect under-provision to be a typical phenomenon in monopolistic markets.

15.2. Oligopoly & Horizontal differentiation

Listen to this TV-commercial from the 70:ies:

> "*Every once in a while your soft drink routine needs a change. We recommend adding, at regular intervals, a 7UP - the un-cola. A clean, crisp change from the every day: Caffein-free 7UP - the un-cola.*"

The point of this commercial is to set 7UP apart from the competition: 7UP isn't necessarily better than Coke and Pepsi, but it is certainly different. (The commercial is available on YouTube. Search for "uncola".)

But differentiation is not a universal principle in the market place. In some markets, the firms appear to cluster around the same point in "product space." Why do all the commercial radio channels in Sweden play the same tunes from the national hit list? Why are all the new pharmacies, that entered the Swedish market following deregulation, opening stores in the city centers and malls where there already are so many stores?

Product choice is intertwined with *market segmentation*, which arises when firms divide markets into more or less homogenous groups of consumers and pitch different offers to different segments. Another related idea is *product positioning* which starts with product choice, but is more focused on positioning a product in the minds of the consumers. A classical example is when Volkswagen introduced the Beatle in the US with its "Think small" campaign. Although there already existed small cars before, no other firm had previously made a point of it and occupied that position in the minds of the consumers.

15.2.1. Hoteling model

Consider a market with two retailers. Each retailer wishes to locate a convenience store to serve consumers who live evenly spread along a street. These location choices give rise to a trade-off. On the one hand, each firms wants to locate its store in the middle, to be as close as possible to as many consumers as possible, in order to maximize the consumers' willingness to pay. Or, rather, each firm wants to locate its store in such a way that it is closer for more people than the rival's store. A better location gives a competitive advantage. On the other hand, if both firms choose the same position, their stores would become perfect substitutes, which would lead to ruinous price competition. That is, location determines the degree of differentiation and the intensity of price competition. The question is how the firms balance these opposing forces?

To analyze this question we will use Hoteling's model of horizontal product differentiation (presented in Chapter 11.1.1). The market consists of a single street of length 1. Two stores sell identical goods, but at different locations. The consumers are uniformly distributed along the entire street. All have unit demand and are willing to pay v for one unit of the good.

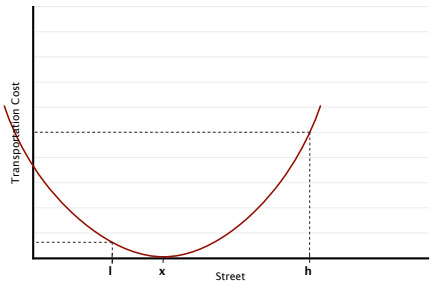
There are two differences compared to the Hoteling model used

in the previous chapter. First, we now need to assume that the firms are not necessarily located at the end points of the street, but may have their stores somewhere between the end-points. One is located at l and the other located at $h > l$. Second, we assume

Figure 15.7.

that the consumers have quadratic transportation costs. Thus, a consumer living at x visiting the store located at l would have to pay the transportation cost $(x-l)^2$. The same consumer would have to pay $(h-x)^2$ if visiting the store at h. Thus, the consumer at x in Figure 15.8 prefers store at l equal prices. Quadratic trans-

Figure 15.8.



portation costs implies that the extra cost of traveling one extra meter is increasing in the distance already travelled and may well be a realistic assumption in many cases. Here, it is primarily made

for technical reasons, however.[1]

The timing of the game is that first the firms simultaneously choose their locations, second they simultaneously set their prices and finally the consumers decide whether they want to buy the good and, if so, from which firm. We solve the model backwards.

[1]The reason for using a quadratic cost function is that the linear cost function gives rise to some technical problems when the firms are not located at the end- points of the street. If a firm would lower its price to the point where it attracts all the consumers located between the two firms, it would also attract all the consumers on the other side of the rival. As a result, the firms demand and profit functions would be discontinuous and under some conditions an equilibrium in pure strategies fail to exist.

A methodological note on timing in game-theoretic models

When modeling product choice, it is common to assume that the firms first (perhaps simultaneously) choose their product positions and then (perhaps simultaneously) choose their prices. Assuming product choice to precede pricing allows us to capture the fact that firms adjust their prices quicker than they adjust their product characteristics. If, for example, firms revise their product design every decade and their prices every quarter, then they will take the product design as given when they set their prices. And, when designing their products, they will consider the effect of the design on future prices. This is exactly what we get in the model with the suggested timing. Thus, the present exercise also conveys a more general game-theoretic insight: when analyzing situation in which people make different types of choices (such as product positioning and pricing) but with different frequencies (e.g. by decade and quarter), the model should let the players make the infrequent decisions first and the frequent decisions later. An alternative is, of course, to set up a more complicated model where all these decisions are repeated with different frequencies. But, fortunately, that is often not necessary.

Stage 3: Demand A consumer living at x and shopping at store l receives utility $u_l(x) = v - p_l - t \cdot (l - x)^2$ where p_l is the price in the store. Buying from store h gives the same consumer utility $u_h(x) = v - p_h - t \cdot (x - h)^2$ where p_h is the price in store h.

Before reading any further, please solve the following Exercise.

Exercise 52. Show that the demand functions for the two stores

are given by:

$$D_l\left(p_l, p_h l, h\right) = \frac{h+l}{2} + \frac{p_h - p_l}{2 \cdot t \cdot (h-l)}$$

and

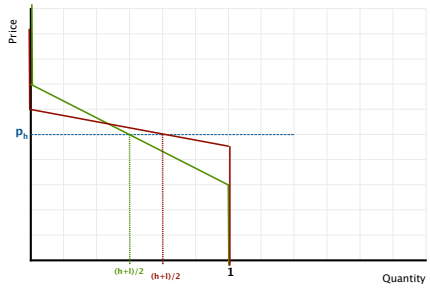
$$D_h\left(p_l, p_h l, h\right) = 1 - \frac{h+l}{2} + \frac{p_l - p_h}{2 \cdot t \cdot (h-l)}.$$

Hint: Start out by finding the location of the consumer who is indifferent between the two stores. Next, think about which consumers prefer which store. A solution is available below.

The demand functions reveal that each firm will sell less the higher price it sets, and that it will sell more the higher the price of the competitor is. Note in particular that the price difference $p_h - p_l$ has a larger impact on demand the closer the two stores are located to each other, i.e. the smaller $h - l$ is. Expressed differently, the closer the stores are, the less differentiated they are, and the more elastic the firms' residual demand functions will be. This feature of demand is absolutely crucial for the coming analysis.

Also note that, keeping prices fixed, a firm's choice of location has two effects on demand, one effect on competitive advantage and one on product differentiation. Take the left store as an example. If the firm would move the left store somewhat to the right and closer to the competitor, the left store would become the closest store to more consumers, which is a benefit. A move to the right would create a competitive advantage (or reduce a competitive disadvantage) for the left firm. At equal prices, the left firm would increase it sales. But, on the other hand, by moving the left store to the right, the difference between the stores would become smaller, making demand more elastic. The move shifts the demand curve from green to red in Figure 15.9. Taking prices as given, a move to the right increases the left stores sales, unless its price is considerably higher than the rival's price. To fully understand the consequences of a relocation of a store, we first need to understand how it affects price

Figure 15.9.



competition.

Stage 2: Price competition In the second stage the two firms simultaneously set their prices, taking into account how prices will affect sales, as described by the demand functions, and taking as given the location of the two stores, chosen in the first stage.

Before reading any further, please solve the following exercise.

Exercise 53. Show that the equilibrium prices are given by

$$P_l(l,h) = c + \frac{2}{3} \cdot t \cdot (h-l) \cdot (1 + \frac{h+l}{2})$$

and

$$P_h(l,h) = c + \frac{2}{3} \cdot t \cdot (h-l) \cdot (2 - \frac{h+l}{2}).$$

Hint: First, write down the two firm's profit functions. Second, find the firms' first-order conditions. Third, solve the two (linear) equations in the two unknowns (*i.e.* the prices).

A solution is available below.

Note that both firms' prices depend on the location of both stores. Again, there are two reasons. The first reason is product differentiation, as captured by $h-l$. If the two stores would be located at the same point, $h-l=0$, they would be perfect substitutes and prices would be competed down to marginal cost, $P_l(l,h) = P_h(l,h) = c$.

The second reason is competitive advantage. If the stores are relatively far to the right (i.e. $\frac{h+l}{2} > \frac{1}{2}$) the left store will be the closest store to more people and therefore have an advantage and charge a higher price than the store on the right.

If the left firm would be located slightly further to the right, it would charge a slightly higher price, that is the increase in competitive advantage outweighs the increase in competition (Exercise: Verify this by evaluating the appropriate derivative.) At the same time the rival would charge a lower price, both due to lower differentiation and due to a loss of competitive advantage.

Stage 1: Product choice In the first stage, the two firms simultaneously choose a location for their stores, taking into account how store location will affect future price competition and sales. Firm l's profit can, as usual, be written as the product of markup and sales:

$$\Pi_l(l,h) = [P_l(l,h) - c] \cdot D_l(P_l(l,h), P_h(l,h), l, h)$$

where $P_l(l,h) - c$ is the markup.

Note that we here wish to emphasize that the firms' profits are functions of their locations, $\Pi_l(l,h)$, and that prices will adjust to the locations chosen by the firms, i.e. $P_l(l,h)$ and $P_h(l,h)$. In particular, firm l's profit depends on its own location through three different channels. To be concrete let us study the effect of moving the left store further to the right, while still being to the left of the other store:

1. The location l affects demand directly, i.e. $\frac{\partial D_l(P_l(l,h),P_h(l,h),l,h)}{\partial l}$. By moving further to the right (still being to the left of the other store), the left store will be the closest store to a larger number of people (a gain in competitive advantage), which increases sales at similar prices.

2. The location l affects the firm's own price, i.e. $\frac{\partial P_l(l,h)}{\partial l}$. Price is increased since the competitive advantage outweighs the differentiation effect. This increase in price will, in turn, affect both the firm's markup and its sales.

3. The location l also affects the competitor's price, i.e. $\frac{\partial P_h(l,h)}{\partial l}$. The competitor's price is reduced both due to the rival's loss in competitive advantage and due to the reduction in differentiation. The change in the competitor's price will, in turn, reduce the firm's sales.

All these effects can be conveniently summarized by taking the derivative of the firm's profit function with respect to its location, i.e.

$$\frac{d\Pi_l}{dl} = (P_l - c)\cdot\frac{\partial D_l}{\partial l} + \left[D_l + (P_l - c)\cdot\frac{\partial D_l}{\partial P_l}\right]\cdot\frac{\partial P_l}{\partial l} + (P_l - c)\cdot\frac{\partial D_l}{\partial P_h}\cdot\frac{\partial P_h}{\partial l}.$$

Each term on the right hand side corresponds (in order) to one of the three channels listed above. Recall, however, that the firm chooses its price optimally. This means that whatever location the firm chooses, it will later set its price such that the first-order condition $D_l + (P_l - c)\cdot\frac{\partial D_l}{\partial P_l} = 0$ is fulfilled. This means that the second term above will be equal to zero, independent of what location the firm chooses. This is an application of the Envelope Theorem. Thus:

$$\frac{d\Pi_l}{dl} = (P_l - c)\cdot\left[\frac{\partial D_l}{\partial l} + \frac{\partial D_l}{\partial P_h}\cdot\frac{\partial P_h}{\partial l}\right].$$

The Envelope Theorem thus teaches us that we only need to consider what effect location has on demand directly (competitive advantage) and how location affects the rival's price.

The next job is to find expressions for all the partial derivatives in the second factor on the right (i.e., the two term in square brackets).

The first term can be written:

$$\frac{\partial D_l}{\partial l} = \frac{1}{2} + \frac{p_h - p_l}{2t\left(h-l\right)^2} = \frac{1}{2} + \frac{\frac{2}{3}t\left(h-l\right)\left(1-h-l\right)}{2t\left(h-l\right)^2} = \frac{2+h-5\cdot l}{6\cdot\left(h-l\right)}.$$

This effect is positive if prices are not too different, but may also negative. We will leave it for now. The second term can be written:

$$\frac{\partial D_l}{\partial P_h}\cdot\frac{\partial P_h}{\partial l} = \frac{1}{2t\left(h-l\right)}\left[-\frac{2}{3}t\left(2-\frac{h+l}{2}\right) - \frac{2}{3}t\left(h-l\right)\left(\frac{1}{2}\right)\right] = -\frac{2\cdot\left(2-l\right)}{6\cdot\left(h-l\right)} < 0.$$

The second term, which must be negative, shows that moving the left store further to the right will provoke the rival to reduce its price, which reduces the left firm's sales.

Summing the two terms yields:

$$\frac{\partial D_l}{\partial l} + \frac{\partial D_l}{\partial P_h}\cdot\frac{\partial P_h}{\partial l} = \frac{2+h-5\cdot l}{6\cdot\left(h-l\right)} - \frac{2\cdot\left(2-l\right)}{6\cdot\left(h-l\right)} = \frac{-2+h-3\cdot l}{6\cdot\left(h-l\right)} < 0.$$

Exercise 54. Derive this result yourself.

The expression must be negative. The denominator is positive since $h > l$ and the nominator is negative since h is at most equal to one. This implies that

$$\frac{d\Pi_l}{dl} < 0$$

since the markup is strictly positive (as can be verified). In other words, the firm wants to locate its store as far to the left as possible, i.e. at $l = 0$. The intuition for this result is that the left firm reduces differentiation by moving to the right. Less differentiation means more price competition and in particular that the rival will lower its price. When the rival lowers its price, the left firm's sales go down. This effect is so strong that it will always be optimal for the left firm to locate as far away from the rival as possible.

Exercise 55. Use similar reasoning to show that the other firm wishes to locate its store as far to the right as possible.

15.2.2. Discussion

Maximal differentiation In effect, we have demonstrated the "principle of maximum differentiation" which asserts that, at least under certain circumstances, firms wish to locate their products as far away from each other as possible, in order to minimize price competition. This result is not trivial since locating at the extreme end of the street means that the firms impose high transportation costs on their customers, which in turn reduces the customers' willingness to pay for the goods. But, under certain conditions, the competition effect dominates the transportation cost effect.

Extension: Locating outside the "city" The analysis above assumed that the firms had to position their stores "inside the city," that is in the interval [0,1]. If allowed, the firms would actually even prefer to locate their stores even further apart, outside the city, despite no consumers living there.

Minimal differentiation If, however, the two firms where not allowed to compete in prices, the choices would be different. The inability to compete in prices could come from some regulation, as in the case in the newly deregulated Swedish pharmacy market. Another example is commercial radio which is free-to-air but earn advertising revenues in proportion to their number of listeners. In such markets, the firms will both choose to locate in the middle of the street to be as close as possible to the customers. And, in fact, many people complain that most commercial radio channels are rather indistinguishable, playing the same kind of music. And it appears that the commercial pharmacies only open new stores in the city centers and large shopping malls where there already are other stores.

Social Welfare A social planner, who would like to minimize the total transportation costs, would locate one store at 1/4 and

the other at 3/4. But a social planner who could affect the location of shops (municipal zoning rights) but not their pricing, would actually prefer the shops to be even closer than that, to promote price competition. There is thus too much differentiation in equilibrium when there is price competition, but too little differentiation in markets subject to price regulation.

First-mover advantage It is often claimed that it is important to be the first firm to set up operations in a new market. The first mover will earn a higher profit than the followers. This difference will often persist over time. An obvious reason for this is that the first mover will attract many customers during the time it has a monopoly position. And, then customers will be reluctant to switch to new entrants since switching is often associated with various costs.

Switching costs may for example include the cost of gathering information and comparing the different alternatives or to learn how to use the other firms' variants. Another reason for the first-mover advantage is that the first firm can select a more favorable position in the product space - a better product design. For example, a retailer which locates its store in the middle of the Hoteling street will be relatively close to most customers. And, when a second firm enters and opens a store, it will select at position away from the firs store. Assume that the "right store" is located at the middle of the street, then the entrant will locate its store even further away from town than when firms choose their locations simultaneously. As a result, the first-mover will be able to both sell more and to charge a higher price than the second-mover.

To analyze the first-mover's choice of location, we would have to consider a model with four stages, namely (i) first-mover selects location, (ii) second-mover selects location, (iii) both firms set prices simultaneously and (iv) consumers select stores. In this case the second-mover will have to take the first store's location as given

when selecting a location in the second period. The first-mover has to consider the effect of its choice on the second-mover's choice.

15.2.3. Solutions to selected exercises

Stage 3: derive demand (Exercise 52) To derive demand, we need to find the location of the consumer who is indifferent between the two stores. The indifferent consumer will be located at some point z which can be found by equating the utility of going to the left store $u_l(x) = u_h(x)$:

$$v - p_l - t \cdot (l - x)^2 = v - p_h - t \cdot (x - h)^2 .$$

Solving this equation for x shows that the indifferent consumer is located at

$$z = \frac{h + l}{2} + \frac{p_h - p_l}{2 \cdot t \cdot (h - l)}.$$

To see this, note that $(p_h - p_l)/t = (l - z)^2 - (h - z)^2$ and use the fact that $x^2 - y^2 = (x - y) \cdot (x + y)$.

Since the left store will receive all customers living to the left of the indifferent customer and the right store will receive all the consumers living to the right of the indifferent customer, the two stores' demand functions are given by:

$$D_l(p_l, p_h l, h) = \frac{h + l}{2} + \frac{p_h - p_l}{2 \cdot t \cdot (h - l)}$$

and

$$D_h(p_l, p_h l, h) = 1 - \frac{h + l}{2} + \frac{p_l - p_h}{2 \cdot t \cdot (h - l)}.$$

Stage 2: Price setting (Exercise 53) Firm l's profit is given by

$$\Pi_l(p_l, p_h l, h) = (p_l - c) \cdot D_l(p_l, p_h l, h) = (p_l - c) \cdot \left(\frac{h + l}{2} + \frac{p_h - p_l}{2 \cdot t \cdot (h - l)} \right).$$

Firm l's first-order condition is given by

$$\frac{\partial \Pi_l\left(p_l, p_h l, h\right)}{\partial p_l} = D_l\left(p_l, p_h l, h\right) + \left(p_l - c\right) \cdot \frac{\partial D_l\left(p_l, p_h l, h\right)}{\partial p_l} =$$

$$= \left(\frac{h+l}{2} + \frac{p_h - p_l}{2 \cdot t \cdot (h-l)}\right) - \left(p_l - c\right) \cdot \frac{1}{2 \cdot t \cdot (h-l)} = 0.$$

Firm l's best-reply function is found by solving the first-order condition for p_l:

$$p_l = \frac{c + (h+l) \cdot t \cdot (h-l)}{2} + \frac{1}{2} \cdot p_h.$$

Using the same procedure, Firm h's best-reply function is given by:

$$p_h = \frac{c + (2-h-l) \cdot t \cdot (h-l)}{2} + \frac{1}{2} \cdot p_l.$$

To find the equilibrium, substitute h's best reply into l's best reply

$$p_l = \frac{c + (h+l) \cdot t \cdot (h-l)}{2} + \frac{1}{2} \cdot \left[\frac{c + (2-h-l) \cdot t \cdot (h-l)}{2} + \frac{1}{2} \cdot p_l\right]$$

and solve this single equation in p_l:

$$p_l = c + t \cdot (h-l) \cdot \frac{2+h+l}{3}.$$

To find h's price substitute l's price into h's best-reply:

$$p_h = \frac{c + (2-h-l) \cdot t \cdot (h-l)}{2} + \frac{1}{2} \cdot \left[c + t \cdot (h-l) \cdot \frac{2+h+l}{3}\right]$$

and solve for price:

$$p_h = c + t \cdot (h-l) \cdot \frac{4-h-l}{3}.$$

To summarize, the equilibrium prices depend on the locations of the stores and may be written as functions of these locations, $P_h\left(l, h\right)$ and $P_l\left(l, h\right)$.

15.3. Oligopoly & Vertical differentiation

Starbucks success comes from pioneering so-called specialty coffee (which is just another word for premium or gourmet coffee) in a country that previously was satisfied with less. That being better is a selling point is not so difficult to understand. But, what about Euro Shopper? Euro Shopper's minced meat often contains 20% fat, rather than only 10% as in most other brands. Fair enough, the price is set accordingly. The odd thing is that Euro Shopper seems to make low quality a selling point, with simple packaging and the ugliest logo on the planet. Similarly, the German grocery retailer Lidl's stores are small and ugly, with little product variety. While other European retailers may once have dismissed them as an anomaly, they have now successfully spread their concept to other countries including Sweden. What is the point of low quality?

(To be continued.)

Research & Development

(To be continued.)

16. Entry and market concentration

Some industries are more concentrated than others. In the UK, for example, a 1970:ies study showed that 10 industries out of 118 had a 5-firm concentration ratio above 90 percent while 40 industries were below 40 percent. But concentration not only varies across markets but also over time. There is for example little doubt that average concentration increased during long periods of the 20th century.[1]

A closer look reveals that the markets that are relatively concentrated in Sweden or in Spain are also relatively concentrated in other countries. One example is the mobile phone industry. In the late 1980:ies almost all OECD countries had only one monopoly network. Today, the same countries have at least three competing networks. It thus appears that it is the characteristics of the different markets that determine their concentration levels. In this chapter, we will discuss how industry characteristics, more precisely, determine market concentration.

Economic theory suggests that market concentration is an important determinant of welfare. High levels of concentration allow firms to exploit scale economies. But high levels of concentration also allow firms to engage in monopolistic practices, which leads to a misallocation of resources and possibly to a more skewed distribution of wealth. Fewer firms may also imply less product variety. A second question, addressed in this chapter, is if there is any reason

[1] B. Curry and K. D. George: Industrial Concentration: A Survey, *The Journal of Industrial Economics*, Vol. 31, No. 3 (Mar., 1983), pp. 203-255.

to believe that market concentration is too high or too low from a social welfare point of view?

Market concentration is the result of many different activities, notably entry and exit as well as mergers and acquisitions and spin-offs. We will here focus on the entry process. Mergers and acquisitions are considered in a later chapter.

16.1. Exogenous entry costs

To study the determinants of market concentration, we need to study how firms enter a market. The normal procedure consists of two steps. First, we study how firm profits depend on the number of firms in the market. For this step, an oligopoly model is used to describe how prices and quantities are determined in equilibrium, as a function of concentration. Second, we need to determine the number of firms. To this end, we use the simplest possible entry model, which assumes that firms enter the market as long as it is profitable to do so, taking into account the costs of setting up their operations.

16.1.1. How profits depend on concentration

To study how profits depend on concentration, i.e. the number of firms in the market, we use an oligopoly model. Here we will apply a linear Cournot model with differentiated goods. Since we looked at Cournot models in detail elsewhere (see chapters 6.4 and 6.5), I will proceed without much explanations here.

Preferences and Technology Consider a market with m identical consumers. The consumers love variety, i.e. they all prefer to consume a little of all products to consume more of a single product. The inverse market demand for product i is linear and given

by

$$p_i = \alpha - \frac{1}{m} \cdot q_i - \sigma \cdot \frac{1}{m} \cdot \sum_{j \neq i} q_j,$$

where $\sigma \in [0, 1]$ is the degree of substitutability between the different goods. When people consume more of a certain good j, they will demand less of another good i.

All firms have the same constant marginal cost, c, and fixed cost F.

Cournot equilibrium Firm i's profit is given by

$$\Pi_i = \left(\alpha - \frac{1}{m} \cdot q_i - \sigma \cdot \frac{1}{m} \cdot \sum_{j \neq i} q_j - c \right) \cdot q_i - F,$$

Solving the Cournot model, it turns out that each firm will produce the same quantity in equilibrium, which is given by

$$q = \frac{\alpha - c}{2 + (n-1) \cdot \sigma} \cdot m.$$

Note that each firm will produce more the larger the market is, m, and less the more firms there are, n. Each firm will also produce less the closer substitutes the goods are, σ. The reason is that the consumers will become satiated quicker when the goods are close substitutes.

The equilibrium price is found by evaluating the inverse demand function at the equilibrium quantity. After some simplifications, the equilibrium price is determined to be

$$p = \lambda \cdot \alpha + (1 - \alpha) \cdot c$$

where

$$\lambda = \frac{1}{2 + (n-1) \cdot \sigma},$$

which is strictly positive but below a half, and the same for all

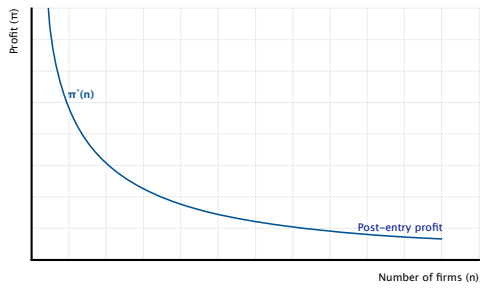
firms. The price will be lower the harder competition there is, i.e. the more firms there are in the market, n, or the better substitutes the goods are, σ. The last point is easily seen since more firms leads to a lower λ which means that the price will move closer to marginal cost and further away from the intercept of the inverse demand function, α.

The equilibrium profits is found by substituting the equilibrium quantities and price back into the firm's profit functions:

$$\pi = m \cdot \left(\frac{\alpha - c}{2 + (n-1) \cdot \sigma} \right)^2 - F.$$

Thus, equilibrium profit is among other things determined by the size of the market, m, and the fixed cost, F, as well as two factors determining the strength of competition, namely the number of firms, n, and the degree of substitutability, σ. Notice in particular that more firms leads to lower prices and to a lower quantity produced by each firm. Therefore, more firms leads to lower profits. Figure 16.1 illustrates how a firm's profit depends on the number of firms in the market.

Figure 16.1.: Entry



Before reading any further, please solve the following exercise.

Exercise 56. Verify that the equilibrium quantities, prices and profits are indeed given by the expressions above. A suggested

solution is available in chapter 16.5.

Measures of concentration

When all firms are identical we can measure concentration by the inverse of the number of firms in the market. This is also true with differentiated goods as long as all firm have the same "vertical quality," here α, and as long as all products are exposed to the same amount of competition, which is true here since all products are equally close substitutes to each other. There are many ways to measure concentration. The easiest measure is simply to sum the market shares of the k largest firms. This measure is called the k-firm concentration ratio, denoted CR_k. When there are n firms in the market and all firms have the same market shares, this measure simplifies to $CR_k = k/n$. Another common measure of concentration is the so-called Herfindahl concentration index, denoted HHI. This index is the sum of the squared market shares of all firms, i.e.

$$HHI = \sum_{i=1}^{n} s_i^2.$$

If all firms have the same market shares:

$$HHI = \sum_{i=1}^{n} \left(\frac{1}{n}\right)^2 = n \cdot \left(\frac{1}{n}\right)^2 = \frac{1}{n}.$$

In particular, $HHI = 1$ in case of monopoly, and $HHI \approx 0$ if there is a very large number of firms in the market.

16.1.2. Entry

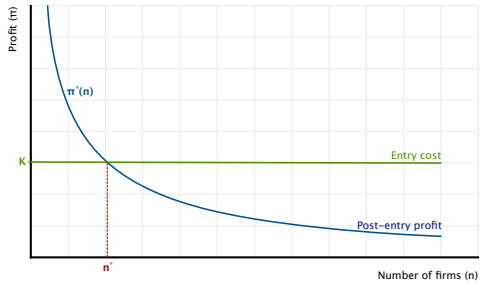
Oligopoly models takes the number of firms to be exogenously given. Now the time has come to determine the number of firms (i.e. concentration) endogenously and to identify the factors that determine how many firms there will be in a market.

To do so, we will use the simplest possible entry model. This entry model assumes that firms are free to enter the market if they wish, but that entry is costly. When entering a market a firm may have to take on various investments such as in buildings and machinery. In some markets it is also necessary to do extensive research and development before entering. In yet other markets it may be necessary to collect information about demand and to advertise the firm's products. We denote these entry costs by K.

Not all of these investments are entry costs, however. The firm may be able to recoup parts of the investments e.g. by selling machinery second-hand, if it were to exit the market. Only the parts of investments that cannot be recouped are considered entry costs. The costs that cannot be recouped are sometimes referred to as sunk in the economic literature. When entering a market a firm may also commit to certain costs in the future. One example may be the costs associated with removing specialized buildings or restore land when the firm exits the market at a later stage. Since this cost is a consequence of the entry decision, they are also considered part of the entry cost.

With this simple model, firms continue to enter the market as long as the post-entry profit exceeds the entry cost. Since entry leads to lower profits, the entry process will come to a stop, when additional entry would lead to a post-entry profits below the entry cost. This entry process is illustrated in Figure 16.2. As above, the blue curve represents how a firm's profit depend on the number of firms in the market. The green horizontal line represents the entry cost. To the left of n^* profit is higher than the cost of entry and

Figure 16.2.: Entry



more firms will get in to the market. This process end when the number of firms is equal to n^*.

That is, if we ignore the fact that the number of firms must be an integer, we can determine n by the condition that profit should be equal to the cost of entry, i.e. $\pi^*(n) = K$ or

$$m \cdot \left(\frac{\alpha - c}{2 + (n-1) \cdot \sigma} \right)^2 - F = K.$$

This equation is known as the zero-profit condition. (If we care about the integer-problem, the number of firms will simply be the largest integer below the n determined by the zero-profit condition.) Solving the zero-profit condition for the equilibrium number of firms we get:

$$n^* = -\frac{2-\sigma}{\sigma} + \frac{\alpha - c}{\sigma} \cdot \sqrt{\frac{m}{K+F}}. \tag{16.1}$$

Before reading any further, please solve the following exercise.

Exercise 57. Verify that the equilibrium number of firms is indeed given by the expression above, by solving the zero-profit condition. A suggested solution is available in chapter 16.5.

16.1.3. Determinants of concentration

Despite its simplicity, the entry model reveals several important factors that determine how concentrated markets are. In particular, it is the right hand side of equation 16.1 that reveals which these factors are.

Entry barriers and scale economies First, the cost structure matters for concentration. There will be fewer firms in the market with higher entry barriers (i.e. if the entry costs K are high) or if there are important economies of scale (i.e. if the fixed cost F is high). By rewriting equation 16.1 slightly as

$$n^* = -\frac{2-\sigma}{\sigma} + \frac{\alpha - c}{\sigma} \cdot \sqrt{\frac{1}{\frac{K}{m} + \frac{F}{m}}}$$

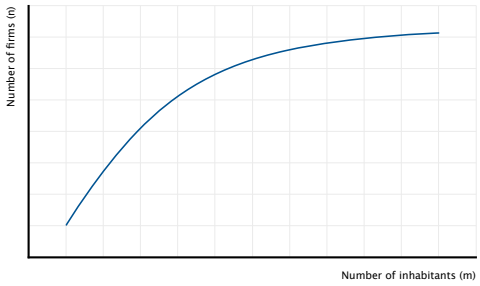
it becomes apparent that what really matters is how large the fixed costs are in relation to market size, i.e. the number of consumers.

Market size Second, the size of the market matters for concentration. There will be more firms in the market, the larger the market is (i.e. the more consumers m there are).

A less obvious result is that the number of firms does not increase in proportion to market size. That is, a market with twice as many consumers will not have twice as many firms. The reason is the increase in competition that follows from lower concentration. Since more firms reduce mark-ups $p - c$, each firm must sell to more consumers to cover its fixed costs. Figure 16.3 shows this relationship. And a numerical example illustrates this point even further. Consumer a hypothetical market characterized by $\sigma = \alpha - c = 1$ and $K + F = 10000$ so that

$$n^* = -1 + \sqrt{\frac{m}{10000}}.$$

Figure 16.3.: Entry



Then, any city with 40 000 inhabitants will have only one firm in the market. A city with 90 000 inhabitants will have two firms. Expressed differently, each duopoly firms needs to serve 45 000 consumers to break even. A monopoly firm only needs 40 000.

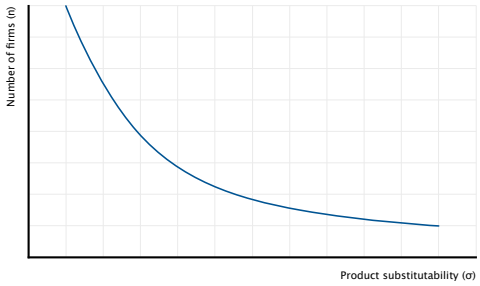
Note that we have abstracted from an important aspect of the problem. The more firms that enter the market, the more difficult it may be to differentiate the new product from other products that are already for sale in the market. As the product space becomes more packed, there is a sharp increase in price competition with every new firm in the market. As a result, the number of firms may increase with market size at an even lower rate than suggested in the model considered here.

Product differentiation Third, product differentiation matters for concentration. If the goods are poor substitutes, i.e. if σ is low, there will be more firms in the market.[2] Figure 16.4 shows this relationship. That is, if competition tends to be weak because products are poor substitutes, competition tends to be strengthened by a larger number of firms.

A similar point is that there will be fewer firms in the market if

[2]To show this, take the derivative of the number of firms with respect to the substitution parameter. Be careful to consider the case when there is at least one firm in the market.

Figure 16.4.: Product differentiation



the firms compete a la Bertrand rather than a la Cournot. With Bertrand competition (absent capacity constraints) firm profit will be lower for any given number of firms in the market. Thus, fewer firms will be able to fit in to the market before profits are zero.

16.2. Entry and Welfare

For the consumers, entry has two beneficial effects in the model considered here. Entry increases the range of products available for the consumers to choose from and it increases price competition, thereby driving price closer to marginal cost. For the firms, entry leads to higher costs, both since economies of scale will be less exploited (i.e. additional fixed costs must be paid) and since there is a direct cost of entry. The question is what happens to the overall level of welfare.

Social welfare is here defined as the sum of the consumer surplus (which measures consumer utility in monetary terms) and firm profits, i.e.

$$W = \left[CS^{Gross} - p \cdot Q\right] - \left[p \cdot Q - TC\right] = CS^{Gross} - TC.$$

The first term is consumer surplus (gross consumer surplus minus the expenditures) and the second term is firm profits (revenues

minus total cost). Since the consumers' expenditures and the firms' revenues cancel, welfare is simply gross consumer surplus minus the cost of production.

16.2.1. Cournot with perfect substitutes

To study the socially optimal number of firms we simplify calculations by assuming that the goods are perfect substitutes (i.e. $\sigma = 1$). Then the equilibrium quantity produced by each firm is given by quantity in equilibrium, which is given by

$$q^*(n) = \frac{\alpha - c}{n+1} \cdot m.$$

Let $Q^*(n) = n \cdot q^*(n)$ be aggregate production, which is a function of the number of firms in the market. Then, social welfare can be written

$$W(n) = \int_0^{Q^*(n)} P(Q) \cdot dQ - [c \cdot Q^*(n) + (F+K) \cdot n]$$

where the first term is the (gross) consumer surplus and the last term is the costs of production and entry. Social welfare clearly depends on how many firms there are in the market, both because the number of firms affects the (gross) consumer surplus and because it affects the total cost.

Let us continue to ignore the fact that the number of firms should be an integer. Thus, to find the socially optimal number of firms

we use the first-order condition, given by[3]

$$\frac{dW(n)}{dn} = [P(Q^*) - c] \cdot \frac{dQ^*(n)}{dn} - (F + K) = 0.$$

The first term is the benefit of increasing the number of firms. This benefit is simply the fact that aggregate production will increase as the number of firms increases. The social value of a small increase in production is equal to the mark-up, that is to say the difference between the value of consuming the extra unit and the cost of producing this extra unit. Expressed differently, a larger number of firms reduces the dead weigh loss associated with limited competition. The second term represents the cost of increasing the number of firms, which is the fixed cost of production and the entry cost.

Before reading any further, please solve the following exercise.

Exercise 58. Verify that the equilibrium number of firms larger than the socially optimal number of firms. A suggested solution is available in chapter 16.5.

This result is probably surprising to most people. One may say that there is too much competition! So, what is the explanation? To understand why there is too much entry, it is instructive to think in terms of the *externalities* caused by a firm's decision to enter the market.

- On the one hand, entry leads to increased competition and lower prices. This effect benefits the consumers and hurts the incumbent firms. We know, however, that the net effect of increased competition is positive. That is, the consumers

[3]To derive this first order condition you need to use the rule for taking derivatives with respect to a limit of integration, i.e

$$\frac{\partial \left[\int_a^b f(x) \cdot dx\right]}{\partial b} = f(b).$$

gain more than the incumbents loose. This *competition effect* implies that there may be too little entry in equilibrium.

- On the other hand, entry means that the entrant takes over market shares from the incumbents. This effect, often referred to as *business-stealing*, implies that profits are transferred from the incumbents to the entrant, who uses them to pay for the costs of entering. The business-stealing effect implies that there might be too much entry in equilibrium.

As it turns out, in a Cournot model with perfect substitutes, the destruction of profits is worse than the benefits of competition, at least at the margin. That is, the last firms that enter the market causes more harm then good. Intuitively, the reason is that the positive competition effect is rather weak when there are already some firms in the market. It also matters that the firms compete a la Cournot, which is a relatively weak form of competition (see below).

This exercise also illustrates the crucial role of the choice of welfare measure. We concluded that there is too much entry because, at the margin, the firms' loss of profits outweighs the consumers' gains. If we would solely care about consumer welfare we would come to the opposite conclusion, that there is too little entry. However, striving for more entry would mean that the firms make losses which would have to be subsidized somehow.

16.2.2. Other modes of competition

If all the firms in the industry succeed to form a cartel, independent of concentration, entry will not increase competition and thus not have any positive effects. Again, there is too much entry in equilibrium.

In other cases, there may actually be too little entry. Think about a market with Bertrand competition and homogenous goods. With

two firms prices are competed down to marginal cost and the whole monopolistic dead weight loss is eliminated. However, if there is a small entry cost (or a small fixed cost) a second firm will not enter the market. In this case the social benefits from entry (a complete elimination of the monopolistic dead weight loss) would have been larger than the small entry cost. Actually, in this case, it might make sense to subsidize entry.

The general conclusion is thus that entry is associated with two externalities, namely the competition and the business-stealing effects, and that depending on the relative strength of these two effects, there may be too little or too much entry in equilibrium. The more competition is increased by entry, the more valuable entry is. In any case, there is no reason to believe that industry concentration would be socially optimal.

16.3. Additional aspects of the entry process

Although the simple entry model presented above delivers several important insights about the determinants of concentration, many other aspects are left out. In some markets entry may be restricted by regulation. One example is that Apoteket AB had a legal monopoly on the Swedish pharmacy market up to 2009. In some markets firms need regulatory approval to enter. For example, in Sweden food retailers need municipal approval to open a store due to the zoning regulation. In yet other industries, firms need to buy one of a limited number of permissions to operate in the market. In other industries, crucial resources may limit entry. One example is the radio spectrum used for mobile phones, radio and television.

Strategic aspects may also be important. Incumbent firm may, for example, engage in activities to keep potential rivals out of the

market. Such entry deterrence may take different forms, but one example may be to fill the space of possible product varieties before anyone else has time to position a product.

16.4. Endogenous Entry Costs

In some product markets, the level of concentration appears to be relatively independent of market size. For example, there are two main providers of beers in Portugal (Centralcer and Unicer) while in the United States, which is 50 times as large, there are only three (Anhauser Bush, Miller, Coors). What can possibly explain this pattern?

We have already noted one part of the explanation: price competition. When there are more firms in a geographical market, they will compete more fiercely in prices and earn lower markups. Therefore, each firm must sell to a larger number of customers. Otherwise it will not be able to cover its fixed costs. As a consequence, the number of firms cannot double when the market size doubles. In the most extreme example, Bertrand competition without product differentiation, there is only room for one firm, independent of the size of the market.

Price competition does not appear to be a good explanation, howver. The reason is that industries that remain concentrated, also in geographical markets with many consumers, tend to be R&D and advertising intensive. In such industries one may suspect that price competition is less intense. R&D often creates product differentiation. Advertising often creates product loyalty, perhaps because consumers perceive products to be more different than they really are. Beer may be a case in point. The difference in taste between different lagers is arguably not that great. When price competition is lax, the number of firms should increase fast in market size.

It has also been noted that both R&D and advertising costs are higher in larger markets. For example, advertising cost as percentage of sale is roughly the same in US and Portuguese beer industry. Thus, such costs may help explain the concentration–market size puzzle. But, why are the firms' costs increasing in market size?

According to the theory of endogenous entry costs, it is because the firms choose to invest more heavily in R&D and advertising in large markets. The aim of this chapter is to explain this idea. To show the importance of endogenous entry costs we consider a model without price competition and first reiterate the analysis when entry costs are exogenous.

16.4.1. Benchmark: No advertising

Consider first a market with exogenous entry costs. There are m customers and let's say that n firms have entered the market, each owning one store. The customers choose stores at random and they do not have the time to shop around: either they buy the good at the store they happen to come to, or they don't buy it at all. Each firm will thus have $\frac{m}{n}$ customers.[4]

Customers attach value v to one unit of the product. No one is interested in consuming a second unit. Since customers do not have the time to shop around, the firms have monopoly power over their customers. They will consequently all charge the monopoly price $p = v$. Each firm's revenue is therefore equal to $\frac{m}{n} \cdot v$.

Assume for simplicity that production does not cost anything, but that each firm has to pay F to set up the shop and enter the market. Each firm will therefore make a profit equal to $\frac{m}{n} \cdot v - F$. The profit of an individual firm is lower the more competitors there are since market shares will then be lower.

[4]The result that each firm has the same share of the market is of course an approximation since it is possible that, say, all customers by chance end up in the same store. But when the number of customers is very large, the outcome should often be close to the approximation.

Entry of new firms will occur until the profit in the market is equal to zero. This means that the number of firms in equilibrium is equal to $n = \frac{v}{F} \cdot m$ That is, the number of firms is proportional to the number of customers, *i.e.* the size of the market. If market size doubles, so will the number of firms. This result is extreme since there is no price competition in this model. If prices would be lower with more firms in the market, the number of firms would still increase with the size of the market, but at a lower rate.

16.4.2. Advertising

Consider now a market where firms must advertise in TV to catch their customers' attention. Assume for simplicity that a customer visits the shop whose advertisement he or she happened to see last. In particular, let us assume that the probability of visiting firm i's shop is proportional to the number of ads the firm runs. If firm i runs r_i ads, the probability that a given customer visits firm i's shop is given by r_i/R where $R = \sum_{j=}^{n} r_j$ is the total number of ads run by all the firms in the market. Since there is a large number of customers, firm i's shop will attract approximately $(r_i/R) \cdot m$ customers.

Assume for simplicity that the customers' valuation of the good, v, is independent of the amount of advertising and that the firms can set the monopoly price, $p = v$. There is no price competition. Firm i's total revenue is therefore equal to $(r_i/R) \cdot m \cdot v$.

Firms incur the cost c for every ad they run. Firm i's profit is therefore equal to

$$\pi_i = (r_i/R) \cdot m \cdot v - r_i \cdot c - F.$$

Clearly the firms' choices of advertising are interdependent. So, in order to determine how many ads they run we need to consider advertising as a game. Let us assume that they make their decisions

simultaneously. Thus, we have defined a static game and can solve for the Nash equilibrium. As it turns out, all firms run the same number of ads, namely

$$r^* = \frac{m}{n} \cdot \frac{v}{c} \cdot \frac{n-1}{n}.$$

The most important feature of the firms' advertising behavior is that, for any given number of rivals, each firm advertises more the larger the market (m) is. Another way to say this is that the firms' advertising costs increase with the size of the market. As a direct result, also advertising costs are increasing in market size:

$$c \cdot r^* = \frac{m}{n} \cdot v \cdot \frac{n-1}{n}.$$

Firms' equilibrium profits can be determined by substituting back the equilibrium level of advertising into the profit function:

$$\pi^* = \frac{m}{n^2} \cdot v - F.$$

Firms' profits are lower the more rivals there are. As before, a part of the reason is that each firm will have a smaller market share when there are many rivals. With endogenous entry costs, there is also a second reason, namely that a larger fraction of potential profit is wasted on advertising when there are many rivals in the market.[5]

Before reading any further, solve the following exercise.

Exercise 59. Compute the equilibrium level of advertising cost. A suggested solution is available in chapter 16.5.

Entry of new firms will occur until the profit in the market is equal to zero. This means that the number of firms in equilibrium

[5] In the simple model considered here, advertising is pure waste. This is not to deny that advertising creates value in many in industries. But this is no the topic for today.

is equal to

$$n^* = \sqrt{\frac{m \cdot v}{F}}.$$

That is, the number of firms increases with market size, but it increases less than in proportion to market size. If the number of customers doubles, the number of firms will increase by approximately 40 percent.

Claim 1. The number of firms increases with market size, but it increases less than in proportion to market size.

Again, the reason is that the firms' costs are increasing with market size, since a larger market increases the firms' advertising competition.

If you would increase the number of customers towards infinity, also the number of firms would increase towards infinity. But there are other—more complicated—models of endogenous entry costs where the number of firms would never exceed a certain finite number, independent of the size of the market. Shaked and Sutton (1987) show that under reasonable conditions there exists a limit on the number of firms in a market, independent of how big the market becomes, if firms compete in both prices and in R&D or advertising.[6]

[6]Ordinary price competition may also give rise to a limit on how fragmented an industry may become as the size of the market increases. The most notable theoretical example is the Bertrand model of price competition with homogenous goods. With constant marginal costs, the price will be equal to marginal cost whenever there are two or more firms. If there are also fixed costs, there would never be more than one firm in the market, independent of the size of the market.

16.5. Solutions to exercises

16.5.1. Cournot equilibrium (Exercise 56)

Best-reply functions Firm i 's profit is given by

$$\Pi_i = \left(\alpha - \frac{1}{m} \cdot q_i - \sigma \cdot \frac{1}{m} \cdot \sum_{j \neq i} q_j - c\right) \cdot q_i - F.$$

The first-order condition is given by

$$\frac{d\Pi_i}{dq_i} = \left(\alpha - \frac{1}{m} \cdot q_i - \sigma \cdot \frac{1}{m} \cdot \sum_{j \neq i} q_j - c\right) - \frac{1}{m} \cdot q_i = 0.$$

Solving the first-order condition for q_i produces the firm's best-reply function:

$$q_i = m \cdot \left(\frac{\alpha - c}{2}\right) - \frac{\sigma}{2} \cdot \sum_{j \neq i} q_j.$$

Equilibrium quantities The equilibrium is defined by a system of n best-reply equations in the n unknown quantities. To solve this system it is convenient to rewrite the best-reply function slightly:

$$\left(1 - \frac{\sigma}{2}\right) \cdot q_i = m \cdot \left(\frac{\alpha - c}{2}\right) - \frac{\sigma}{2} \cdot \sum_{j=1}^{n} q_j.$$

Define the aggregate output $Q = \sum_{j=1}^{n} q_j$, substitute, and simplify:

$$q_i = m \cdot \left(\frac{\alpha - c}{2 - \sigma}\right) - \frac{\sigma}{2 - \sigma} \cdot Q.$$

Next, sum all the "best-reply" functions:

$$\sum_{i=1}^{n} q_i = n \cdot \left[m \cdot \left(\frac{\alpha - c}{2 - \sigma}\right) - \frac{\sigma}{2 - \sigma} \cdot Q\right],$$

i.e.

$$Q = n \cdot \left[m \cdot \left(\frac{\alpha - c}{2 - \sigma} \right) - \frac{\sigma}{2 - \sigma} \cdot Q \right],$$

which is may be solved for the aggregate output:

$$Q = n \cdot m \cdot \left(\frac{\alpha - c}{2 + (n-1) \cdot \sigma} \right).$$

The aggregate quantity can be substituted back into the best reply function:

$$q_i = m \cdot \left(\frac{\alpha - c}{2 - \sigma} \right) - \frac{\sigma}{2 - \sigma} \cdot \left[n \cdot m \cdot \left(\frac{\alpha - c}{2 + (n-1) \cdot \sigma} \right) \right]$$

which may be simplified to

$$q_i = m \cdot \frac{\alpha - c}{2 + (n-1) \cdot \sigma},$$

which is the same for all firms.

Equilibrium price To find the equilibrium price, substitute the equilibrium quantities back into the indirect demand functions:

$$p_i = \alpha - \frac{1}{m} \cdot q_i - \sigma \cdot \frac{1}{m} \cdot \sum_{j \neq i} q_j$$

i.e.

$$p_i = \alpha - \frac{1 + \sigma \cdot (n-1)}{2 + (n-1) \cdot \sigma} \cdot (\alpha - c)$$

which is the same for all firms and may be rewritten as

$$p = \lambda \cdot \alpha + (1 - \lambda) \cdot c$$

where

$$\lambda = \frac{1}{2 + (n-1) \cdot \sigma}.$$

Equilibrium profits To speed up the computation of the equilibrium profit, we first note that by the first-order condition

$$\left(\alpha - \frac{1}{m} \cdot q_i - \sigma \cdot \frac{1}{m} \cdot \sum_{j \neq i} q_j - c\right) = \frac{1}{m} \cdot q_i$$

Thus, the equilibrium profit can be written:

$$\Pi_i = \frac{1}{m} \cdot q_i^2 - F.$$

Substitute for the equilibrium quantities:

$$\Pi_i = m \cdot \left(\frac{\alpha - c}{2 + (n-1) \cdot \sigma}\right)^2 - F.$$

16.5.2. Entry equilibrium with Cournot competition (Exercise 57)

The equilibrium profit is given by

$$\pi^*(n) = m \cdot \left(\frac{\alpha - c}{2 + (n-1) \cdot \sigma}\right)^2 - F.$$

The equilibrium condition (i.e. the zero-profit condition) is:

$$\pi^*(n) = K.$$

Thus:

$$m \cdot \left(\frac{\alpha - c}{2 + (n-1) \cdot \sigma}\right)^2 - F = K.$$

This condition may be rewritten as

$$\left(\frac{\alpha - c}{2 + (n-1) \cdot \sigma}\right)^2 = \frac{K + F}{m} \tag{16.2}$$

and

$$\frac{\alpha - c}{2 + (n-1) \cdot \sigma} = \sqrt{\frac{K+F}{m}}$$

which may be solved for the equilibrium number of firms:

$$n^* = -\frac{2-\sigma}{\sigma} + \frac{\alpha - c}{\sigma} \cdot \sqrt{\frac{m}{K+F}}$$

16.5.3. Socially optimal number of Cournot firms (Exercise 58)

Recall that with perfect substitutes, social welfare is given by

$$W(n) = \int_0^{Q^*(n)} P(Q) \cdot dQ - \left[c \cdot Q^*(n) + (F+K) \cdot n\right].$$

The first-order condition is given by[7]

$$\frac{dW(n)}{dn} = \left[P\left(Q^*\right) - c\right] \cdot \frac{dQ^*(n)}{dn} - (F+K) = 0.$$

When goods are perfect substitutes

$$Q^*(n) = n \cdot m \cdot \left(\frac{\alpha - c}{n+1}\right)$$

and thus

$$\frac{dQ^*(n)}{dn} = m \cdot \frac{\alpha - c}{(n+1)^2}.$$

Also recall that by the first-order condition

$$p - c = \frac{1}{m} \cdot q.$$

[7]To derive this first order condition you need to use the rule for taking derivatives with respect to a limit of integration, i.e

$$\frac{\partial \left[\int_a^b f(x) \cdot dx\right]}{\partial b} = f(b).$$

Thus

$$\frac{dW(n)}{dn} = \frac{1}{m} \cdot q \cdot m \cdot \frac{\alpha - c}{(n+1)^2} - (F + K) = 0.$$

Thus, the socially optimal number of firms, n^W, is characterized by

$$\left(n^W + 1\right)^3 = (\alpha - c)^2 \cdot \frac{m}{F + K}.$$

When goods are perfect substitutes, the equilibrium number of firms (see equation 16.2) is characterized by

$$(n^* + 1)^2 = (\alpha - c)^2 \cdot \frac{m}{F + K}.$$

we can conclude that there are too many firms in equilibrium. This follows from the fact that

$$\left(n^W + 1\right)^3 = (n^* + 1)^2.$$

Since

$$\left(n^W + 1\right)^3 = (n^* + 1)^2 < (n^* + 1)^3,$$

we must have that

$$\left(n^W + 1\right) < (n^* + 1).$$

16.5.4. Advertising equilibrium (Exercise 59)

Recall that firm i's profit is equal to

$$\pi_i = (r_i/R) \cdot m \cdot v - r_i \cdot c.$$

The first-order condition is thus given by

$$\frac{d\pi_i}{dr_i} = \frac{m \cdot v}{R} - r_i \cdot \frac{m \cdot v}{R^2} - c = 0$$

which may also be written

$$r_i = R - \frac{c}{m \cdot v} \cdot R^2,$$

which is similar to a best-reply function. Note that firm i's optimal advertising level depends on how much all other firms advertise. Therefore we need to solve all first order conditions at the same time. There are n such conditions (one for each firm) and there are n unknown variables, namely the firms advertising levels. The easiest way to solve this system of equations is to sum all the first order conditions. Then we get

$$R = n \cdot \left[R - \frac{c}{m \cdot v} \cdot R^2 \right],$$

i.e.

$$(n-1) \cdot R = \frac{n \cdot c}{m \cdot v} \cdot R^2.$$

Solving this equation for the aggregate level of advertising,

$$R = m \cdot \frac{v}{c} \cdot \frac{n-1}{n}.$$

Substituting this expression back into the "best-reply" function yields:

$$r^* = \frac{m}{n} \cdot \frac{v}{c} \cdot \frac{n-1}{n}.$$

A quicker way to solve this problem is to conclude that the right hand side of

$$r_i = R - \frac{c}{m \cdot v} \cdot R^2,$$

is the same for all firms. Thus, all the individual advertising levels must be the same, i.e. $r_i = r$. Then $R = n \cdot r$ and

$$r = n \cdot r - \frac{c}{m \cdot v} \cdot n^2 \cdot r^2,$$

which can be solved for r. To find the cost, simply multiply by the unit cost.

Part VI.

Advanced Commercial Strategies

17. Two-part tariffs

If you wish to visit Gröna Lund in Stockholm you will have to pay both for entering the amusement park and then also for each ride. In some restaurants you can eat as much as you want for a fixed price. It is rumored that Apple earns a high markup on their iPods but a much lower markup on selling the actual songs on iTunes. Telecom companies often charge both a monthly fee as well as a price per call. Why don't firms simply charge per unit of service they offer?

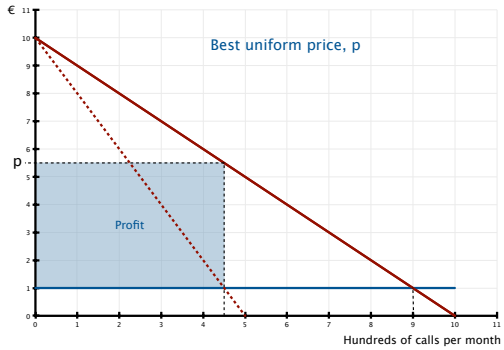
17.1. Monopoly, revisited

Recall the monopolist's dilemma. To sell more, the monopolist must lower its price not only on the extra units, but on all the infra-marginal units as well. As a result, not even monopoly firms can capture the whole potential surplus in a market. The monopolist will set a price sufficiently low to leave the consumer with a surplus on the infra-marginal units. But the price will also be sufficiently high to create a dead-weight loss at the same time. From the monopolist's perspective, both the consumers' surplus and the dead-weight loss are lost profits. Is it possible to for the monopolist to grab a larger part of the potential surplus in the market by using more sophisticated pricing strategies, such as two-part tariffs?

Model Consider a phone company with many identical consumers, all with linear demand, and assume that the phone company has constant marginal cost. Each consumer has to pay a fixed fee F

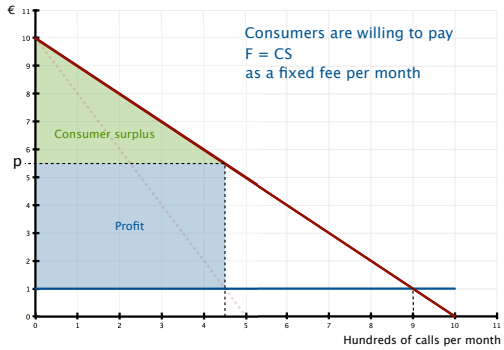
per month as well as a price p per call they wish to make. For simplicity the phone company is a monopolist. The market is described in the Figure. The figure also shows the profit maximizing price, if we disregard the possibility of charging a monthly fee. The

Figure 17.1.: Uniform price



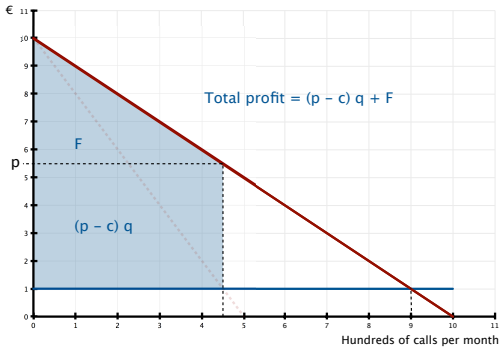
question is now how much the consumers would be willing to pay per month to have the privilege of making calls at the price p. The answer is that they would be willing to pay an amount equal to their consumer surplus. The firm's profit would then be equal to

Figure 17.2.: Consumer surplus



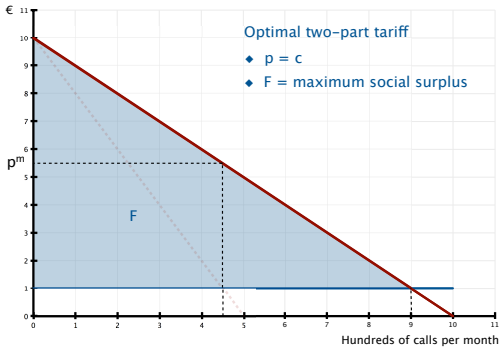
$\pi = (p - c) \cdot q + F$. This profit is clearly higher and probably much higher. In fact, the monopolist succeeds to capture the complete social surplus created by the production and consumption of the q units.

Figure 17.3.: A two-part tariff



The question is now if the monopolist could do even better than this? And the answer is clear, that by lowering its price below the monopoly price, the monopolist can increase the social surplus, which it can capture by increasing the fixed fee instead. To maximize its profit, the monopolist should set the price per unit to maximize the social surplus in the market. That is, the monopolist should charge a price per unit equal to the marginal cost. The optimal fixed fee is then equal to the complete social surplus created at this price. Expressed differently, the optimal fixed fee is equal to the maximum social surplus in the market.

Figure 17.4.: Optimal two-part tariff



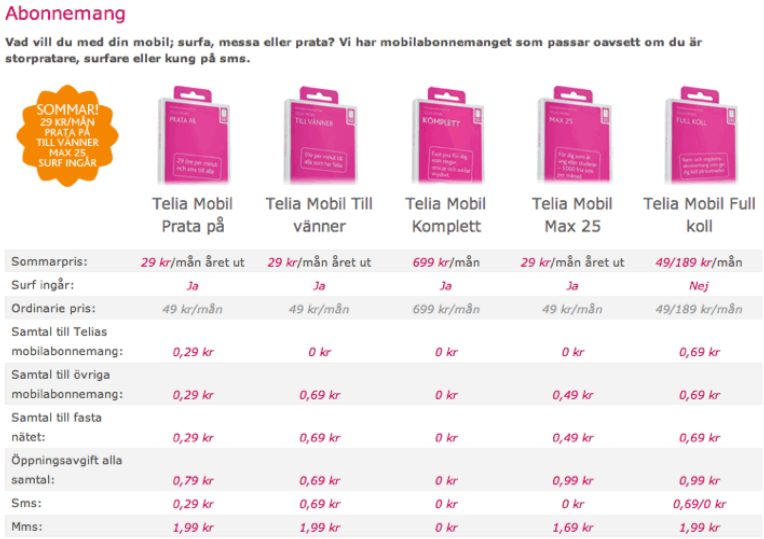
Discussion

In what ways would the Swedish system of value-based pricing be improved by use of two-part tariffs? What information would be needed to compute the welfare gains?

18. Second Degree Price Discrimination

Telia Telias's prices for mobile phone subscribers are summarized in a table on their website. Below you can find the prices from September 2010. Telia thus offers its customers a choice between

Figure 18.1.: Telia



Abonnemang

Vad vill du med din mobil; surfa, messa eller prata? Vi har mobilabonnemanget som passar oavsett om du är storpratare, surfare eller kung på sms.

SOMMAR! 29 KR/MÅN PRATA PÅ TILL VÄNNER MAX 25 SURF INGÅR

	Telia Mobil Prata på	Telia Mobil Till vänner	Telia Mobil Komplett	Telia Mobil Max 25	Telia Mobil Full koll
Sommarpris:	29 kr/mån året ut	29 kr/mån året ut	699 kr/mån	29 kr/mån året ut	49/189 kr/mån
Surf ingår:	Ja	Ja	Ja	Ja	Nej
Ordinarie pris:	49 kr/mån	49 kr/mån	699 kr/mån	49 kr/mån	49/189 kr/mån
Samtal till Telias mobilabonnemang:	0,29 kr	0 kr	0 kr	0 kr	0,69 kr
Samtal till övriga mobilabonnemang:	0,29 kr	0,69 kr	0 kr	0,49 kr	0,69 kr
Samtal till fasta nätet:	0,29 kr	0,69 kr	0 kr	0,49 kr	0,69 kr
Öppningsavgift alla samtal:	0,79 kr	0,69 kr	0 kr	0,99 kr	0,99 kr
Sms:	0,29 kr	0,69 kr	0 kr	0 kr	0,69/0 kr
Mms:	1,99 kr	1,99 kr	0 kr	1,69 kr	1,99 kr

five different pricing plans, called "Komplett," "Prata på" and so on. Although these different plans are presumably targeted to different consumer groups, each person can make his own choice.

A second noteworthy feature is that Telia is charging both a monthly subscription fee and different usage fees, per call, per

minute, per text message (SMS) and so on. If you choose "Komplett," Telia will charge you SEK 699 per month, but then all the calls are free. There is no opening charge and no minute fee. If you choose "Prata på" the monthly charge is only SEK 49. Instead you will have to pay SEK 0.79 for every call you make and then SEK 0.29 per minute you talk. That is, around SEK 1.40 for a call lasting two minutes.

A third feature (largely following from the above) is that different people will pay different prices per call. A person with "Prata på" who only makes one two- minute call per month, will pay SEK 50.40 for that call. A person with "Prata på" who makes 400 two-minute calls per month will pay SEK 1.62 per call. And, someone who chose "Komplett" and makes 700 two-minute calls pays only SEK 1 per call. The more calls you make, the cheaper is the average price. In effect, Telia offers their customers volume discounts.

A fourth feature is that different people will indeed choose different subscriptions. Anyone making fewer than 464 two-minute calls per month should thus prefer "Prata på" to "Komplett," while people who use their mobile phones much more often should go for "Komplett."

The question is why Telia bothers with such complex pricing? Why not just charge everyone a uniform price per minute. The price of milk is the same independent of who you are and how many liters you drink in a month.

Microsoft A similar practice is that Microsoft markets several different versions of its famous Office suite. As indicated by the names of the different versions, they are targeted to different consumer groups like students and professionals. In the students' version, which sells at a lower price, some features are disabled. The question is why Microsoft doesn't allow students to use some of the more advanced features? It is probably more costly for Microsoft to disable these features than to just let the students have them.

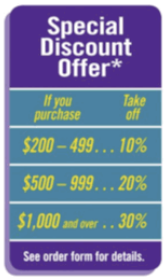
Figure 18.2.: Microsoft

18.1. Setting up a Simple Example: Cost and Demand

In this lecture we will take a close look at quantity discounts, i.e. the practice of charging a lower price per unit the more the customer buys. This practice is very similar to Telia's pricing discussed above, but slightly easier to analyze. Quantity discounts can be offered for different reasons. In some markets quantity discounts reflect that the unit transportation cost is lower for larger orders. This is not a possible explanation for why Telia offers quantity discounts, since transportation costs are essentially zero. Instead we will argue that Telia uses quantity discounts as a clever way to extract a larger share of the consumers' willingness to pay for phone services.

Then you will be asked to work with versioning (as practiced by Microsoft) in one of the problem sets. The logic is very similar.

Figure 18.3.: Quantity discounts

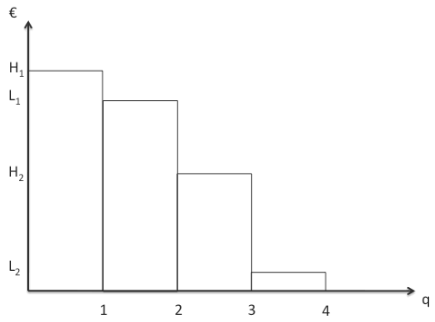


Consider a market served by only one firm. The monopolist first sets price and then the consumers can decide whether they want to buy or not. There are two (types of) consumers called, High and Low. Both High and Low have downward sloping demand. Low is willing to pay L_1 for one unit of the good and $L_2 < L_1$ for a second unit. High is willing to pay H_1 for one unit of the good and $H_2 < H_1$ for a second unit. To avoid introducing more notation than necessary, we are going to assume that there is only one person of each type. It is also assumed that High has a higher willingness to pay for the good than Low. High is willing to pay more for the first unit, $H_1 > L_1$, and also for the second unit, $H_2 > L_2$. To fix ideas, let us also assume that Low is willing to pay more for his first unit than High is willing to pay for a second unit, $L_1 > H_2$. Let us also assume that L_1 is relatively close to H_1 while L_2 is relatively close to cost c.

Market demand is then summarized in Figure. [1]The marginal

[1]In this numerical example, it is assumed that $H_1 = 10$, $L_1 = 9$, $H_2 = 5.5$,

Figure 18.4.



cost of producing the good is constant and equal to $c < L_2$. Thus, the socially efficient production is four units.

18.2. Exogenous Segmentation Revisited

Assume first that the monopolist can distinguish between the two types of consumers. This may be because High lives in one country and Low in another country and that they cannot travel to the other country to buy the good. This is an example of geographical market segmentation. Alternatively High and Low may live in the same country, but be easy to recognize for the firm. High may be old and low may be young customers. Then the firm can make the two groups different offers, at least if reselling between them is costly.

$L_2 = 1$ and $c = 0$.

18.2.1. Pricing to market

Since the monopolist can treat the two groups as two different markets, the monopolist can charge different prices, despite the fact that it is the same good and despite the fact that there may be no differences in the cost of serving the two groups.

Charging different prices for the same good in different markets is called *3:rd degree price discrimination.*

In the High country, the monopolist chooses between charging H_1 and selling one unit or charging H_2 and selling two units. Assuming that the incremental revenue of the second unit is larger than incremental cost, $2 \cdot H_2 - H_1 > c$, the monopolist sets price H_2.[2] This is indeed the case with the numerical example given above (in the footnote).

In the Low country, the monopolist chooses between charging L_1 and selling one unit or charging L_2 and selling two units. Assuming that the incremental cost of the second unit is larger than the incremental revenue, $c > 2 \cdot L_2 - L_1$, the monopolist sets price L_1. This is indeed the case with the numerical example given above (in the footnote).

Note that the demand (willingness to pay) in country High is unambiguously larger than in country Low since $H_1 > L_1$ and $H_2 > L_2$. Still, the monopolist chooses to charge a higher price in Low than in high, since $L_1 > H_2$. That price is higher in a country with lower demand is probably unusual. But it may happen. The reason is that the demand in country Low is much more inelastic, i.e. the firm has to reduce price sharply to sell a second unit. This is an extreme example of the inverse elasticity rule, which requires the monopolist to charge a higher markup the more inelastic demand is.

Charging different prices in different countries is very common,

[2] It may be instructive to rewrite incremental (i.e. marginal) revenue as $H_2 + (H_2 - H_1)$ where the first term is the price of the second unit and the second term is the reduction in the price of the first unit.

however. One example is the price of cars in Europe. The markup of Mercedes is high in Germany and markup of Fiat is high in Italy. This suggests that consumers have a preference for the cars from their own country. Another example is (on patent) pharmaceuticals which are much more expensive in the United States than they are in Canada. In fact the price of some drugs is twice as high in the U.S. Part of the reason why prices can be much higher in the U.S. is that it is illegal to buy drugs in Canada and resell them in the U.S. The pharmaceutical companies can thus treat the two countries as separate markets. Bu also within the European Union, which is supposed to have a common market, the prices of drugs differ between countries. Drugs are typically cheaper in Greece and Spain than they are in Sweden. Even though the European treaty allows parallel trade (to buy where it is cheap for reselling where it is expensive) the price differences remain for some reason.

18.2.2. Packaging

The monopolist can do better than this, however. In the High country the consumer is left with a surplus. In the Low country there is a dead weight loss.

When discussing monopoly pricing it is often implicitly assumed that the monopolist sells the good unit by unit. An alternative would be to sell packages of (say) two units. In fact, it is a quite common practice that firms sell their goods in such multi-unit packages. This is probably often for logistic reasons, but it may also be a strategy for extracting a higher surplus from the consumers.

The idea is that the monopolist does not allow the customers to buy the highly valuable first unit, unless they also buy the less valuable second unit. That serves to eliminate the dead weight loss. When buying the two-unit package, the consumer is charged a price equal to the sum of his willingness to pay for the first and the second unit. This serves to extract the consumer's surplus.

More generally the idea is to tailor-make offers to every individual consumer. Every consumer should be offered to buy the efficient quantity (which typically differs between consumers) and to pay a price equal to that individual's willingness to pay for that quantity.

In the simple example we consider here, the monopolist offers both consumers to buy a two-unit package. The monopolist can then charge High the price $H_1 + H_2$; similarly Low gets the offer to buy two units of the good for $L_1 + L_2$. Both High and Low would accept their offers, still assuming that it is too costly for Low to buy two two-unit packages in order to resell one package to High (arbitrage).

In this case, four units of the good will be produced and consumed, which means that the outcome maximizes the social surplus. Since the monopolist extracts the full social surplus (the consumer surplus is zero) this is the best possible outcome that the monopolist could ever hope to achieve. Hence, the name perfect (or first-degree) price discrimination.[3]

18.2.3. Two-part tariffs

An alternative method to implement the same outcome is for the monopolist to use a so-called two-part tariff. In particular, the monopolist can offer the consumers in this high-demand country to buy the product at a price equal to marginal cost, $p = c$. Then, the consumers will voluntarily choose to consume the socially efficient quantity, in this case two units. For the right to buy the good at this low price, the monopolist can also charge the consumers a fixed

[3]Perfect price discrimination, in the sense of extracting the full social surplus, is rare in practice. The problem with perfect price discrimination is that while firms often know market demand well, they rarely know the willingness to pay of individual customers (at least in consumer markets). It may for example be difficult for the monopolist to tell how much more the elderly are willing to pay than young. And to estimate demand for subgroups, the monopolist would have to gather much more information than to estimate demand for a market as a whole.

fee. The highest possible fee that the consumers are willing to pay is the net consumer surplus, i.e. $F = H_1 + H_2 - 2 \cdot c$. The monopolist can make a similar offer also to the low-demand consumers. Since the monopolist induces all consumers to buy the efficient quantity and is able to extract the full surplus, also two-part tariffs lead to 1:st degree price discrimination when the firm has full information about the individual consumers' demand.

Note that the monopolist can earn a higher profit with both packaging and two-part tariffs than with normal uniform pricing. The reason is that the monopolist uses two instruments instead of only one. With a two-part tariff, the monopolist uses the price per unit to induce the consumers to buy the efficient quantity and the fixed fee to extract their willingness to pay. With packaging, the monopolist uses the package size to induce the consumers to buy the efficient quantity and the price to extract their willingness to pay. With regular uniform pricing, the monopolist must use only one instrument to achieve both goals, to induce as efficient consumption as possible and to extract the largest possible share of the surplus. This leads to a trade-off.

18.3. Endogenous Segmentation

In many markets it is difficult for producers to sell their products at different prices in different countries. One reason is that the consumers can travel and buy the good in the low-price country. Another reason is that firms specialized in parallel trading can buy the good in the low-price country and resell it in the high-price country. Arbitrage may also make it impossible to charge different prices from different consumers within the same country. It may be that rich people are willing to pay more than poor people, but if the monopolist cannot tell who is who, the price must be the same. And even if the monopolist could observe some signal of the

customer's income, reselling often makes 1st and 3rd degree price discrimination impossible.

18.3.1. Uniform Pricing

Assume now that the monopolist is unable to distinguish between High and Low customers. When the monopolist is restricted to charge a uniform price, the optimal price must be either to set H_1 to sell one unit, or to set L_1 to sell two units, or to set H_2 to sell three units or to set L_2 to sell four units. Let's assume that the profit is maximized when two units are sold at the price L_1. This is the case if $H_1 > 2 \cdot L_1 - H_1 > c > 3 \cdot H_2 - 2 \cdot L_1 > 4 \cdot L_2 - 3 \cdot H_2$. These inequalities are indeed satisfied by the numerical example. In this case, monopoly pricing leads to a dead weight loss: both High and Low are prevented from consuming a second unit, despite having a willingness to pay for the second unit that exceeds the cost of production. Of particular interest is to note that social welfare (the sum of the monopolist's profit and the consumer surplus) is actually higher with 3rd degree price discrimination than it is with uniform prices. The monopolist sells more when it can price discriminate.

But, the monopolist is sometimes able to do better than this, by adjusting the package size. Let us assume that the monopolist only sells two-unit packages (and no one-unit packages) and charges a uniform per-package price, the same to both types of consumers. The best price would then be either to charge $H_1 + H_2$ to sell a two-unit package to High only, or to charge $L_1 + L_2$ to sell a two-unit package to both consumers. It is easy to see that selling two-unit packages may be better or worse than selling one-unit packages depending on the circumstances. With the numerical example, it turns out to be better to sell two-unit packages at the lower price.

18.3.2. Quantity Discounting – Second Degree Price Discrimination

Absent the ability to distinguish between High and Low, it appears that the monopolist has to choose between different uniform prices and different uniform package sizes. That is, it appears that the monopolist must choose between either selling only two-unit packages at the price $H_1 + H_2$ or $L_1 + L_2$ or only selling single units at the price L_1, L_2, H_1 or H_2 . But, actually the monopolist can sometimes do better than this, by offering the consumers the choice between different package sizes at different unit prices. In particular, the monopolist can offer a quantity discount, by selling single units at a high price and two-unit packages at a low price per unit! The monopolist doesn't know who is who, but by fixing the prices at appropriate levels, the monopolist can make the different consumers voluntarily choose different offers. Those with a high willingness to pay buy large packages at a high price, as long as the per-unit price is low. Those with a low willingness to pay buy small packages at a low price. In order for this to work, it is necessary that the monopolist can prevent people from buying two-unit packages and resell single units.

18.3.2.1. Basic Idea

To analyze quantity discounts it is convenient to introduce the idea of a *contract*. In the market studied here, a contact simply specifies the number of units to be traded, q, and the price, p. A contract can thus be written $c = (q, p)$. Above we have seen several examples of contracts, with different quantities (one or two units) and different prices.

But it was always assumed that the firm only offers a market one type of contract and that all customers in that market had to choose either this contract or no contract at all. But firms can also let the consumers choose from a *menu of contracts*, which is a collection

of different contracts. One should then think of each contract as designed for a certain type of consumers.

Ideally the monopolist would like to offer the consumers the two contracts $c_H = (q_H, p_H) = (2, H_1 + H_2)$ and $c_L = (q_L, p_L) = (2, L_1 + L_2)$ hoping that High chooses c_H and that Low chooses c_L. If the consumers would choose the contracts designed for them, the outcome would be the same as with perfect price discrimination. And, clearly, each consumer would (weakly) prefer to accept the contract designed for its type rather than not trading with the firm at all. A contract satisfying this property is said to be *individually rational*. Still, the monopolist will be disappointed if it offers this particular menu of contracts. The reason is that High will choose the contract intended for Low, which provides the same number of goods but at a lower price. We say that the two offers are not *incentive compatible* when one consumer prefers the contract designed for someone else.

To make the contracts incentive compatible, the monopolist must make c_H more attractive than c_L for High. One option would be to completely give up price discrimination and offer High the same price as Low, i.e. $p_H = p_L = L_1 + L_2$.

An alternative is to make the contract intended for Low less attractive to High. One cannot increase the price offered to Low, without loosing all the sales to Low. But one can reduce the number of units from two to one, if the price is lowered from $L_1 + L_2$ to L_1 at the same time. This change leaves Low's utility from choosing c_L unchanged (equal to zero) but reduces the attractiveness of the same contract to High, since $H_1 - L_1 < H_1 + H_2 - L_1 - L_2$.

By offering $c_L = (1, L_1)$ the monopolist looses a part of its profits from dealing with Low, namely $L_2 - c$. The advantage is that High's price can be increased. In order for High to prefer $c_H = (2, p_H)$ over $c_L = (1, L_1)$ it is require that $H_1 + H_2 - p_H \geq H_1 - L_1$, allowing the monopolist to charge $p_H = L_1 + H_2$. Hence, by reducing q_L by one unit the monopolist can increase p_H from $L_1 + L_2$ to $L_1 + H_2$.

This solution means that the monopolist offers a quantity discount: consumers are offered to buy a first unit at the price L_1 and a second unit at the price $H_2 < L_1$. Such a quantity discount implies price discrimination: High's average price for the two-unit bundle is lower than Low's price for one unit.

Note that Low is left without a consumer surplus. High, on the other hand, is granted as surplus of

$$(H_1 + H_2) - (L_1 + H_2) = H_1 - L_1 > 0.$$

This result is quite general. When consumers have private information (about their willingness to pay), firms are not able to extract the whole social surplus. It is the customers who have an incentive to disguise as some other type that must be left with some surplus in order to discourage such mimicry. In the example here it is High that have an incentive to mimic Low and consequently it is High that has to be left with a surplus.

Also note that High is allowed to consume two units, the socially efficient quantity, while Low only gets one unit, which is suboptimal. By selling fewer units to Low, Low's price per unit can be increased from $(L_1 + L_2)/2$ to L_2. This is not the reason, however. Rather, the reason is that the reduction in q_L by one unit allows the monopolist to increase p_H from $L_1 + L_2$ to $L_1 + H_2$. Also this result is quite general. If other groups of consumers have an incentive to disguise as a certain type of consumer, firms will offer this type of consumer less attractive contracts.

18.3.2.2. Optimality

The pricing schedule with a quantity discount that was just derived is actually the best pricing schedule (with quantity discount) that the monopolist could construct. To see this let p_1 be the price the monopolist charges for one unit and p_2 the price for a two-unit package.

Participation constraints In order for High to prefer the two-unit package over not buying anything it is necessary that $H_1 + H_2 - p_2 \geq 0$. This is referred to as High's participation constraint or High's individual rationality constraint. Similarly, Low's individual rationality constraint is that $L_1 - p_1 \geq 0$.

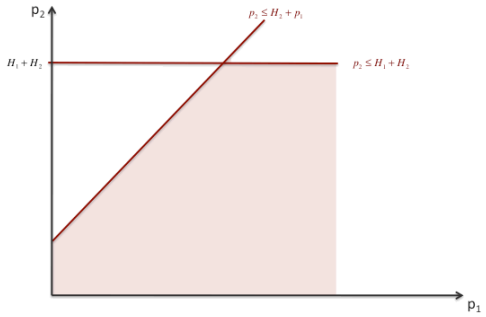
Incentive compatibility constraints In order for High to prefer the two-unit package to the one-unit package it is necessary that $H_1 + H_2 - p_2 \geq H_1 - p_1$. This is referred to as High's incentive compatibility constraint. Similarly, Low's incentive compatibility constraint requires that $L_1 - p_1 \geq L_1 + L_2 - p_2$.

Choice set So, for High to buy two units it is required that

$$\begin{aligned} IR_H: &\quad p_2 \leq H_1 + H_2 \\ IC_H: &\quad p_2 \leq H_2 + p_1 \end{aligned}$$

In Figure 18.5, High's individual rationality constraint is the hori-

Figure 18.5.: High



zontal line, which requires p_2 to be sufficiently low. High's incentive compatibility constraint is the upward sloping line, which requires

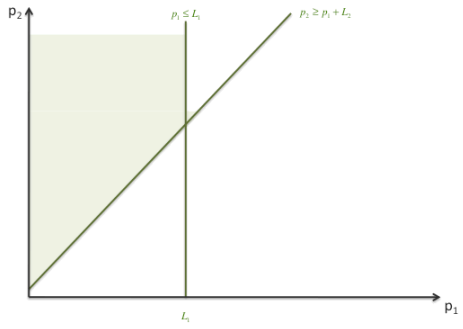
p_2 to be sufficiently low compared to p_1. To satisfy both these restrictions, the pair of prices must be in the shaded area below the two lines.

For Low to buy one unit it is required that

$$\begin{aligned} IR_L: &\quad p_1 \leq L_1 \\ IC_L: &\quad p_1 \leq -L_2 + p_2 \end{aligned}$$

In Figure 18.6, Low's individual rationality constraint is the ver-

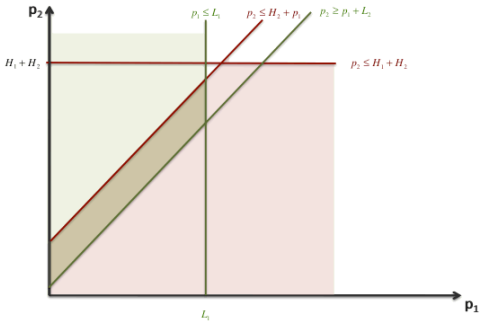
Figure 18.6.: Low



tical line requiring p1 to be sufficiently low. Low's incentive compatibility constraint is the upward sloping line requiring p1 to be sufficiently low relative to p2 . To satisfy both these restrictions, the pair of prices must be in the shaded area to the left of the two lines.

The monopolist's choice set (the pair of prices satisfying all the four constraints) is described by the brownish area in Figure 18.7. Note that if the one-unit offer is individually rational for Low ($L_1 - p_1 \geq 0$) and if the two-unit bundle is incentive compatible for High ($H_1 + H_2 - p_2 \geq H_1 - p_1$), it automatically follows that the two-unit bundle must be individually rational for High since $H_1 + H_2 - p_2 \geq$

Figure 18.7.: Choice set



$H_1 - p_1 > L_1 - p_1 \geq 0$. Hence we need not bother with High's individual rationality constraint. In Figure 18.7, this is visualized by the fact that any pair of prices in the choice set is clearly below the red horizontal line.

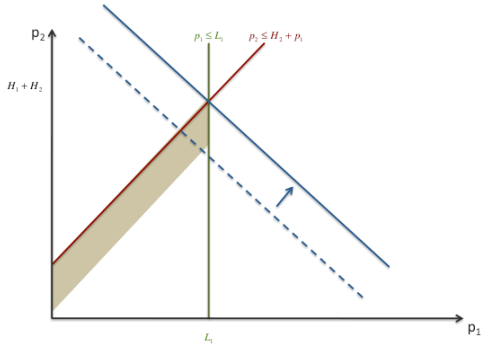
Maximization Note also that the monopolist always wants to charge the highest possible prices in the choice set. More precisely, the monopolist wants to maximize its revenues, which are given by $p_1 + p_2$. An iso-revenue curve is a straight line with slope minus one. Higher revenues are associated with a line further away from the origin. In Figure 18.8, the optimal price-pair is thus given by the intersection of High's incentive compatibility constraint and Low's individual rationality constraint.

The optimal prices are given by

$$\begin{aligned} p_1 &= L_1 \\ p_2 &= L_1 + H_2. \end{aligned}$$

The final step for the monopolist is to compare the best outcome with price discrimination with the best outcome under uniform pric-

Figure 18.8.: Choice set



ing. The comparison can go either way, but with the numerical example the monopolist prefers 2nd degree price discrimination.

18.3.2.3. Implementing the quantity discounts using a menu of pricing plans

The analysis above suggest that the monopolist should offer consumers the choice between buying one unit for L_1 or two units for $L_1 + H_2$. High will then buy two units and Low will buy one unit. The same outcome can, however, be achieved by offering a menu of different pricing plans, with different fixed fees and usage fees.

Let us again think about Telia and the mobile phone market, now in the light of the simple model with two consumers. Let us assume that Telia offers consumers two pricing plans, one with a high monthly fee but a low price per calls, the other with no monthly fee but a high price per call.

Once consumers have subscribed to a pricing plan, they are allowed to make any number of calls they wish. Assume now that High has chosen to subscribe to pricing plan H and paid the monthly fee. He would then make two calls. The reason is that High's will-

Table 18.1.: A simplified version of Telia's menu

	Plan H	Plan L
Monthly fee	$L_1 + H_2 - 2 \cdot c$	0
Price per call	c	L_1

ingness to pay for both the first and the second call is higher than the usage fee (which is equal to marginal cost). In this case High's utility would be $(H_1 + H_2) - (L_1 + H_2 - 2 \cdot c) - 2 \cdot c = H_1 - L_1 > 0$. The first term on the left hand side is High's consumption value. The second term is the monthly fee and the third is the sum of all usage fees paid for consuming the two calls.

Assume now that High has chosen to subscribe to pricing plan L. He would then only make one call. The reason is that High's willingness to pay for the second call is lower than the high price Telia charges for calls. In this case High's utility would be $H_1 - L_1$. This means High will choose plan H. Pricing plan H is both individually rational and incentive compatible for High. A similar analysis would reveal that Low would choose pricing plan L and make only one call.

The conclusion is that quantity discounts can be implemented using a menu of pricing plans, allowing consumers to choose how much to buy, instead of a menu of packages with predetermined number of units. In the simple model considered here, the two options are equivalent.

Also note that consumers with a high willingness to pay choose pricing plans with a high monthly fee and a low charge per call, while consumers with a low willingness to pay choose low monthly fees and high prices per call. The simple model here thus provides a possible interpretation of at least some of the complexities in Telia's pricing behavior in the mobile phone market.

18.4. More Examples of Screening

Second degree price discrimination is just an example of a much more general principle, called *screening*. Screening is used by a party, often referred to as the principal (here the firm), who has incomplete information about another party, often referred to as the agent (here the customers). The idea is to let the informed party choose from a set of well-defined alternatives, constructed in such a way that each type voluntarily chooses the alternative designed for him.

Another example is quality discrimination, by reducing the quality of goods offered to consumers with a low willingness to pay. Such practices are normally referred to as *versioning*. It is common to sell computer programs with some functions disabled at a reduced price. If the quality is reduced sufficiently, consumers with a high willingness to pay will continue to buy the full-featured version. Another example of versioning is when the French railway operator dismounted the roofs from 3rd class carriages. The reason was not to make life difficult to 3rd class customers, but to discourage the 2nd class customers to switch. Yet another example may be the services airline companies offer first class passengers but not in economy class.

Buyers of insurance often know more about their personal risks of incurring a loss than the insurance companies do. Low-risk customers wish to distinguish themselves from high-risk applicants in order to obtain insurance coverage at terms more favorable than would be available if they were pooled with the high risks. The desired separation is attained through a screening mechanism. The insurance companies offer customers a menu of contracts. One contract provides high risks with full insurance, while the other contract incorporates a lower average price for coverage but also includes a deductible that is unacceptable to high risks.

Similarly, in credit markets, borrowers may get different condi-

Figure 18.9.: Can you spot any discrimination in this advertisement?

tions depending on how much of their own money they are willing to put at risk in their business projects. An entrepreneur willing to risk his whole fortune reveals that he has very favorable information about the prospects. Such an entrepreneur will be offered partial funding at a favorable interest rate. Entrepreneurs requiring full funding (thus not taking any risks) will have to pay a much higher interest rate, if they get a loan at all. The problem is that a high interest rate will increase the risk of failure even further.

Part VII.

Competition Policy

19. What is a market?

In ordinary language we often use the term market to refer to an industry or even a whole sector of the economy. We may for example be talking about the "car market," certainly including both Volvo and Volkswagen and perhaps also Tata Motors and Lamborghini. Sometimes the market may be the world market, sometimes only the national market is included.

In microeconomics on the other hand each good is thought of as a separate market and since goods are distinguished not only by their physical characteristics but also by their exact location, time of delivery and the "state of nature," each market is extremely narrow. We would for example distinguish between different Volvo S80 markets depending on which engine the car is equipped with and depending on the hour of delivery.

In applied industrial economics it is often natural to define markets more narrowly than in casual conversations but also to be more inclusive than suggested by the idealized notion of a market in microeconomics. The role of a market definition is to set the scene for the discussions to come: exactly which goods are we talking about and which buyers and sellers do we want to include in the discussion?

Most often we want to include all goods and firms that some set of consumers consider as close competitors. Clearly, a Volvo S80 with 140 horsepowers and one with 170 horsepowers are close substitutes; the two specifications can be said to compete with each other. But are Volvo and Lamborghini part of the same market?

A systematic approach to market definition The most systematic approach to think about market definitions has been developed not by academic economists but in competition policy. So this is the approach that we are going to follow here.[1] In competition policy markets are defined with a very particular purpose in mind, namely to assess how much market power a certain firm or a certain group of firms has. Market power is often measured by the firm's ability to charge a price above marginal cost without loosing all its customers. To delineate a firm's market is therefore to start investigating the constraints on the firm's pricing decisions. Expressed differently, to define markets is a first step to describe how competition works in some part of the economy.

People more or less explicitly define markets also in many other situations, such as when economists are doing empirical studies or when a financial analyst wants to assess a firm's value. Business managers need to think in terms of markets when they make pricing and investment decisions. I am convinced that the competition policy approach to market definitions will often prove useful also in such situations.

As we will see, the competition policy approach has many merits, but it is also plagued by many practical and conceptual problems.

An important insight The microeconomic approach to market definition focuses on how substitutable consumers consider different goods to be. Goods that consumers find (very) good substitutes are part of the same market, while less substitutable goods are considered part of different markets. In short, market definitions are solely based on the cross price-elasticities of demand. Such demand substitution is also the starting point for the competition

[1] For more information, the interested reader is referred to the European Commission's Notice on market definitions (97/C 372/03), the Horizontal Merger Guidelines issued by the U.S. Department of Justice and the Federal Trade Commission and the U.K. Office of Fair Trading's Competition Law Guidelines.

policy approach, but as we soon will see also many other aspects are considered important, eg so-called supply substitution, the opportunities for price discrimination, substitution chains and substitution further down the value chain when defining intermediate goods markets.

19.1. Substitution

In competition policy a market is thought of as a collection of goods worth monopolizing. The goods within a market defined this way are relatively close substitutes and compete for customers. But they do not compete much with goods outside the market. To identify such a collection of goods, one needs to think in terms of substitution possibilities.

19.1.1. Demand substitution

The key disciplinary force on firms' pricing is demand substitution. Volvo would have little real influence over its own price if its customers can easily switch to other cars such as Saab. Therefore, identifying a firm's market is to identify the best alternative products and the best alternative suppliers.

One way to determine the range of products that the consumers view as substitutes is to make the following thought experiment.

We start out with some market segment as a candidate for a market definition. Let us say that Volvo V70 is our candidate.

Assume now that Volvo permanently increases the price of Volvo V70 by a small amount, while all other prices remain constant. What are the likely reactions of the customers to such a price increase? Probably some of the customers would switch to Saab 9-5 and other cars. If sufficiently many customers switch, the price increase will be unprofitable as a result of the lost sales. Then, the market segment V70 does not constitute a market in itself, since

Saab 9-5 and the other cars to which the customers switch restrain the pricing of V70.

We then select the car that would attract most customers in case of an increased price on V70, say Saab 9-5, and conclude that at least this product is on the same market as V70. Our new candidate market definition consists of two cars, Volvo V70 and Saab 9-5.

The next step of the thought experiment is to permanently increase the price of both Volvo V70 and Saab 9-5 by a small amount at the same time, while all other prices remain constant. Now the customers may switch from V70 and 9-5 to other cars including Volkswagen Passat to such an extent that the price increase in unprofitable. If Passat catches most of the switching customers, we will say that also Passat is on the same market.

If a hypothetical simultaneous increase in the price of all three cars would be profitable, the process stops. The market is defined to consist of Volvo V70, Saab 9-5 and Volkswagen Passat. The consumers consider these three cars to be so similar that it would be impossible to profitably increase the price of only one of them. In contrast, a simultaneous increase in the price of these three cars would not make consumers switch to other cars since they are sufficiently different. It would be profitable to monopolize a market consisting of these three cars.

Since the thought experiment involves a small but significant nontransitory price increase, it is usually referred to as the SSNIP-test. It is alternatively referred to as the "hypothetical monopolist test" since the essence of the test is to establish the smallest range of goods such that a hypothetical monopolist controlling that group of products could profitably sustain supra-competitive prices.

19.1.2. Supply substitution

Another disciplinary force on a firm's pricing is supply substitution. If a firm would attempt to increase its prices permanently, rival

firms may change the specification of their products or extend their sales into neighboring geographical areas to catch the customers that want to leave the now more expensive supplier.[2]

Example: Paper For instance, paper is supplied in a range of different qualities, from standard writing paper to high quality papers used in art books. Different qualities of paper are not good substitutes from the customers' point of view: Art books will not be printed on standard writing paper and newspapers will not be printed on thick and glossy paper.

Many paper plants, on the other hand, can manufacture many of the different qualities. Production can be adjusted with negligible costs and within a short time span. Paper manufacturers are therefore able to compete for orders for many different qualities, at least if orders are placed with sufficient lead time to allow for modification of production plans. Under such circumstances, the various qualities of paper may be thought of as part of the same market.

SSNIP-test To analyze the extent of such supply substitution, one can use a similar thought experiment as before. In response to the hypothetical permanent increase in the price of a certain product, firms currently not active on the market (as defined by demand substitution) may be able to switch production and start marketing products that the customers view as good substitutes. If it is profitable for a certain rival firm to do so, and if customers would switch to the rival's product to such an extent that the price increase is unprofitable, then also this rival is on the same market.

Note that for the rival to be on the same market, the rival must

[2]The European Commission does take supply substitution into account when defining markets. The U.S. Department of Justice and the Federal Trade Commission do not consider supply substitution when defining markets. They consider supply substitution as a later stage of their competitive analysis.

be able to make the switch sufficiently quickly. Otherwise, the price increase will be profitable after all.

Supply substitution vs. entry Also firms that may start to market substitutable goods some time after a price has been permanently increased may have a disciplinary effect on pricing. The longer it takes to switch the smaller is the effect, however. At some point one has to draw the line and argue that some firms are not active on the relevant market.

This does not mean, however, that competition authorities would ignore these firms when assessing the competitive conditions in the market. This would, however, be phrased as future entry of new firms. The exact distinction between supply substitution within a market and entry is of course arbitrary.

19.2. Dimensions of market definition

A market is typically thought of as having a product dimension and a geographical dimension. The definition of the product market is a description of the physical characteristics of the products included in the market. The definition of the geographical market is a list of delivery points.

19.2.1. Product dimension

Horizontal differentiation Coke and Pepsi contain similar ingredients; they also lock, smell and taste similar. More precisely, Coke and Pepsi are examples of two horizontally differentiated products. They are somewhat different and some consumers prefer Coke and others prefer Pepsi. If consumers consider them sufficiently similar, they are part of the same market. Otherwise not.

In fact, Coke and Pepsi may be so similar that it could be reasonable to presume that they compete in the same product market,

even absent detailed information about the customers' substitution behavior.

Pure intuition may sometimes be misleading, however. Most people would say that water is water and that different brands of mineral water have even more similar characteristics than Coke and Pepsi. Still, the fact that mineral water is much more expensive than tap water and that some brands of mineral water sell at a much higher price than other brands, indicate that consumers do not find all water substitutable. And it is the consumers' substitution behavior that is the basis for market definitions. It is likely that tap water and mineral water are considered separate markets, even in Sweden where the quality of tap water is very high. Different mineral waters may be on the same market, however.

Vertical differentiation In some cases most consumers would agree that one of the products is superior to the other. Most people would for example agree that BMW and Mercedes are better cars than, say, Volkswagen and Opel. We are then talking about vertical product differentiation. If the consumers consider the quality differences to be relatively important, two different markets must be defined, one for regular cars and one for luxury cars.

Useful evidence One may often come up with a reasonable candidate definition of the market based on product characteristics. But a number of quantitative tests can also be used as input in a market definition. The most straight forward evidence would be econometric studies of demand to estimate price and cross-price elasticities. Other tests are based on the similarity of price movements over time, the analysis of causality between price series and similarity of price levels and/or their convergence.

19.2.2. Geographical dimension

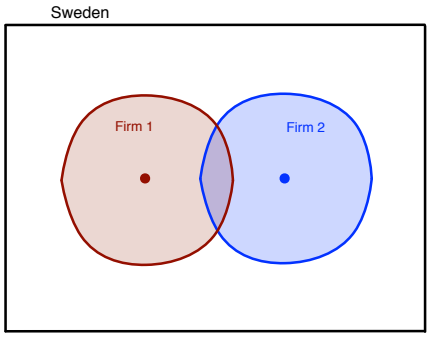
The geographical market is list of delivery points. And the basic issue is the same as when defining a product market. If two different firms sell the same good at two different locations they are on the same geographical market if the customers consider delivery at these two points good substitutes. They are on the same market if the customers would switch from one location to the other as a result of a small permanent price increase. If customers would switch, price increases would not be profitable.

Supply substitution may also be important for geographical market definitions. A firm may currently not be active as a supplier in a certain area. But if this firm could easily switch its production capacity from a nearby area to the relevant area, this firm would be considered to be part of the relevant market.

Transportation costs An important determinant of the extent of the geographical market is transportation costs. For example, asphalt firms' abilities to supply asphalt are severely limited by the location of their production facilities since asphalt can only be transported approximately 120 kilometers. This means that a customer who wants a job done at a certain location can usually only choose from a small set of nearby firms. Reversely, when we want to define the relevant geographical market of a certain asphalt firm we will typically find that there are only a few rivals that can deliver at the same locations. Only the firms that can, are active on the relevant geographical market.

In the figure, Sweden is assumed to have two firms with one production plant each. The geographical area that can be supplied by each firm is a circle around each plant. If we are interested in defining firm 1:s geographical market, we would need to conclude that there are two different relevant markets: one in which it is a monopolist (the light red area) and one in which it has competition

Figure 19.1.: Asphalt



from firm 2 (the darker red/blue area). The reason for this definition is that the consumers living in the dark gray area consider the two firms substitutes, while the customers living in the light gray area do not consider the second firm an alternative to firm 1.

National borders Another important aspect of geographical market definition is country borders. It may be difficult for both customers and sellers to transfer goods from one country to another. This may for example have to do with tariffs, import quotas, safety regulations and language differences. In fact, according to recent empirical studies, there is much more trade within national borders, e.g. between different Canadian provinces, than across borders, e.g. trade between Canadian provinces and US states.

19.2.3. Other dimensions

The product dimension and the geographical dimensions are in practice the most important aspects of market definition, at least in competition policy cases. In principle, however, also other dimen-

sions could be important. One example is the *time of delivery*. A seat on an airplane from Stockholm to Brussels at 7 in the morning is a very different good from a seat on a plane on the same route at 7 in the evening. There may also be reason to distinguish between working days and weekends as well as between the summer season and the winter season.

Another example is the *state of nature*. In Sweden, many customers need more electricity when the temperature happens to be low. It is also difficult for electricity generating firms to switch supply between different points in time and between different states of nature since electricity is not easily storable.

For a similar reason contracts written at different points in time may from an economic point of view be considered different goods even if the time of delivery is the same. A well-known example is that it may be substantially cheaper to book a flight in good time before departure than it is to book it the same day. The main difference is that both the buyer and the seller will have different amounts of information at different points in time. If people are uncertain about the future cost and demand conditions and if this uncertainty is expected to be resolved between two different dates, contracting at those two dates will not be considered good substitutes by the buyers and by the sellers. The two different contracts will entail different distribution of risk between the buyers and the sellers.

19.3. Topics

19.3.1. Uniform pricing vs. price discrimination

In a famous competition policy case, called *United Brands*, the European Court of Justice considered bananas to constitute a separate market from other fruit. The reason was that babies and old people without teeth cannot switch to other sorts of fruit if the price of

bananas would increase.

This decision has later been criticized and even ridiculed. Although many customers would not be able to avoid paying the higher prices, most people have teeth and can switch to apples and pears. Thus, even if a single firm would supply all bananas, this monopolist would not be able to profitably increase the price of bananas in a shameless effort to exploit the toothless.

This critique hinges crucially on the fact that the banana suppliers cannot charge different prices from customers with and without teeth. If they could charge different prices, a hypothetical banana monopolist would be able to profitably increase the price of bananas sold to the toothless. It would then be correct to define a separate market for bananas sold to the toothless.

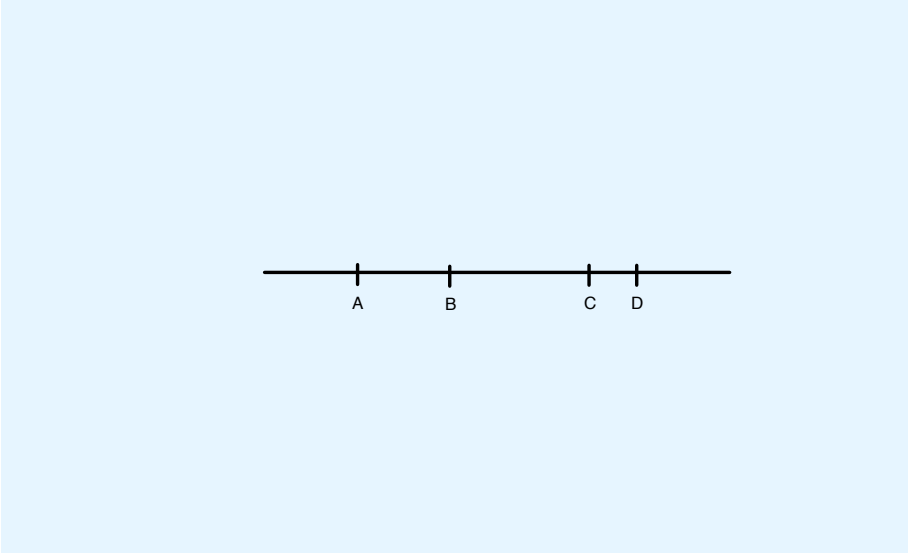
Also recall the above discussion about asphalt. It was implicitly assumed that the firms can charge different prices to different consumers depending on location. Firm 1 could charge a lower price from the customers in the dark gray area of the figure in order to meet the competition from Firm 2, and a high price from the consumers in the light gray area where there is no competition. Absent this possibility, there would only be one relevant market, consisting of the whole geographical area where asphalt is supplied.

19.3.2. Indirect vs. direct competition

Two products need not compete directly with each other to be on the same market. There may exist a third product "lying in-between" and competing with the first two products making it natural to define the market to consist of all three products. We then talk about a chain of substitution.

An example may be food retailing. Three convenience stores may be located along the same street. Some customers live between stores A and B and choose between those two. Other customers live between stores B and C, and so on.

Figure 19.2.: Indirect competition



Even if no customers choose between stores A and C, they may still constrain each other's pricing decisions and therefore be on the same market. To see this, suppose that A would reduce its price. Then B would have to reduce its price in order not to loose its customers living on the street between A and B. But, then also the customers living between B and C would switch from C to B, forcing C to lower its price as well.

Another way to describe this is to use the same market definition procedure as above. This procedure would first conclude that A and B are on the same market since it would be unprofitable for A to permanently increase its price by a small amount, as customers would then switch to B. Increasing the price of A and B at the same time would also be unprofitable, assuming that sufficiently many customers would switch from B to C.

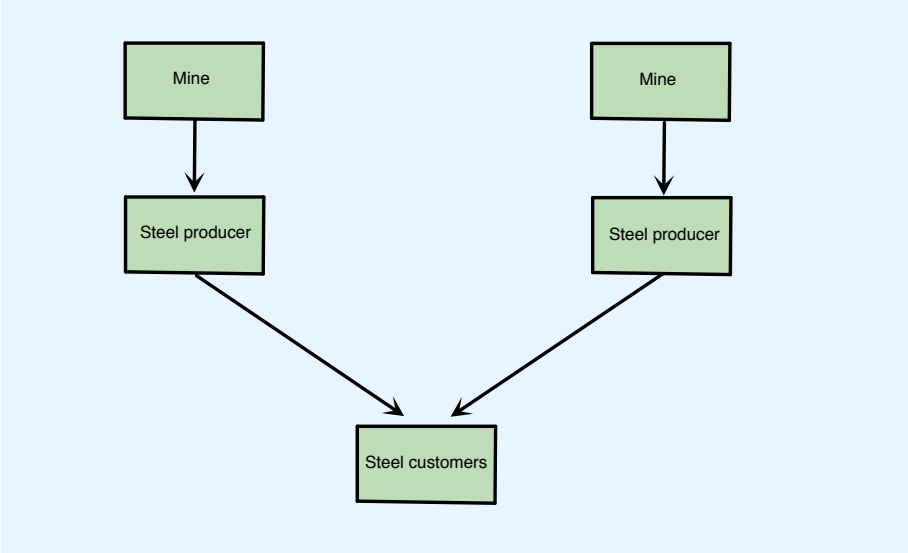
If we are interested in defining firm A:s market, we will sooner or later reach a limit when additional stores will not be included to the market. Increasing the price of A, B and C at same time may be profitable even if C will loose some customers to D. The

reason is that only a few customers, those who live between C and D have the possibility to avoid paying the higher price. Those who live between A and C are stuck with the higher prices.

19.3.3. Wholesale vs. retail markets

Producers of steel normally buy their iron ore from mines relatively close by. It would often make little sense to export a relatively homogenous input of low value-to-volume between different regions. Does this mean that the relevant geographical markets for mining are narrow? Not necessarily.

Figure 19.3.



The reason is that steel, which has a higher value-to-volume, competes over larger geographical areas. A rise in the price of ore in a certain region may therefore reduce the ability of the steel producers in that region to compete in the world market. As a result of demand substitution at the next level of the value chain, also input markets may be fairly wide.

19.3.4. Current price level vs. competitive price level

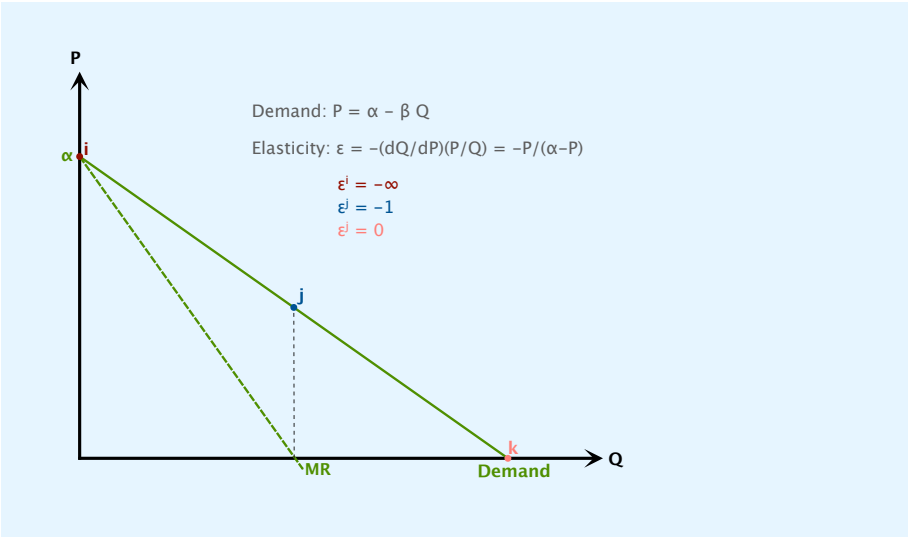
In a famous American antitrust case, the US Government argued that DuPont was a monopoly producer of cellophane. The firm, on the other hand, argued that they had only a modest share of the more broadly defined market for "flexible wrappings." They argued that a price increase would make their customers switch to e.g. aluminum. Which is the correct market definition, cellophane or flexible wrappings? With the first definition DuPont would be thought of as a monopolist; with the second as a firm under competition.

To substantiate their claims, DuPont cited econometric evidence that suggested that DuPont's demand was rather elastic. The court accepted DuPont's wider market definition, thereby committing what is now known as the *cellophane fallacy*. The court failed to see that DuPont had already increased its price to a level above cost. At this high price level, further increases would be unprofitable as a result of customer switching.

The lesson is that monopoly firms, or firms with market power in general, will increase price as long as the gain in terms of a higher markup will outweigh the loss in terms of lost customers. The higher the price is, the lower are sales, and therefore the gain from an increased markup will be lower. At some point further price increases will therefore not be profitable. Even a monopoly firm will set price in such a way that it operates on the elastic part of its demand curve.

Consider the market depicted in the figure below. Clearly the price elasticity varies along the demand curve. In fact, with a linear demand the elasticity varies between zero and minus infinity. A monopoly firm chooses to operate at point M on the demand curve, i.e. on the elastic part. It may thus appear as if the firm is under tough competition at the prevailing price. If we instead start at the

Figure 19.4.



cost level, i.e. the competitive price level, and study whether price increases would be profitable to the firm, the conclusion would be different. Cellophane would then be considered a market worth monopolizing, and DuPont would have been considered a monopolist.

Exercise 60. Compute the elasticity of demand at points M and C and compare.

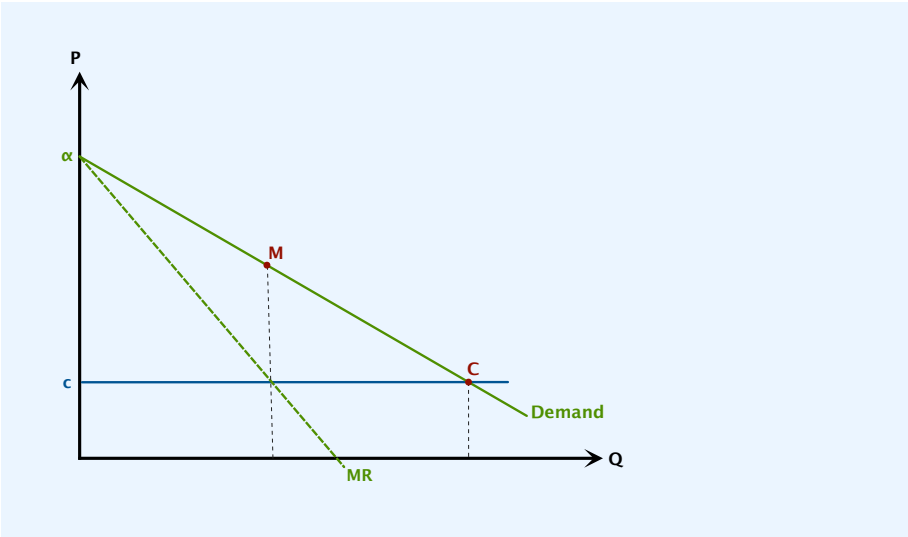
19.3.5. After markets

Should one define separate markets for razor blades or should razor blades be considered part of a system consisting of razors and blades?

When pricing razor blades, a company like Gillette must distinguish between captive and non-captive consumers. Some people have already bought Gillette razors. These customers also have to buy the blades from the same company; other blades wouldn't fit. A small increase in the price of blades would not make them switch to other brands.

Other customers have not yet bought their razors. They may

Figure 19.5.



choose between different razors primarily based on the price of the razor itself, or they may be more sophisticated and consider the lifetime cost of the system. Increasing the price of razor blades may induce customers to start buying competing companies' razors and blades.

The correct definition must thus depend on the frequency with which people buy new razors and the extent to which they consider the lifetime cost.

20. Cartels

Competition forces firms to charge prices below the monopoly level. In the extreme case of the Bertrand trap, prices are forced down to marginal cost even if there are only two firms competing in the market. Since competition is ruinous, why don't firms cooperate instead?

A cartel is an agreement between should-be competitors to cooperate. Oligopoly firms may agree to charge prices that are higher than the ones they would set when they make their choices independently of each other (*i.e.* when they compete). They may for instance agree to all charge the monopoly price and to share the market equally between themselves. As a result of the higher prices, wealth is transferred from the customers to the firms. And since the firms are only able to increase prices by reducing sales, this transfer of wealth is achieved by reducing the efficiency of the market. They create or increase the dead weight loss.

Luckily, however, cartels agreements are difficult to negotiate and they are difficult to enforce. While all the firms may have an incentive to cooperate, each firm will also have an incentive to cheat on the agreement, to undercut the competitors' prices and poach their customers. An important issue is then if and how the cartel members can enforce their agreements.

The normal procedure when business partners don't fulfill their contractual obligation is to sue them (or at least threaten to do so). But cartel agreements are illegal according to the competition rules (e.g. Article 101 of the European treaty). The reason is of course that cartels by reducing competition harm the customers

and reduce market efficiency. Firms can therefore not take those who break a cartel agreement to court.

Article 101 of the Treaty on the Functioning of the European Union:

1. The following shall be prohibited as incompatible with the common market: all agreements between undertakings, decisions by associations of undertakings and concerted practices which may affect trade between Member States and which have as their object or effect the prevention, restriction or distortion of competition within the common market, and in particular those which:
   a) directly or indirectly fix purchase or selling prices or any other trading conditions;
   b) limit or control production, markets, technical development, or investment;
   c) share markets or sources of supply;
   d) apply dissimilar conditions to equivalent transactions with other trading parties, thereby placing them at a competitive disadvantage;
   e) make the conclusion of contracts subject to acceptance by the other parties of supplementary obligations which, by their nature or according to commercial usage, have no connection with the subject of such contracts.
2. Any agreements or decisions prohibited pursuant to this article shall be automatically void.

Instead, the cartelists must enforce their agreement themselves. One way would be for the cartelists to threaten cheaters with retaliation in the future. Retaliation may e.g. take the form of a price war. Or, expressed differently, the cartelists may promise co-operators continued cooperation in the future.

The oligopoly models discussed so far presume the market to exist only one period. Such static models can therefore not be used to study cartel formation. Rather, we need to acknowledge the fact that, in reality, most oligopolistic interaction takes place over long time spans and that firms have the possibility to react to each other's past behavior. This ability to react on each other's past behavior can change the competitive situation dramatically. Firms might be able to punish cheaters with price war and reward co-operators with continued cooperation. But threats and promises must be credible to have an effect. Firms must have an incentive to start a price war if a deviation should occur and they must have an incentive to continue previously uninterrupted cooperation.

This lecture will study the issue of credibility in some detail. The main questions are (i) if it is possible for oligopoly firms to form a stable cartel, (ii) what the conditions for this are, and (iii) how to do it, i.e. what strategies to use?

20.1. How a cartel can enforce its agreement

The easiest way to model repeated interaction is to build a so-called repeated game which, as the name suggests, simply is a game where the same oligopoly game is repeated several periods. A repeated oligopoly game is defined by the following assumptions:

- The market exists for some number of time periods $t = 1, 2, 3, ..., T$.
- Each period is described by the same "stage game." The stage

game may *e.g.* be the static Bertrand duopoly game or some other static oligopoly model that describes the market well. Let us stick to the Bertrand model. That is, in each period the two firms simultaneously set their prices, p_1^t and p_2^t, and then the customers buy from the firm with the lowest prices. The firms earn profits $\pi_1\left(p_1^t, p_2^t\right)$ and $\pi_2\left(p_1^t, p_2^t\right)$ in period t.

- In any time period, the firms can observe all previous prices charged by the competitor.
- A firm's payoff is the sum of discounted profits. That is, firm i's payoff is given by $\Pi_i = \sum_{t=1}^{T} \delta^{t-1} \cdot \pi_i\left(p_1^t, p_2^t\right)$ where $\delta < 1$ is the discount factor, which is assumed to be the same for both firms.

We have now defined an extensive form game. The next question is what the subgame perfect equilibria (SPE) of the game are?

Recall that an equilibrium is a collection of strategies, one for each player, and that a strategy is a complete plan of action for a player. A strategy for firm i thus specifies firm i's pricing in every period. Moreover, a strategy does not only allow a firm to charge different prices in different time periods. It also allows the firm to charge different prices in any time period depending on what has happened in the past.

Let me elaborate on the last point. It is assumed that the firms can observe everything that happened in the market in all earlier periods. In particular, the firms are able to observe their rival's past prices. The firms can therefore make their price decision at time t contingent on the prices set in all previous periods. A firm may for example charge a high price in period 10 if the competitor charged a high price in period 9 and a low price otherwise. That is, the prices in period t can be made contingent on the *history* of play.

Expressed differently, we can define several different decision nodes

for every period ($t > 1$), with each decision node corresponding to a particular history of price choices. A firm will thus be at different decision nodes in period 2 depending on whether the rival charged e.g. €5.50 or €5.70 in period 1. Likewise, the firm will be at different decision nodes in period 2 depending on the firm's own prices in the previous period. Recall that a strategy may prescribe different prices for different nodes since the player knows which node he is at.

There may not be any so-to-say exogenous reasons to set different prices in different time periods. If cost and demand conditions are stable over time, it may appear natural to set the same prices in all time periods. But if one firm, say A, for some yet unexplained reason, decides to charge different prices depending on what has happened in the past, then it may make sense for a competitor, say B, to also charge different prices depending on what has happened in the past. And if B does it, also A may after all have a reason to do it.

20.1.1. Finitely Repeated Games

For simplicity, let us start out analyzing the twice repeated Bertrand duopoly. That is, assume, $T = 2$. To find the SPE we need to find the subgames. Actually, in period 2, there is a subgame for every possible history. Expressed differently every pair of first-period prices (p_1^1, p_2^1) defines a possible history and therefore a separate subgame. In addition, the full game is also a subgame.

However, when period 2 arrives, independent of what happened in the first period, the situation is exactly the same as in the static Bertrand model discussed above. Therefore, the Nash equilibrium of every period-2 subgame requires the firms to set prices equal to marginal cost, *i.e.* $(p_1^2, p_2^2) = (c, c)$.

Now, knowing that prices will be equal to marginal cost in the second period, independent of the prices in period 1, also period 1

becomes equivalent to the static Bertrand game. The firms must set prices equal to marginal cost, *i.e.* $(p_1^1, p_2^1) = (c, c)$. Intuitively, the firms will not be kind to each other in period 1 since they know that they will not be rewarded in period 2.

To sum up, there exists a unique SPE in the two-period game. The equilibrium strategy requires firm i to set price equal to marginal cost in the first period and to set price equal to marginal cost in the second period, independent of the prices charged in the first period. The outcome is that both firms charge price equal to cost in all periods.

The same result holds if the Bertrand stage-game is repeated three times ($T = 3$). Prices must be set equal to cost in every subgame starting in the third period. Prices must also be set equal to cost in the second period, independent of what happened in the first period. Finally, prices must be set equal to cost in the first period.

The same result holds if the Bertrand stage-game is repeated a finite number of times. In fact, any finitely repeated game, with a stage game that has a unique Nash equilibrium, has a unique SPE, namely the Nash equilibrium repeated in each stage.

20.1.2. Infinitely Repeated Games

In a *super-game*, the stage game is repeated infinitely many times. When the game is repeated without an end, we cannot use backwards induction to find the subgame perfect equilibria (SPE). There is no last period to start from. The logic of the previous section does not apply.

To identify the SPE:s in super-games, we need to work in a completely different way. We need to (i) guess what the equilibrium strategies look like and then (ii) prove that these strategies indeed form an equilibrium.

As it turns out, there are many SPE:s in super-games. We will

not try to find all of them. Instead, we will focus exclusively on the key issue here. The key issue is whether there exist any strategies which sustain cooperation between the firms as a SPE?

The trigger strategy Consider the following strategy for firm i:

- In period 1, set the monopoly price, i.e. $p_i^1 = p^m$.
- In period $t > 1$, set the monopoly price, *i.e.* $p_i^t = p^m$, provided that both firms have charged the monopoly price in all previous periods.
- In period $t > 1$, set price equal to marginal cost, *i.e.* $p_i^t = c$, if any firm has deviated from the monopoly price in any previous period.

Notice that these three instructions form a strategy, since they describe what firm i should do whenever it has to act. It is a complete plan of action.

Note that if *both* firms, for some reason, follow the trigger strategy, then both firms will set the monopoly price in all periods. They will start out charging the monopoly price. And since both charged the monopoly price in period 1, they will also charge the monopoly price in period 2. And since both charged the monopoly price in periods 1 and 2, they will also charge the monopoly price in period 3. And, so on.

However, if a firm would deviate from the trigger strategy once, then the trigger strategy prescribes a firm to set price equal to marginal cost forever after. An eternal price war breaks out.

How to prove that the trigger strategy is a subgame perfect equilibrium To prove that both firms adopting the trigger strategy is a SPE, we need to show that no firm has an incentive to deviate from the trigger strategy, given that the other firm sticks

to the trigger strategy. That is, we need to show that the trigger strategy is a best reply to itself.

To show that a firm does not have an incentive to deviate from the trigger strategy, one has to show that the firm does not have an incentive to deviate in any subgame. This may sound as a formidable task since there are so many different subgames, infinitely many actually. Already in period 2 the firms may find themselves in the beginning of many different subgames, depending on exactly what prices the they charged in the first period.

Fortunately, the trigger strategy only distinguishes between two types of histories, implying that the firms are either in a cooperative phase (the two first contingencies or bullet points) or in a price war (the third contingency). This simplifies the analysis substantially since it means that we can classify all subgames as belonging to one of two classes of subgames.[1] We therefore only need to prove that the firms do not have an incentive to deviate in any of these two "phases." They do not have an incentive to deviate in a price-war subgame and they do not have an incentive to deviate in a cooperative subgame.[2]

Incentives to fight a price war Assume that at least one of the firms has charged a price different from the monopoly price at some point in the past. Then the firms are in a price war. The question is if a firm really has an incentive to fight the war and to forever charge price equal to cost, assuming that the other firm will forever

[1] The details of the history don't matter. If a deviation has occurred, both firms are instructed to fight a price war. It doesn't matter who deviated or what price the deviator charged or when the deviation occurred. Similarly if a deviation has not yet occurred, both firms are instructed to charge the monopoly price. It doesn't matter what time period it is.

[2] In addition, it is actually sufficient to prove that no player has an incentive to deviate in a single period. We therefore don't need to consider complicated deviations where a firm e.g. charges prices that are increasing or decreasing over time. This is called the one-stage-deviation principle. The principle is applicable to all repeated games with observed actions. To prove this result is beyond this lecture, however.

charge price equal to cost. Expressed differently, what incentives does a firm have to deviate from the trigger strategy during a price war (given that the other firm doesn't deviate)?

Let's say that firm i contemplates to deviate during a single period s. Firm i assumes that the competitor will follow the trigger strategy both in period s and in all future periods. Since it is a deviation during a single period, also firm i will follow the trigger strategy from period $s+1$ and onwards. That is, from period $s+1$ and onwards both firms will charge price equal to cost and earn zero profits, independent of what firm i does in period s.

What is then the best price for firm i to set in period s? Since the competitor will set price equal to cost, firm i cannot increase its profit by deviating from price equal to cost in period s.

That is, it is not profitable for the firm to deviate from the price war prescribed by the trigger strategy. Expressed differently, the price war is credible.

Incentives to charge the monopoly price Now consider a firm's incentive to deviate during the cooperative phase. Let's say that the deviation would occur during a single period s. The firm then understand that from period $s+1$ and onwards, both firms will follow the trigger strategy and set prices equal to cost. Profit will be zero from period $s+1$. The firm also understands that the competitor will set price equal to the monopoly price in period s. The best deviation the firm could make would therefore be to undercut its rival slightly, *i.e.* to set $p_i^s = p^m - \varepsilon$. Then, the deviating firm would essentially set the monopoly price and serve all customers. It would essentially earn the monopoly profit, π^m , during a single period. Then, the firm's payoff in the rest of the game would be

$$\Pi^{Deviation} = \pi^m.$$

With no deviation the firm would earn half the monopoly profit in period s and all periods onwards. (Since past payoffs are of no consequence, we may simplify the analysis by renumbering the periods, by setting $s = 1$.) Hence the firm's payoff in the rest of the game would be

$$\Pi^{NoDeviation} = \sum_{t=1}^{\infty} \delta^{t-1} \cdot \frac{\pi^m}{2}$$

which can be rewritten as:

$$\Pi^{NoDeviation} = \frac{\pi^m}{2} \cdot \sum_{t=1}^{\infty} \delta^{t-1} = \frac{1}{1-\delta} \cdot \frac{\pi^m}{2}.$$

In order for the firm not to deviate, it is required that

$$\Pi^{Deviation} \leq \Pi^{NoDeviation}$$

i.e.

$$\frac{1}{1-\delta} \cdot \frac{\pi^m}{2} \geq \pi^m,$$

or

$$\delta \geq \frac{1}{2}.$$

That is, if the firms are sufficiently patient and value the future, they will not deviate from the trigger strategy.

Other equilibria The trigger strategy equilibrium is not the only equilibrium when the firms are patient, however. Both firms following the competitive strategy, which requires a firm to set price equal to marginal cost independent of history (i.e. no matter what), is also a SPE.

Exercise 61. Prove this claim! (Hint: the proof is very similar to the proof that price wars are credible.)

20.1.3. Folk Theorems

The so-called Folk theorems demonstrate that the results above are "typical" for a large class of games. The Folk theorem characterizes the subgame perfect equilibria in super-games regardless of the stage game. In particular, any possible outcome, such that each player receives a payoff at least as large as the payoff he would get in the Nash equilibrium of the stage game, can be sustained as a SPE in a super-game if the discount factor is sufficiently close to one.

20.1.4. Interpretation of Infinite Time Horizon

Interpreted literally, an infinite time-horizon is not a very realistic assumption; every market is bound to vanish at some point in the future. An alternative assumption is that the lifetime is unknown. Instead, firms hold expectations about the probability that the market survives from one period to the next. For example, the firms may believe that the survival probability is 95 percent. In this case, there is a risk (5 percent probability) that the market will only persist one period. However, most likely (95 percent probability), the market will survive an additional period. Having survived two periods, it is likely that it will survive three periods. Having survived three periods, it is likely that it will survive four periods, and so on. Thus, assuming a fixed survival probability (above zero) it appears as if the market might continue forever. There is no last period in which cooperation necessarily breaks down. Even though there does not exist any last period, the expected lifetime is finite. If the survival rate is 95 percent, the expected lifetime is 20 periods. If the survival rate is 99 percent, the expected lifetime is 100 periods. If the market is expected to be long-lived (which is equivalent to a high survival probability) collusion is self-enforcing. If the market is expected to be long-lived, a duopoly firm that does not cheat can expect to earn half the monopoly profit in many periods. That

alternative may be preferred to earning the whole monopoly profit during just one period.

20.1.5. Exercise

Consider an infinitely lived symmetric oligopoly market. The interaction can be described by some stage game but we will not go into the details of this game. We just assume that if all the firms would play the Nash equilibrium of this stage game whey would all the profit π^N. If the firms would form a cartel agreement, they would all earn the profit π^C in every period they all abide by the agreement. If a firm would deviate from the agreement during a single period, while all the other firms stick the agreement, the deviating firm would earn the profit π^D during that period. The other firms would earn the profit π^S during the same period.

Exercise 62. Prove that a cartel is self-enforcing if $\delta \geq \frac{\pi^D - \pi^C}{\pi^D - \pi^N}$.

20.2. Factors Facilitating Cartels

The theory of super-games teaches us that oligopoly firms may sometimes be able to enforce cartel agreements, simply by threatening those who cheat with a price war. The same theory also teaches that whenever a cartel is self-enforcing it will also be an equilibrium for the firms to compete with each other. Unfortunately economists cannot say which equilibrium is the more plausible. Should we then predict that firms form cartels or that they compete? We have no answer to this question today.

The economic theory of cartels can still be very useful. What the theory shows is that cartels are possible, but only under certain conditions. We have already seen that the possibility of collusion depends on the discount factor (or *interest rate*). Collusion is possible *only if* the firms are sufficiently patient (only if the interest

rate is sufficiently low). But there are many other factors that are important for the firms' ability to fix prices. We will go through the most important factors below. This list of facilitating factors may guide policy in various ways. Competition authorities may for instance use the list to help identify potentially problematic markets, which can then be monitored more closely than other markets. The list can also be used in individual cartel cases to assess how plausible allegations about a cartel are.

20.2.1. Concentration

The single most important factor for a cartel to work is probably high market concentration. Collusion is impossible in fragmented markets, since the incentive to cheat is too strong and the harshest possible punishment is too weak.

Consider for simplicity a Bertrand oligopoly with homogenous goods. Assume now that there are n firms instead of only two as we assumed above. Then, if firms cooperate, they will earn a one n'th of the monopoly profit each (that is π^m/n). A cheating firm undercuts the monopoly price by one cent and will earn (almost) the whole monopoly profit during one period (that is π^m). The punishment is a price war and zero profits forever after. The short-term gain from cheating is thus

$$G = \pi^m - \pi^m \cdot \frac{1}{n} = \frac{n-1}{n} \cdot \pi^m > 0.$$

Clearly the temptation to cheat (*i.e.* the short-term gain from cheating) is larger the more firms there are in the market. The punishment, i.e. the per-period loss during the subsequent price war, is

$$L = 0 - \pi^m \cdot \frac{1}{n} < 0.$$

Clearly the punishment is weaker in a fragmented market.

The value of cheating can be written:

$$V = G + \frac{\delta}{1-\delta} \cdot L,$$

since G is the gain today and L is the per-period loss which occurs in every period starting tomorrow. Thus the value of cheating is given by

$$V = \frac{n-1}{n} \cdot \pi^m - \frac{\delta}{1-\delta} \cdot \pi^m \cdot \frac{1}{n} = \left[(n-1) - \frac{\delta}{1-\delta}\right] \cdot \pi^m \cdot \frac{1}{n}.$$

It is easy to see that cheating is unprofitable if, and only if,

$$n \leq \frac{1}{1-\delta}.$$

That is, a cartel agreement can only be enforced if there are sufficiently few firms in the market. In more fragmented markets, the firms are forced to compete.

20.2.2. Entry Barriers

Closely related to concentration are barriers to entry. And it is easy to see that even if a market is concentrated, the possibility of new entry can render cartels unstable. Consider first a market where it is know that new firms will soon enter, independent of whether there is a cartel or not. Then, even if the gains from cheating today may be small (since the market is concentrated) future entry reduces the strength of the punishment for cheaters. For example, if there are two firms in the market today and entry of a third firm will occur in the next period, the gains from cheating today will be half the monopoly profit, but the punishment will only be the loss of a third of the monopoly profit. In addition, cartels means higher prices and profits, which may in itself attract new entry, reducing the incumbents incentives to form a cartel.

20.2.3. Frequency of Interaction

Another crucial condition for cartels is that the firms interact frequently. This point is of special interest for public procurement. National and local governments procure all sorts of goods at regular intervals. Examples include medicines and supplies for schools, e.g. milk. The milk producers are then invited to submit their bids, *i.e.* the prices they wish to charge in case they win the contract. Like in any other market the milk producers have an incentive to form a cartel, or a *bidding ring* as it is also called in auction markets. If the contracts are for short durations and the auctions are held often, it is easier for the firms to sustain a cartel. The reason is that the cheater can be punished more quickly.

To see this point more formally, think of time as continuous but divided into short periods of length Δ. At the beginning of each period, there is a procurement auction and the winner gets to supply the market during the period. Then the a firm's profit may be written $\Pi_i = \sum_{t=1}^{T} \delta^{t-1} \cdot \pi_i \left(p_1^t, p_2^t\right)$, exactly as above and the discount factor is given by

$$\delta = e^{-r \cdot \Delta}$$

where r is the (continuous-time) interest rate. Note in particular that if the time periods are short (Δ small) then the discount factor is high. In the extreme, if $\Delta \approx 0$, then $\delta \approx 1$. The important conclusion is that short periods, or frequent interaction, increases the discount factor and thus facilitates collusion.

Anyone organizing procurements on a regular basis should therefore prefer long contract duration and infrequent auctions to discourage bidding rings.

20.2.4. Growing Markets

Collusion is easier in growing markets. If the punishment is metered out in a market that is larger than it was when the cheating occurred, the punishment will be more costly in relative terms. To see this more formally, consider a duopoly and let's say that the monopoly profit today is only π^m per period and that it will grow to be $\Pi^m > g \cdot \pi^m$ per period in the near future, say starting the next period. Then the gain from cheating today is

$$G = \frac{n-1}{n} \cdot \pi^m > 0$$

and the per-period loss during the subsequent price war, is

$$L = -\Pi^m \cdot \frac{1}{n} = -g \cdot \pi^m \cdot \frac{1}{n} < 0.$$

Collusion will then only be possible today if

$$\frac{n-1}{n} \cdot \pi^m > \frac{\delta}{1-\delta} \cdot g \cdot \pi^m \cdot \frac{1}{n}$$

i.e.

$$(n-1) \cdot \frac{1-\delta}{\delta} > g.$$

That is, the market must not grow too fast.

But collusion may also be more difficult to sustain in growth markets. One reason is that growing markets also require new investments to meet the future higher demand. New investments must typically be financed, at least partly, from the firms' own current profits. When firms need the money today rather than tomorrow, they should be more tempted to cheat on cartel agreements. Another reason is that growing markets may attract new entry.

20.2.5. Business Cycles

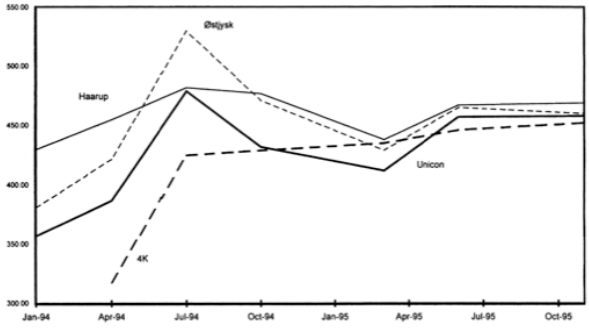
Collusion is more difficult in markets that are more exposed to business cycles. The reason is that collusion is difficult during downturns, when the market is shrinking. Then collusion will break down. The logic is the reverse of the argument concerning growing markets (see the previous section). But then collusion will not be possible during upswings either since the firms know that they will stop collude anyway as soon as the downturn comes, independent of whether they have behaved nicely to each other during the upswing. That is, during the upswing, the logic is the same as for markets with finite time.

20.2.6. Transparency

The Danish competition authority was worried about the competitive conditions in the Danish cement market and decided to do something about it. Since most firms treat contracts as business secrets, it may be difficult for customers to compare prices. But, in order for competition to work, the authority reasoned, the customers must be well informed. If they do not know what prices different firms charge from different customers they will not be able to buy from the cheapest one, and there will be little reason for the firms to lower their prices. The authority therefore decided in October 1993 that all transaction prices on cement should be published to enable price comparisons. The result was not what the competition authority had expected. Within a year prices had increased by 15-20 percent, as described by the figure below. During the same period most other prices where relatively constant, so the price increased could not be blamed on general inflation. What had happened?

Why Transparency may lead to Increased Prices A possible explanation is that the increased transparency helped the cement

Figure 20.1.: Danish cement



producers to organize a cartel. In order to sustain collusion, firms must be able to detect cheating. But in the cement market, as in many other markets, firms cannot observe the prices that their competitors charge in each transaction. And even if the firms can observe prices, the rivals can also cheat by giving extra service or increase the quality of the product and so on. Undetected cheating cannot be punished, and then cartels cannot be formed. The Danish competition authority later realized its mistake and stopped publishing prices in 1996.

Information Exchange It is common that firms exchange information with each other, for example through trade associations. It should be clear that such exchange may facilitate collusion. Quite naturally, competition authorities are therefore suspicious about information exchange. And certain forms are actually illegal. Aggregate historical information is okay. But firms are not allowed to exchange information about recent individual prices.

20.2.7. Other Aspects

The theory of super-games also leaves out several important factors. In the example above, the firms punish by charging a low price in the future. Such a punishment can be very costly. If one

firm deviates, the other firms cannot continue their cooperation. However, there may exist other ways to punish a deviator. For example by excluding the deviating firm from R&D joint ventures, standardization committees and so on. In order to cooperate, firms must be able to agree on which price to collude on. If firms are very different, for example if they have different costs, they will have different preferences on which price to chose. Intuitively, such differences may make agreements more difficult.

20.3. Transparency, revisited

The example from the Danish cement industry suggests that it is more difficult for firms to collude if they cannot observe each other's prices. But, a lack of transparency doesn't necessarily make collusion completely impossible. Some limited cooperation may still be possible, but this cooperation will require a somewhat more refined strategy than the trigger strategy.

Even if the firms cannot observe each other's prices, they can monitor their own sales. Low sales may then be taken as a signal that the competitor has cheated and charged a lower price. But sales may be low also for other reasons, e.g. that market demand is low. When a firm cannot tell if its sales are low as a result of a competitor cheating or as a result of low market demand, the firm does not have a choice - it must start a price war. Because if it doesn't the competitor will for sure cheat.

When market demand fluctuates, the firms will observe low sales from time to time and they will actually have to fight price wars from time to time. These price wars will be costly. To maximize profits the firms will agree to fight as short price wars as possible. They should only be long enough to deter cheating.

This logic is somewhat paradoxical. A successful cartel agreement will, by definition, deter cheating. The firms will then know

that whenever they observe low sales, it is because of low market demand, not because of cheating. Still, they have to initiate a "punishment" every time their sales are low. Because otherwise the competitors will cheat. One may think of this as a "reversed burden of proof:" Firms will punish each other, unless there is clear evidence that all rivals honored the agreement.

A Formal Model Consider a market with two firms producing perfect substitutes with the same constant marginal cost, c. If the firms compete, there is Bertrand competition resulting in zero profits. Assume that market demand is uncertain. In every period demand is low, say zero for simplicity, with probability α, and high with the remaining probability $1-\alpha$. That is:

$$d(p) = \begin{cases} 0, & \textit{with probability } \alpha \\ \\ D(p), & \textit{with probability } 1-\alpha \end{cases}$$

Let p^m be the monopoly price when demand is high. Similarly, π^m is the monopoly profit when demand is high. The firms cannot observe market demand, neither before nor after they set their prices. The firms are also unable to observe the prices their rival has charged. As a result, the firms face a so-called *signal extraction* problem. They cannot tell if their sales are low (zero) because market demand is low, or because the rival has charged a price below their own price.

Price War Strategy Clearly the firms are unable to use the trigger strategy in this market, since they cannot condition their behavior on the rivals past prices. To sustain collusion they may instead consider a somewhat more complicated strategy, called the price-war strategy (PWS). Exactly as the trigger strategy, PWS has two phases, namely the collusive phase and the price war phase, and it

prescribes the same actions:

- In collusive phase, set the monopoly price, p^m.
- In the price war phase, set price equal to marginal cost, c.

What differs is the transitions between the two phases. PWS prescribes the following transitions:

- Start out in collusive phase.
- If the profit turns out to be $\pi^m/2$, stay in the collusive phase.
- Otherwise, switch to the price war phase for T periods and then switch back to collusive phase.

The question is now under what conditions the firms can sustain collusion using this price war strategy. In particular, what is the shortest duration T of the price war, necessary to deter cheating?

It is easy to see that setting price equal to marginal cost during the price war is a credible threat, since this behavior is the Nash equilibrium of the stage game. The logic of the proof is the same as in the case of the trigger strategy.

Value functions - some general remarks To study the incentives to cheat we will have to define the value of a firm under different circumstances. To do this, it is convenient to first recall two different methods of computing the present value of a firm. The direct way of writing down the value function of a firm, earning π in every period, is

$$V = \pi + \delta \cdot \pi + \delta^2 \cdot \pi + ...$$

This expression can be simplified to give $V = \frac{\pi}{1-\delta}$.[3]

[3]To see this, first multiply the equation by δ to get

$$\delta \cdot V = \delta \cdot \pi + \delta^2 \cdot \pi + \delta^3 \cdot \pi + ...$$

Alternatively, we can define the value function recursively as

$$V = \pi + \delta \cdot V$$

where the first term on the right hand side is the profit earned during the current period and the second term captures the fact that the firm will have the same value V tomorrow and that the current value of owning the firm tomorrow must be discounted by δ. Again, it is easy to see that also this expression can be simplified to give $V = \frac{\pi}{1-\delta}$.

Firm values, in the collusive phase and during price war Now back to defining the value of the colluding firms. We will first define the value of a firm when both firms follow the PWS. Clearly the value of a firm is different depending on whether the firms are in a collusive phase or in a price war. We will therefore define two value functions, one for the beginning of a period when the firms find themselves in the collusive phase, V^+, and one for the beginning of the first period of price war, V^-. In the beginning of a period when the firms are in a collusive phase the value of a firm is equal to

$$V^+ = (1-\alpha) \cdot \left[\frac{\pi^m}{2} + \delta \cdot V^+ \right] + \alpha \cdot \left[\delta \cdot V^- \right]$$

The first term on the right hand side reflects the possibility that demand will be high during the current period. This event occurs with probability $1 - \alpha$. In this case both firms will charge the monopoly price, and each firm earns half of the monopoly profit during the current period. As a result the firms will find themselves

---

Second, take the difference between the two equations:

$$V - \delta \cdot V = \pi - \delta \cdot \pi + \delta \cdot \pi - \delta^2 \cdot \pi + \delta^2 \cdot \pi - \delta^3 \cdot \pi + ...$$

i.e

$$(1-\delta) \cdot V = \pi.$$

in the collusive phase also the next period. Therefore their value will be the same in the beginning of the next period. To compute the current value of being in the collusive phase tomorrow we simply discount V^+ by δ. The second term on the right hand side reflects the possibility that demand will be low during the current period. This event occurs with probability α. In this case the firms will make a zero profit in the current period despite both firms charging the monopoly price. As a result the firms will find themselves in the beginning of a price war in the next period. To compute the current value of being in the beginning of the price war phase tomorrow we simply discount V by δ.

In the beginning of the price war phase the value of a firm is given by

$$V^- = \delta^T \cdot V^+.$$

In the beginning of the price war phase the firm has to endure T periods without profit. After the price war is over, the firms will switch back to the collusive phase. Then their value will again be given by V^+. But since this happens T periods later, the current value must be discounted.

The two value functions are interdependent. Thus we need to consider them as a system of two equations in two unknowns. In particular we solve for the value of being in the collusive phase by using the second equation to substitute for V^- in the first equation, *i.e.*

$$V^+ = (1-\alpha) \cdot \left[\frac{\pi^m}{2} + \delta \cdot V^+ \right] + \alpha \cdot \left[\delta \cdot \delta^T \cdot V^+ \right].$$

Next we solve for the value, *i.e.*

$$V^+ = \frac{(1-\alpha)}{(1-\delta) + \alpha \cdot \delta \cdot [1-\delta^T]} \cdot \frac{\pi^m}{2}.$$

Similarly one may solve for V^-. But this is of less interest.

Value of cheating Note that the value of a firm (in collusive phase) is higher, the shorter is the price war they will fight following a period of low demand. This means that the firms will wish to agree on the shortest possible price war. The question is now what incentive a firm has to deviate and undercut the monopoly price in the collusive phase. The optimal one-period deviation is to charge slightly under the monopoly price, thereby capturing essentially the whole monopoly profit during the current period. That, however, will trigger the start of a price war in the next period, implying that the value of a cheating firm is

$$V^D = (1-\alpha)\cdot\left[\pi^m + \delta\cdot V^-\right] + \alpha\cdot\left[\delta\cdot V^-\right].$$

Length of Price Wars In order for the firms not to cheat it is necessary that $V^+ \geq V^D$. This condition can be rewritten as[4]

$$2\cdot\frac{\delta}{1-\delta}\cdot\left(1-\delta^T\right)\cdot\left(\frac{1}{2}-\alpha\right)\geq 1.$$

Notice that in order for the left hand side to be positive it is necessary that $\alpha < \frac{1}{2}$. That is, the probability of high demand must be relatively high in order to sustain collusion. We assume this condition to be fulfilled.

The left hand side depends on the length of the price war. In particular, the higher T is the larger is the left hand side. If $T = 0$, the left hand side is zero and the condition is violated. This means that it is necessary to fight a price war of some duration to sustain collusion. For example, if $\delta = \frac{3}{4}$ and that $\alpha = \frac{1}{4}$, the no-cheating

[4]Notice that $V^+ \geq V^D$ can be rewritten as $\frac{\pi^m}{2} + \delta\cdot V^+ \geq \pi^m + \delta\cdot V^-$ since $V^+ = (1-\alpha)\cdot\left[\frac{\pi^m}{2} + \delta\cdot V^+\right] + \alpha\cdot[\delta\cdot V^-]$ and $V^D = (1-\alpha)\cdot[\pi^m + \delta\cdot V^-] + \alpha\cdot[\delta\cdot V^-]$. The no-cheating condition can be rewritten as $\delta\cdot(V^+ - V^-) \geq \frac{\pi^m}{2}$ or $\frac{\delta\cdot\left(1-\delta^T\right)\cdot(1-\alpha)}{(1-\delta)+\alpha\cdot\delta\cdot[1-\delta^T]} \geq 1$ which can be simplified to the expression in the main text.

condition simplifies to

$$\left(\frac{3}{4}\right)^T \leq \frac{1}{3}.$$

Thus, in this case, a punishment of four periods will suffice. And since punishments are costly, the firms would agree on four periods.[5]

Summary It is more difficult to form a cartel in a market that is not transparent, meaning that the firms cannot observe each other's prices. Then, they cannot observe cheating. It is also difficult to infer cheating indirectly by observing one's own sales since market demand typically is uncertain. It may still be possible for the firm to collude. This will however require firms to actually fight price wars when they observe low sales. The firms will have to punish each other unless there is proof of no collusion. For this reason it is important for the firms to reduce the length of the price war as much as possible, while still providing sufficient incentives to deter cheating. Less patient firms have more incentives to cheat and must therefore be punished more severely, by a longer price war.

20.4. Fighting Cartels

Since cartels transfer wealth from consumers to firms and since they reduce market efficiency, cartels have been made illegal. This means first of all that cartel members cannot sue rivals who break the agreement. We have seen, however, that under some conditions cartel firms can enforce their agreements themselves, for example by punishing cheaters with price wars. Just denying cartels help to enforce their agreements is thus not sufficient. For this reason competition authorities have been set up with the capability to

[5]More generally, it is straight forward to show that if

$$\delta \cdot (1 - \alpha) \geq \frac{1}{2}$$

there exists some T^* which is large enough for a cartel to be stable.

investigate suspected cartels and to impose fines in case there is sufficient proof of the cartel. In addition the cartel firms may have to pay damages to their customers.

One may ask, however, what the point of all this is? If market conditions made a cartel possible before detection, one may fear that the cartel will simply continue its operations also afterwards. After the legal process is over and the fine has been paid, then market may look exactly the same as before. The competition authorities do not have the mandate to change the industry structure, *e.g.* by breaking up firms to reduce concentration, to reduce the risk of future collusion. Thus, if it was profitable for the firms to form a cartel before, it may well be profitable to resume the cooperation as soon as possible again.

Dismantling of the cartel It seems reasonable, however, to believe that detected cartels will sometimes dismantle. Following the investigation and court procedures much information about the industry and the cartel will be known to the competition authority and it may therefore be easier for the authority to monitor the industry. If this is indeed the case, fighting cartels will improve the market outcome.

Deterrence Ideally, however, a judicial system should deter crimes, not just punish them. That is, the risk of being caught and punished should ideally eliminate the firms' incentives to form a cartel in the first place. The question is if this indeed is the case? Standard economic arguments suggest that crimes are deterred if the expected punishment (the probability of being caught times the fine) is at least equal to the benefits of committing the crime (the increased profit as a result of the cartel). It is easy to demonstrate that this condition indeed applies also to cartel deterrence.

To see this, let π^C be the per-period profit of a firm when there is a cartel, let π^N be the per-period profit of a firm when there is no

cartel (the profit in the Nash equilibrium of the stage game), and let π^D be the profit that a firm can make during one period if it cheats on the cartel. Let p be the probability of being caught and F the fine (and damages) that a firm has to pay if caught.[6]

The value of a cartel firm is then $V^C = \pi^C - p \cdot F + \delta \cdot V^C$, assuming that the cartel would continue after being caught. That is, $V^C = \frac{1}{1-\delta} \cdot \left[\pi^C - p \cdot F\right]$. The value if cheating is $V^D = \pi^D + \frac{\delta}{1-\delta} \cdot \pi^N$. A cartel is stable if, and only if, $V^C \geq V^D$. This condition can be simplified to

$$\delta \geq \frac{\pi^D - \pi^C}{\pi^D - \pi^N} + \frac{p \cdot F}{\pi^D - \pi^N}$$

Without competition authorities, $p \cdot F = 0$, and the condition is the same as before. With competition authorities, $p \cdot F > 0$, cartels have become more difficult in the sense that the minimum discount factor has increased. Cartels may then be deterred at least in some markets. If the expected cost of the punishment is set larger than the benefits of the crime, i.e.

$$p \cdot F > \pi^C - \pi^N \tag{20.1}$$

the condition for a cartel agreement to be self-enforcing simplifies to $\delta > 1$, which is never satisfied. That is, there is complete deterrence.[7]

[6] It is assumed that F is the same, independent of how long the cartel has been active. The reason may e.g. be that it is difficult for competition authorities to prove that the cartel was active before the current period.

[7] In case cartels dismantle following successful conviction, the condition for deterrence is actually less stringent. The value of a cartel firm is then $V^C = \pi^C + (1-p) \cdot \delta \cdot V^C + p \cdot \left[\frac{\delta}{1-\delta} \cdot \pi^N - F\right]$, which simplifies to $V^C = \frac{1}{1-(1-p)\cdot\delta} \cdot \left[\pi^C + p \cdot \left[\frac{\delta}{1-\delta} \cdot \pi^N - F\right]\right]$. The value if cheating is $V^D = \pi^D + \frac{\delta}{1-\delta} \cdot \pi^N$. Cartels are deterred if, and only if, $V^C < V^D$. This condition can be simplified to

$$p \cdot F > \left(\pi^C - \pi^N\right) - \left[1 - (1-p) \cdot \delta\right] \cdot \left(\pi^D - \pi^N\right).$$

The first term on the right hand side is the same as when the cartel continues after discovery. Since the second term on the right hand side is negative,

Empirical assessments Based on this criterion, equation 20.1, some economists have argued that cartel deterrence today might not be sufficient. Based on evidence from the U.S. it has been estimated that cartels may face a 10-15% detection probability per year. This probability is estimated from the age of detected cartels. (For example, if the detection probability is 5%, then the average age of detected cartels is 20 years.) With a 15% detection rate (per year) the fine must be almost 7 times larger than the yearly gains from cartelization $\frac{F}{\pi^C - \pi^N} > \frac{1}{p} = \frac{1}{0.15} \approx 6.7$. However, it has also been estimated that fines and damages amount to approximately 115% of the costs resulting from the cartel. If we assume that the cartel firms' gains equal the inflicted cost, we have $\frac{F}{\pi^C - \pi^N} = 1.15$, and it is clear that the numbers do not match.

the punishment can be less sever when discovery leads to dismantling of the cartel.

21. Horizontal mergers

Mergers and acquisitions turn around trillions of U.S. dollars every year and are important determinants of how the economy works. More precisely, mergers and acquisitions are instruments for restructuring the economy, to change the way goods and services are produced and marketed, to adapt the economy to new technologies and liberalized international trade. Mergers and acquisitions shape the competitive conditions in our markets. But even a quick look in the business press reveals that mergers and acquisitions are often both puzzling and controversial.

21.1. Puzzles and Controversies

Do mergers benefit the owners? Most puzzling, perhaps, is all the empirical evidence that mergers often fail. Some studies find that as many as two thirds of all mergers result in lower profits for the merging firms. But despite this dismal evidence, merger activity remains at a huge level, by any measure. For instance, the worldwide value of mergers and acquisitions exceeded 3.4 trillion US dollars in 1999. And nearly half of all major US corporations received a takeover offer during the 1980s. What could possibly be the motive behind all these risky deals? Equally puzzling is that, while mergers often appear to lower profits, share prices typically increase. Share prices and profits thus seem to contradict each other, since a firm's stock market value should reflect the firm's profit opportunities. So, what should we believe? Do mergers and acquisitions benefit the firms' owners or not?

Do they benefit the rest of us? It is also not clear whether mergers benefit the rest of us. Mergers may perhaps help firms reduce costs by producing goods and services more efficiently, but they may also help firms monopolize their markets. For this reason, any two large companies that wish to merge have to seek approval from competition authorities before they can proceed. And sometimes authorities say no. These decisions often turn out highly controversial.

GE and Honeywell Take the merger between the two giant American corporations, GE and Honeywell, in 2000. The U.S. authorities (the Department of Justice) thought that the merger would benefit competition and approved the deal. But then the European Commission said no. The Commission was worried that the merger would reduce competition. One of the fears was that a merged GE-Honeywell would be able to offer customers bundles of GE:s airplane engines and Honeywell's aviation electronics and that competitors would not be able to match such offers. The European decision to block the merger between the two American companies triggered a debate between American and European political leaders. President Bush expressed his "concern" in public and Commissioner Monti responded that he "deplore[d] attempts to... trigger political intervention."

Volvo and Scania The same year, the European Commission also prohibited the Swedish truck manufacturer Volvo's acquisition of the competing Swedish truck manufacturer Scania. The Commission argued that a merger between Volvo and Scania would reduce competition in both the Swedish and Finish truck markets. This and similar interventions triggered a heated political debate about EU merger control and market definitions in several Member States. Representatives of smaller countries declared that, in effect, EU merger policy makes it impossible for companies in small

countries to merge and obtain a leading global position. Following the Volvo/Scania decision, there was almost complete consensus among the political parties in Sweden on this view, with only the Liberal party expressing support for the Commission's decision. These claims have been rebutted by EU officials, who argue that companies in smaller countries can expand by merging with companies operating in other countries. According to this line of reasoning, the Volvo/Renault operation and the strategic partnership concluded by Scania/Volkswagen, following the prohibition of the Volvo/Scania merger, clearly showed that there were alternative ways for these companies to merge.

Issues In this chapter we will start analyze the most basic economic issues concerning mergers. How do mergers affect competition and prices? What are the effects on economies of scale and the cost of production? Do mergers increase profits? Do they benefit or harm consumers? Are mergers good for welfare? Should we control mergers and, if so, how should the control system be designed?

21.2. Effects of Mergers

21.2.1. Effects of Mergers on Competition

A merger between two firm competing with each other in the same market, a so-called horizontal merger[1], will reduce competition in that market. The most direct reason for the loss of competition is of course the fact that the two firms participating in the merger will stop competing with each other. More precisely the merging firms will have an incentive to increase their prices and to reduce their output. This phenomenon goes under many different names.

[1] A merger between competing firms is often called "horizontal" while a merger between two firms that having a buyer-seller relation is called "vertical." Mergers between firms without any obvious relationship are called "conglomerate".

Competition authorities often refer to it as the "unilateral effect" of the merger. I will refer to it as "internalization." Others call it the "portfolio effect".

In addition to the loss of competition between the merging firms, a merger can also make it easier for all the firms remaining in the market to form a cartel. Cartels can namely only be formed if the market is sufficiently concentrated (as will be discussed in a later lecture). Competition authorities often refer to this as the "coordinated effects" of a merger. During this lecture, we will focus on internalization (unilateral effects) only.

Internalization To see why horizontal mergers increase prices, consider a merger between two competitors named A and B. Before the merger, when firm A contemplates increasing its price, the firm faces two adverse effects, a loss of sales to firm B and a loss of sales to other rivals in the same market or firms in completely different markets. When the merged firm contemplates increasing the price of product A, it would only face one these adverse effect, namely the loss of sales to other rivals and firms in other markets. That some customers turn to product B, when the price of A is increased, is no longer a loss, since the same firm owns both products. In short: following a horizontal merger, price increases reduce sales less and are therefore more attractive.

What determines the strength of internalization? We can gain some additional insights about the firms' incentives to increase prices by comparing the firms' first-order conditions before and after the merger. Before the merger firm A's profit is given by

$$\pi_A(p_A, p_B) = (p_A - c_A) \cdot D_A(p_A, p_B)$$

where p_A and p_B are the prices of the two firms, c_A is A's marginal cost of production (which is assumed to be constant) and $D_A(p_A, p_B)$

is the demand for A's product, which depends on the prices of both the competing products. In equilibrium, firm A would set its price to satisfy the first-order condition:

$$\frac{\partial \pi_A (p_A, p_B)}{\partial p_A} = D_A (p_A, p_B) + (p_A - c_A) \cdot \frac{\partial D_A (p_A, p_B)}{\partial p_A} = 0$$

where $\frac{\partial D_A(p_A,p_B)}{\partial p_A} < 0$ is the loss of sales when price is increased. Firm B's pre-merger price satisfies a similar condition.

After the merger, the integrated firm's profit is given by

$$\pi (p_A, p_B) = (p_A - c_A) \cdot D_A (p_A, p_B) + (p_B - c_B) \cdot D_B (p_B, p_A)$$

assuming that the firm continues to produce and market both product A and B. The merged firm would set the price of product A to satisfy the first-order condition:

$$\frac{\partial \pi (p_A, p_B)}{\partial p_A} = D_A (p_A, p_B) + (p_A - c_A) \cdot \frac{\partial D_A (p_A, p_B)}{\partial p_A} + (p_B - c_B) \cdot \frac{\partial D_B (p_B, p_A)}{\partial p_A} = 0.$$

Note that even if an increase in the price of product A leads to a loss of sales of product A, given by $\frac{\partial D_A(p_A,p_B)}{\partial p_A} < 0$, the sales of B is increased, as indicated by $\frac{\partial D_B(p_B,p_A)}{\partial p_A} > 0$. That is, the merged firm regains some of the lost sales.

To assess the strength of the internalization effect one may evaluate the merged firm's first marginal profit $\frac{\partial \pi(p_A,p_B)}{\partial p_A}$ at the equilibrium prices before the merger. At these prices,

$$\frac{\partial \pi (p_A, p_B)}{\partial p_A} = (p_B - c_B) \cdot \frac{\partial D_B (p_B, p_A)}{\partial p_A} > 0$$

since the first two terms of the merged firm's marginal profit coincide with firm A's first-order condition and thus equal zero (at the pre-merger prices). That is, a small increase in p_A from the level prevailing in the market before the merger would increase the profit, $\pi (p_A, p_B)$.

The formula also reveals that the strength of this effect is determined by two factors, namely the markup before the merger $(p_B - c_B)$ and how close substitutes the two products are, as measured by the cross-price derivative, $\frac{\partial D_B(p_B,p_A)}{\partial p_A}$. If the markups are high already before the merger, for example because competition is weak already before the merger, then the merged entity will have stronger incentives to increase prices. Moreover, if the products of the two firms are close substitutes, *i.e.* if $\frac{\partial D_B(p_B,p_A)}{\partial p_A}$ is large, meaning that the two firms must have competed fiercely before the merger, then the merged entity will have stronger incentives to increase the prices.

In fact, when competition authorities evaluate the likely competitive effects of a merger they are more concerned with a merger in a market with little competition already before the merger and they are more concerned if the merging firms' products are close substitutes.

The Diversion Ratio One may also rewrite the above first-order conditions on "elasticity-form," to reveal the importance of the so-called diversion ratio, which is a more intuitive measure of how close competitors the two products A and B are. Recall that before the merger, firm A's pricing can be described by the inverse-elasticity rule:

$$\frac{p_A - c_A}{p_A} = -\frac{1}{\varepsilon_{AA}}$$

where ε_{AA} is the own-price elasticity of demand for product A.

Similarly, the merged firm's first-order condition can be rewritten as:

$$\frac{p_A - c_A}{p_A} + \frac{p_B - c_B}{p_B} \cdot \frac{\varepsilon_{BA}}{\varepsilon_{AA}} \cdot \frac{D_B}{D_A} = -\frac{1}{\varepsilon_{AA}}$$

where ε_{BA} is the percentage change in the demand for product B if the price of A is increased by one percent (i.e. the cross-price elasticity of demand).

Exercise 63. Rewrite the merged firm's first-order condition on this form.

The *diversion ratio* from A to B quantifies how much of the displaced demand for product A that switches to product B, when the price of A is increased by one percent. The diversion ratio is given by:

$$\delta_{AB} = -\frac{\varepsilon_{BA}}{\varepsilon_{AA}} \cdot \frac{D_B}{D_A}.$$

Expressed differently, the diversion ratio is the fraction of product A:s demand which treats product B as a second choice. If $\delta_{AB} = 0$ the two products do not compete with one another at all. If $\delta_{AB} = 1$, all the consumers who stop buying A switch to B instead.

To simplify the analysis, assume now that the two products A and B are symmetric, in the sense of having the same cost of production, i.e. $c_A = c_B = c$, and the same demand, *i.e.* $D_A(p_A, p_B) = D_B(p_B, p_A)$, which implies that $\varepsilon_{AA} = \varepsilon_{BB} = \varepsilon$ and $\varepsilon_{BA} = \varepsilon_{AB}$. Then, after the merger the prices of the two products $p_A = p_B = p$ are characterized by

$$\frac{p - c}{p} \cdot (1 - \delta_{AB}) = -\frac{1}{\varepsilon}.$$

Hence, the change in the markup (or Lerner index) caused by the merger is given

$$\frac{p - c}{p} = -\frac{1}{\varepsilon} \cdot \frac{\delta_{AB}}{1 - \delta_{AB}}.$$

This formula shows that a merger will lead to a larger increase in market power (as measured by the Lerner index) the less elastic is demand and the larger is the diversion ratio between the two firms' products. If the two products do not compete, *i.e.* $\delta_{AB} = 0$, the merger does not affect market power at all. On the other hand, if most of the losses of sales of product A would be regained as increased sales of product B, *i.e.* $\delta_{AB} \approx 1$, the merger would lead to a very drastic increase in market power. Intuitively, if the merged

firm wouldn't loose almost any sales by increasing its price, it would indeed increase prices a lot.

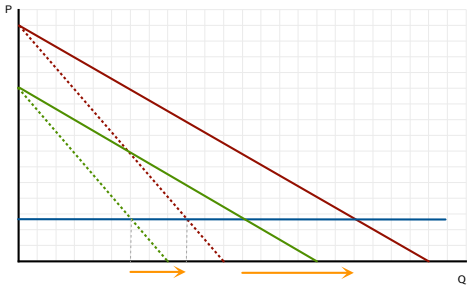
Outsiders' Response To fully assess the effects of the merger it is also important to understand that a merger will also trigger a response from other firms in the industry. In particular, when the merging firms increase their prices and reduce their output, the residual demand for the rivals increase. With increased residual demand the rivals, or the outsiders to the mergers, will have an incentive to increase their prices and expand their production. An important issue for determining the overall effect of the merger is then, to what extent the outsiders will simply increase their prices and to what extent they will increase production instead. The outsiders' response is determined by many different factors. Outsiders will increase sales more if:

- Outsiders don't have capacity constraints
- Outsiders have low costs
- Outsiders compete intensely with each other
- New firms can enter the market easily

If the outsiders are unable to increase sales, for example as a result of binding capacity constraints, they will instead increase their prices much. In such a case the merger gives not only the insiders, but also the outsiders, increased market power. The merger is more anticompetitive. On the other hand, if the outsiders do not have capacity constraints and if they compete hard with each other, they will increase sales rather than prices. In such a case the merger is less anticompetitive.

The importance of the intensity of competition among the outsiders can be illustrated using Figure 21.1. Here it is assumed that all the outsiders produce a homogenous good (which may be differentiated from the goods produced by the insiders). The merger is

Figure 21.1.: Outsiders' response



described by an increase in the residual demand for the outsiders, from the green demand curve to the red. If there is intense competition, the outsiders will not be able to increase their price. Instead the sales are expanded, as described long orange arrow on the right. If there is no competition among the outsiders, the same increase in residual demand will trigger a price increase and a much smaller increase in sales, as described by the short orange arrow on the left.

Summary A horizontal merger is a merger between former rivals. Such a merger reduces competition, *i.e.* it increases the merging firms' market power. Competition will be reduced more if there is less competition already before the merger and if the merging firms' products are close substitutes. The merger will also trigger a response from the firms not participating in the merger ("outsiders"). The more intense competition is between the outsiders, the less anticompetitive the merger will be.

In this section, we have analyzed internalization and the outsiders' response as if they are two separate issues. But, actually, they are not. For this reason we will soon return to the issue using a formal model, which will enable us to study both issues at the same time.

21.2.2. Effects of Mergers on Productive Efficiency

If horizontal mergers only would result in a loss of competition, there would be no reason to allow them. But most horizontal mergers probably also generate some efficiency gains for example in the form of lower production costs.

Efficiency Gains Horizontal mergers allow the new firm to *rationalize production* by reallocating production from less efficient plants to more efficient plants. In the extreme, very efficient plants may even be shut down completely.

Horizontal mergers may also allow firms to better exploit *economies of scale*. It may for example be possible to avoid the duplication of certain fixed costs when two head offices become one. The merged firm may also avoid the duplication of development costs by using the same components in different products. For example, different brands of cars owned by the same company often use the same components such as engines and transmission systems. It may also be possible to lower costs by increased specialization and longer production runs. That is, instead of producing a variety of different components in several plants, each plant may specialize in producing only a few components.

Horizontal mergers may also promote *technical progress* by pooling existing knowledge or by coordinating R&D efforts.

Finally mergers may *reduce slack* by replacing inefficient managements. In fact, the mere threat of hostile takeover may discipline managements that typically enjoy considerable discretion in large corporations.

Efficiency Losses Mergers may also, however, reduce productive efficiency for various reasons. One reason is that a larger organization may be more difficult to manage. But also the merger in

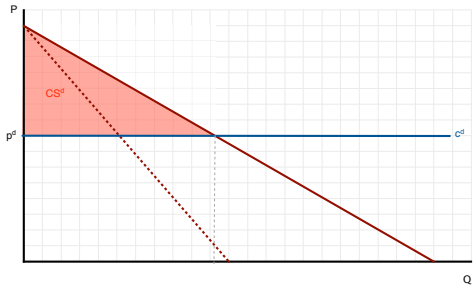
itself may create problems of melting together different company cultures. In addition, the loss of competition in the product market may allow firm more slack.

21.2.3. Effects of Mergers on Welfare - The Tradeoff

Horizontal mergers reduce competition but may also reduce production costs. The loss of competition benefits all the firms in the industry but harms consumers. More efficient production benefits the merging firms, not only by lowering costs. In fact a cost reduction can also make the merging firms more competitive. A sufficiently large reduction in the merging firms' marginal cost will induce the integrated firm to lower its price to poach customers from rivals. Cost savings are thus beneficial for the consumers but harmful to competitors. The question is now what the overall effect will be on the different parties?

Total Welfare Consider a merger with the most drastic change in the competitive conditions. Before the merger there are two firms competing a la Bertrand. The equilibrium duopoly price is equal to the marginal cost of the duopoly firms, $p^d = c^d$. Figure 21.2 illustrates the total welfare created in the market, which is equal to

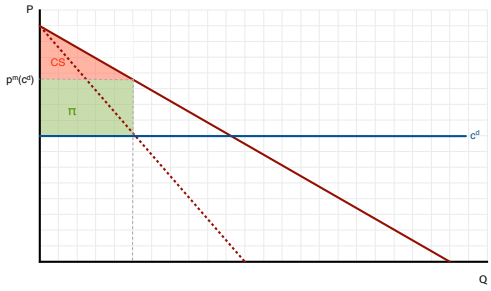
Figure 21.2.: Before merger



the consumer surplus, CS^d. There are no profits. That is, under duopoly, total welfare is given by $W^d = CS^d$.

The merger transforms the market into monopoly. If the cost remain unaffected by the merger, the monopolist will increase price to $p^m\left(c^d\right)$. As shown by Figure 21.3, the consumer surplus is smaller

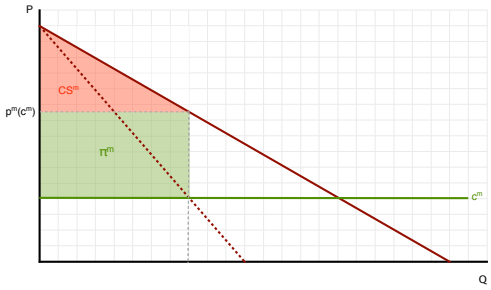
Figure 21.3.: After merger, absent cost savings



but profits are larger. The merger will clearly result in a deadweight loss, which is to say that the firms gain less than the consumers lose.

But, the merger may also allow the firm to produce more efficiently, perhaps to reduce marginal cost from c^d to c^m. As a consequence, the monopolist will also lower its price from $p^m\left(c^d\right)$ to $p^m\left(c^m\right)$. The end outcome is illustrated by Figure 21.4, Under

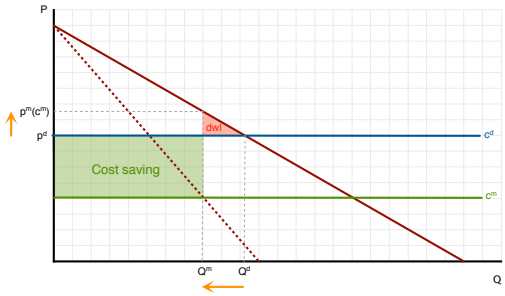
Figure 21.4.: After merger and cost savings



monopoly, total welfare is given by $W^m = CS^m + \pi^m$.

Comparing the monopoly outcome with the duopoly outcome, there are two effects on total welfare, as revealed by Figure 21.5.

Figure 21.5.: Comparison



1. There is a reduction in production costs, implying that society saves resources. The total value of this cost saving is equal to the green rectangle in the diagram.

2. There is a reduction in production which creates a dead weight loss. This loss is equal to the area of the pink triangle.

Notice that the reduction in marginal cost not only allows society to save resources. The reduction in marginal cost also implies that the dead weigh loss created by the merger is smaller than it otherwise would have been.

As the figure is drawn, the cost savings are actually larger than the deadweight loss. That is, despite the loss of competition, the merger is beneficial from a social welfare point of view. Of course, if we would have assumed a more modest reduction in cost, the merger would have been socially inefficient.

Exercise 64. What happens to the cost saving (the green rectangle) and the dead weight loss (the pink triangle) if the marginal cost is not reduced as much as in the figure above? Explain these two effects in words. Draw a figure of a socially inefficient merger.

Also note that if we know the demand curve and if we can predict how costs will be changed as a result of the merger, then we can actually predict the welfare effect of the merger. That is, we can reproduce the figure above. Expressed differently, if we know market demand, we can compute the minimum necessary cost saving for the merger to increase social welfare.

Appendix: Total welfare with formulas Recall that total welfare is defined as the sum of the *net* consumers' surplus and the firms' profits. Let P be the inverse demand function and Q the output, then total welfare is given by:

$$W = \left[\int_0^Q P(x) \cdot dx - P(Q) \cdot Q\right] + [P(Q) \cdot Q - c \cdot Q]$$

which may be rewritten as the difference between the *gross* consumers' surplus and production costs:

$$W = \int_0^Q P(x) \cdot dx - c \cdot Q.$$

That is, the consumers' expenditures and the firms' revenues cancel. The change in total welfare, caused by the merger, is therefore given by

$$\triangle W = \left[\int_0^{Q^m} P(x) \cdot dx - c^m \cdot Q^m\right] - \left[\int_0^{Q^d} P(x) \cdot dx - c^d \cdot Q^d\right]$$

which may be rewritten as

$$\triangle W = -\int_{Q^m}^{Q^d} P(x) \cdot dx + c^d \cdot \left(Q^d - Q^m\right) + \left(c^d - c^m\right) \cdot Q^m$$

or

$$\triangle W = -\int_{Q^m}^{Q^d} P(x) \cdot dx + \int_{Q^m}^{Q^d} c^d \cdot dx + \left(c^d - c^m\right) \cdot Q^m$$

or

$$\triangle W = -\int_{Q^m}^{Q^d} \left[P(x) - c^d\right] \cdot dx + \left(c^d - c^m\right) \cdot Q^m$$

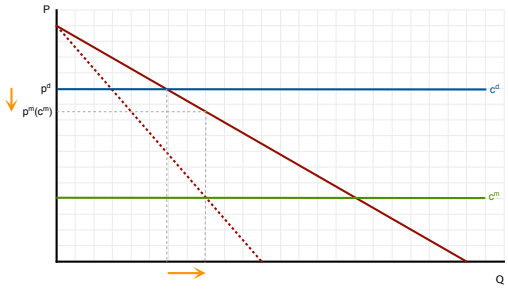
where the first term on the right hand side is the dead weight loss created by the merger (pink triangle) and the second term is the cost saving (green rectangle).

Consumer Welfare Even if the merger illustrated in the figure above is efficient from a total welfare point of view, it actually implies a redistribution of wealth from consumers to the firms. The reason is that the price is increased as a result of the loss of competition, despite the fact that marginal cost is lowered.

Most competition authorities around the world, including the European Commission and the Department of Justice and the Federal Trade Commission in the U.S. only allow merger that are believed to benefit the consumers. That is the merger above, even if it is socially efficient, would be blocked. The question is then, could a horizontal merger ever be cleared?

The answer is yes, as illustrated in Figure 21.6. If the cost saving

Figure 21.6.: Comparison



is large enough, a horizontal merger may actually reduce price and benefit consumers. To complicate things, competition authorities would call such a merger pro-competitive, despite the reduction in competition (as economists typically use this term).

Note that it is crucial that it is marginal costs that are lowered for consumers to gain. If the merger would only lead to a saving of fixed costs, the firm would not have an incentive to lower its price. Also note that the figure suggests that the reduction of marginal costs must be quite substantial for a merger leading to monopoly to benefit consumers.

Merger Policy Large firms that wish to merge must secure permission from the competition authorities, such as the European Commission or the Department of Justice and the Federal Trade Commission in the U.S. The competition authority will allow the merger in case the authority believes the merger will increase competition in the relevant territory, for example the U.S. or the E.U. Member States. If the authority believes that the merger will reduce competition, it will block the merger.

Both the European and the American competition authorities use a so-called *consumer welfare standard*, which means that they consider a merger anticompetitive if it harms the consumers, for example by increasing prices, and pro-competitive if the merger would benefit the consumers, for example by reducing prices.

There are two approaches to analyze the tradeoff between the anticompetitive effects (in the economic sense of the word) and the efficiency gains. The two approaches use different amounts of information.

- The first approach doesn't have a name, but may be called the *indirect method*. In this case the competition authority only assesses the likely anticompetitive effects, for example based on how much competition there is in the market before

the merger (as measured by e.g. the Hirfindhal index) and on estimates of the diversion ratio. If the anticompetitive effects appear to be large, the competition authority would simply presume that the anticompetitive effects dominate the efficiency gains and thus block the merger. The advantage with this approach is that the competition authority doesn't need to gather any evidence on the future cost savings that might come as a result of the merger.

- The second approach has a well-established name, namely the *efficiency defense*. In this case the competition authority not only gathers evidence about the effects on competition but also evidence on the efficiency gains. In practice it is the firms that have the burden to convince the competition authority that the merger will produce efficiency gains and that these gains are sufficiently large to dominate the anticompetitive effects found by the competition authority. The disadvantage with the efficiency defense is that it is more complicated and costly to assess efficiency gains on a case-by-case approach rather than to use presumptions about the average efficiency gains from mergers. The advantage with the efficiency defense is that the policy will not be based on presumptions about the average level of efficiency gains, but rather on evidence about the likely efficiency gains in the individual case.

If decisions are based on presumptions only, then mergers with relatively large anticompetitive effects will be blocked even if they would have generated large efficiency gains (error of type 1). Likewise mergers with modest anticompetitive effects would be allowed even if they don't generate any efficiency gains at all (error of type 2). That is, with an efficiency defense, the quality of the decisions should be better, but the cost of decision-making should be higher.

21.2.4. Formal Analysis using Cournot

The above qualitative analysis lacks in rigor. The most obvious shortcoming is that it is impossible to analyze the merging firms pricing behavior without at the same time incorporating the outsiders' response, and vice versa. Clearly, also the competitive effects and the efficiency gains are interdependent. Large cost saving make the merging firms more competitive and may put the outsiders at a disadvantage.

To include all these aspects in the analysis at the same time, it is necessary to analyze the effect of a merger using some oligopoly model. In this lecture we will only illustrate the idea and use the most simple oligopoly model, namely a linear Cournot model with homogenous goods. We will use this model to study the merging firms incentives to reduce output (thereby increasing price) and the rival firms reaction at the same time. We will allow for the possibility that the merging firms can lower their costs and study what the effect will be on the equilibrium price, output, profits and consumer welfare. The exercise is quite simple. To simulate the effect of the merger, we compute the Cournot equilibrium before the merger and the Cournot equilibrium after the merger and compare the two outcomes.

Actually, increasingly often, competition authorities rely on such formal models together with empirically estimated demand and cost functions. This is often referred to as merger simulation. When competition authorities use merger simulations they must clearly be more careful than we are here. The results of the analysis will depend on exactly which oligopoly model that is used. The Cournot model and the Bertrand model give different results. Thus, in real world cases, it is important to use the model that represents the actual market the best. It is clearly also important that demand and cost functions can be estimated accurately. But even with an appropriately chosen oligopoly model and well estimated demand

and cost functions, assessments of the likely efficiency gains will necessarily involve some guesswork.

Model Consider a market for a homogenous good with linear inverse demand given by $p = a - Q$ where Q is aggregate output. Assume that all firms have constant, but possibly different, marginal costs. Firm i's marginal cost is c_i. Firm i's profit can then be written $\pi_i = (a - Q - c_i) \cdot q_i$.

Some of the firms in the market have decided to merge. I will refer to the merging firms as the "insiders" and to all other firms in the same market as the "outsiders." There are $n_I \geq 2$ insiders. For simplicity, we assume that all the insiders have the same marginal cost c_I before the merger. The insiders' aggregate production is given by $Q_I = \sum_{i \in I} q_i$, where I is also used to denote the "set" of insiders. There are also $n_O \geq 1$ outsiders, all with marginal cost c_O. The outsiders' aggregate production is given by $Q_O = \sum_{i \in O} q_i$. Clearly the total output in the market is simply the sum of the aggregate production of the insiders and the outsiders, $Q = Q_I + Q_O$.

Pre-merger Equilibrium Although this is a standard Cournot model we will actually derive the equilibrium in a slightly different way than usual. In particular, we will focus on the aggregate production of the insiders, Q_I, and the aggregate production of the outsiders, Q_O. To do this, we will first compute how much the insiders produce in equilibrium, given any possible total output produced by the outsiders. (This "equilibrium relationship" is akin to a normal best-reply function.) Then we will compute the equilibrium aggregate output of the outsiders as a function of the aggregate output of the insiders. (Also this "equilibrium relationship" is akin to a best-reply function.) Finally we will combine these two equilibrium relationships to compute the overall equilibrium. Note however that this is just another way to compute the

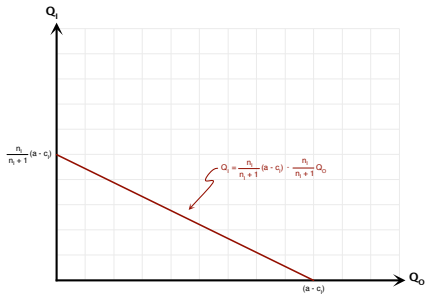
standard Cournot equilibrium, which you can verify by doing it the standard way.

An insider firm's profit before the merger is given by $\pi_i = (a - Q - c_I) \cdot q_i$. The first-order condition is given by $\frac{\partial \pi_i}{\partial q_i} = (a - Q - c_I) - q_i = 0$. (Derive the first-order condition as an exercise!) The aggregate equilibrium output of the insiders is then given by $Q_I = \sum_{i \in I} q_i = \sum_{i \in I} (a - Q - c_I) = n_I \cdot (a - Q - c_I)$. Using the fact that $Q = Q_I + Q_O$, we have $Q_I = n_I \cdot (a - Q_I - Q_O - c_I)$ and thus:

$$Q_I = \frac{n_I}{n_I + 1} \cdot (a - c_I) - \frac{n_I}{n_I + 1} \cdot Q_O.$$

Note that this function is very similar to a standard best-reply function. The difference is here that it captures the optimal collective behavior of one group of players (the insiders) as a function of the behavior of the collective behavior of another group of players (the outsiders). We may call it a "sub-group equilibrium-reply function". It is easy to see that it is linear as illustrated in Figure 21.7. Note

Figure 21.7.: Insiders' sub-group equilibrium-reply function



that the more the outsiders produce, the less the insiders would produce in equilibrium.[2]

The outsiders' equilibrium aggregate production as a function of the insiders' aggregate output can be derived in the perfectly

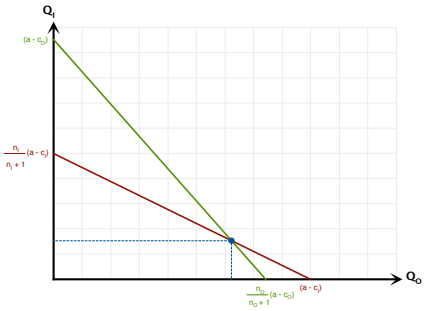
[2]The two groups aggregate output levels are strategic substitutes.

analogous way. It is given by

$$Q_O = \frac{n_O}{n_O + 1} \cdot (a - c_O) - \frac{n_O}{n_O + 1} \cdot Q_I.$$

The result is illustrated in Figure 21.8.

Figure 21.8.: Insiders' and outsiders sub-group equilibrium-reply functions



Clearly, in equilibrium, both groups should be producing their equilibrium aggregate output given the aggregate output by the other group. The equilibrium is therefore illustrated by the point of intersection between the two lines in the figure.

The algebra The equilibrium is defined by the two equations

$$Q_I = \frac{n_I}{n_I + 1} \cdot (a - c_I) - \frac{n_I}{n_I + 1} \cdot Q_O$$

and

$$Q_O = \frac{n_O}{n_O + 1} \cdot (a - c_O) - \frac{n_O}{n_O + 1} \cdot Q_I$$

which may be solved for the two unknown aggregate production levels Q_I and Q_O. The result is given by

$$Q_I = n_I \cdot \left[\frac{n_O + 1}{n_I + n_O + 1} \cdot (a - c_I) - \frac{n_O}{n_I + n_O + 1} \cdot (a - c_O) \right]$$

and

$$Q_O = n_O \cdot \left[\frac{n_I + 1}{n_I + n_O + 1} \cdot (a - c_O) - \frac{n_I}{n_I + n_O + 1} \cdot (a - c_I) \right]$$

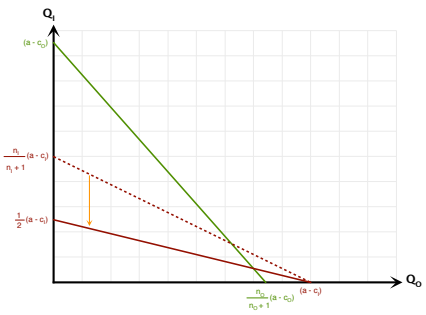
Total output in the market is given by

$$Q = \frac{n_I}{n_I + n_O + 1} \cdot (a - c_I) + \frac{n_O}{n_I + n_O + 1} \cdot (a - c_O)$$

The equilibrium price is obtained by substituting total output into the inverse demand function. (Derive aggregate production as an exercise!)

Post-merger Equilibrium After the merger, the there is only one insider, *i.e.* $n_I = 1$. This means that the insiders' sub-group equilibrium-reply function is pushed inwards, signifying the fact that the insiders want to produce less, for any given output produced by the outsiders. The result is illustrated in Figure 21.9. This is the internalization effect discussed above.

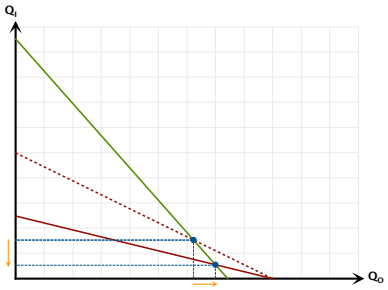
Figure 21.9.: Merger



The outsiders "sub-group equilibrium-reply function" is unaffected by the reduction of the number of insider firms (see equations above). Clearly, the outsiders only care about how much the insiders produce, not how many they are per se. Thus the effect of the merger can be described by the move between the pre-merger

and post-merger equilibria in in Figure 21.10.

Figure 21.10.: How equilibrium moves as result of merger



The insiders reduce their output to increase the market price. The outsiders benefit from increased residual demand and increase their aggregate output somewhat. This is the outsiders' response. Naturally this will tend to push the market price down. But, it is easy to see in the figure that total output is reduced in the market, since the insiders reduce output more than the outsiders increase their aggregate output. This means that the equilibrium price must be increased.

The Algebra Assume for the moment that insiders and outsiders have the same marginal cost, $c_I = c_O = c$. Then total output is given by

$$Q = \frac{n_I + n_O}{n_I + n_O + 1} \cdot (a - c)\,.$$

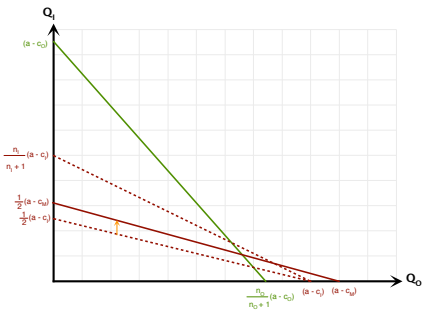
Clearly, when the number of firms is reduced as a result of the merger, total output is reduced, which must result in a higher price.

Exercise 65. Prove that a merger without any efficiency gains reduces aggregate production and increases price.

Efficiency Gains Assume now that the merged firm can reduce marginal cost, for example as a result of pooling knowledge. In

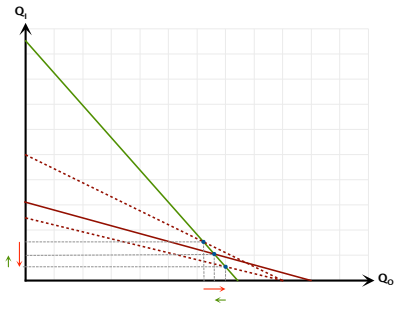
particular, the cost is reduced to $c_M < c_I$. This effect is illustrated in Figure 21.11.

Figure 21.11.: Cost reduction



Lower cost implies that the merged firm wants to expand production (for any given aggregate output by the outsiders) as shown by the outward shift of the "best-reply function". In equilibrium the cost reduction will also imply that the merged firm increase output, as described by the green arrow below. In response the outsiders will reduce their output somewhat. This effect is illustrated in Figure 21.12. The figure is based on the assumption that the reduction

Figure 21.12.: How equilibrium moves as a result of merger and cost reduction



is marginal cost is relatively modest which implies that anticompetitive effect (red arrows) dominate the pro-competitive effect (green arrows). In this case total output is reduced and prices increased.

If however the cost reduction is sufficiently large, the merged firm will actually produce more than the insiders did before the merger. In this case the outsiders will reduce output, but total output is increased. This possibility is illustrated in the in Figure 21.13. If

Figure 21.13.: Large cost reduction

the cost reduction is large enough, the anticompetitive effect (red arrows) is dominated by the pro-competitive effect (green arrows). In this case total output is increased. (Exercise: What happens to price in this case? How do you know?)

21.2.5. Profitability Paradox

One of the most surprising results of merger analysis is that a merger can actually be unprofitable. One would have guessed that a merged firm should always be able to replicate what the previously independent firms used to do, guaranteeing that profits should be at least as high as before the merger. This intuition is not correct, however.

The reason is that the merger also triggers a reaction from the outsiders. If two Cournot firms merge they will have the incentive to reduce output. The outsiders know this. The outsiders therefore will have an incentive to expand their output. And the outsiders' expansion of output implies that the insiders' residual demand is

reduced. That is, a merger may reduce the insiders' residual demand.

To see that a merger may be unprofitable, consider a Cournot market with inverse demand $P = 1-Q$ and identical firms with constant marginal cost c.Total equilibrium output is given by $Q(n) = \frac{n}{n+1} \cdot (a-c)$ and the equilibrium price is given by $P(n) = \frac{1}{n+1} \cdot a + \frac{n}{n+1} \cdot c$. The equilibrium profit of an individual firm is given $\pi(n) = \frac{1}{(n+1)^2} \cdot (a-c)^2$.

Assume that two firms merge. Before the merger their aggregate profit is given by

$$2 \cdot \pi(n) = \frac{2}{(n+1)^2} \cdot (a-c)^2 .$$

After the merger, the merged firm earns

$$\pi(n-1) = \frac{1}{n^2} \cdot (a-c)^2 .$$

The merger reduces profit if $\pi(n-1) < 2 \cdot \pi(n)$, which is the case if $(n+1)^2 < 2 \cdot n^2$. It is easy to verify that this condition is indeed fulfilled whenever $n > 3$.

That is, a merger from Cournot duopoly to monopoly would be profitable. But a merger from Cournot triopoly to Cournot duopoly would be unprofitable for the insiders. This means that some form of efficiency gains are necessary for mergers to be profitable in Cournot markets. We will get back to this point below.

21.3. Merger Formation

The traditional perspective on mergers, sometimes referred to as exogenous merger theory, is to only study the effects of single mergers, viewed in isolation. As we have just seen, the traditional perspective is useful for understanding how a particular merger influences the prices customers have to pay, and how the profits in the in-

dustry are affected, compared to the situation before the merger. But many other questions such as when to expect international takeovers and when to expect national firms to merge instead, why we need to control mergers, and why the competition authorities only care about the consumers, traditional analysis cannot answer.

To understand the firms' incentives to merge, the analysis must be broadened; individual mergers must be put into context. There are often alternatives to any specific deal, such as mergers with other partners. Some mergers are mutually exclusive, giving rise to takeover battles, as when Cingular and Vodafone both bid for AT&T Wireless. Other mergers may be complementary: If one firm buys another, the acquisition of third company by a fourth may be more profitable, leading to a merger wave.

Not only are there many different alternative mergers, with different consequences for the merging parties, but all the different alternatives also affect the firms outside the merger. Externalities may be positive, as when some firms join to reduce competition in the market. Also the outsiders benefit from lower competition. Externalities may be negative, as when a new combination of assets makes the merging firms more competitive. Recall the European Commission's worries that a merger between GE and Honeywell would "bundle" engines and avionics in packages that other firms couldn't match.

The firms and their owners fare better if they consider all possible alternatives, and if they take into account the possible moves and countermoves that their rivals may take. To delineate the firms' incentives in such a highly interactive environment, one cannot focus on any particular merger, or any particular firm's situation. The incentives of one firm are very much influenced by what it believes its rivals will do. The incentives of the rivals will, in turn, depend on what they believe their competitors will do. The situation must be examined as a whole.

The new framework In recent times a new framework has emerged to study these problems. The so-called endogenous merger theory takes the results from traditional theory as a stepping-stone. Equipped with an idea of how different mergers would affect the profits of the firms in an industry, the fundamental question is which firms will merge, when they will make their bids, and how will the parties split the surpluses (if any) from these transactions?

Endogenous merger theory is novel also for another reason. It builds on the theory of coalition formation, a branch of game theory that has not yet found many applications in economics. This lecture will cover a few examples of results from this theory and explain how they can help us understand some current public policy issues.

21.3.1. Do we need to Control Mergers?

Merger policy is costly. There are of course some direct costs such as paying the wages for the public servants working at the competition authorities. Also the firms are incurred costs for notifying mergers to the competition authorities and for gathering evidence to convince the authorities that the merger is pro-competitive. Then there are also hidden costs, such as when competition authorities take the wrong decisions and block mergers that would have been socially efficient, or when firms don't even initiate mergers that they believe would raise objections. Also the disciplining effect of merger on executive slack can be lessened if it is known that takeovers are unlikely as a result of merger policy.

And, actually, merger policy might not even be necessary. The market might solve the potential problem by itself.

Merger Profitability Consider a potential merger without efficiency gains. If the merger is undertaken, the merged entity would increase prices (since marginal cost is not reduced) and reduce output. Existing competitors would then increase their output and

new firms would enter the market attracted by the higher prices. As a result of the outsiders' reaction, the insiders' residual demand is reduced. Such mergers are often unprofitable.

Consider now a potential merger with efficiency gains enabling the merged firm to reduce marginal cost, and assume that the cost reduction is sufficiently large that the merged firm will have an incentive to lower price and expand production, relative to the situation before the merger. Such mergers are often profitable.

Since anticompetitive mergers are often unprofitable and pro-competitive mergers are profitable, one may assume that the mergers that are actually initiated should typically be pro-competitive. Had this been the complete story, merger policy would have been superfluous.

Merger Holdup Unfortunately, however, it is straightforward to construct theoretical examples demonstrating that mergers can be anticompetitive and profitable at the same time. An example is a merger that does not reduce marginal cost, implying that price will be increased. Such a merger may be profitable *e.g.* if the merging parties can reduce administration and lower other fixed costs. Thus, even if the market might solve part of the problem, we might still fear that also some anticompetitive mergers would occur absent merger control.

But, actually, there is an argument also against this fear. If mergers are anticompetitive in a certain market, it means that the merging firms would reduce their output in order to increase price. But also the outsiders would gain from the price increase without having to reduce their output. On the contrary they typically increase production. Thus, the outsiders are in the best possible situation, enjoying a free ride.

But if all firms hope for the free ride, merger activity will be held up. Again, the market may provide little incentives for anticompetitive mergers. The next section introduces some simple theoretical

models to show that the hold-up problem does reduce, but does not eliminate completely, the incentives for anticompetitive mergers.

Formal Analysis of Merger Holdup (Simultaneous Bids) Consider a market with three symmetric firms, each earning the triopoly profit π^3. A merger among all three firms to form a monopoly would be profitable due to reduced competition and since there is no outsider who could expand production. That the merger to monopoly is profitable is the same as saying that the monopoly profit π^1 is larger than three times the triopoly profit, *i.e.*

$$\pi^1 > 3 \cdot \pi^3.$$

This inequality holds true in almost any oligopoly model, suggesting that it is probably true in most real-world markets.

Assume that a merger between only two firms, to create a duopoly, would be anticompetitive. This means that the merger would increase the price, which would benefit the outsider. For simplicity we assume that both firms following such a merger would earn the same duopoly profit π^2. This assumption is clearly unrealistic, but it simplifies the analysis. And it is easy to show that the same results can be obtained without the assumption. In any case, the fact that a merger between two firms is anticompetitive and beneficial to the outsider means that

$$\pi^2 > \pi^3.$$

Now we are interested in investigating if the firms can arrange a merger to monopoly.

To study this issue we consider a very simple merger game. The rules of this game stipulate that one firm, called the acquirer, is allowed to make a bid for the two competitors and that this bid b must be the same to both firms. Then the two targets are allowed

to either accept or reject their offers. These decisions are taken simultaneously by the two targets.

To find the subgame perfect equilibrium of this game, we first have to analyze the second period. The second period can be described by the following payoff matrix. If both targets accept, they

Table 21.1.: A merger game

	Accept	Reject
Accept	b, b	b, π^2
Reject	π^2, b	π^3, π^3

will both get the bid. If both firms reject, the market remains a triopoly and both firms will get the triopoly profit. If one firm accepts and the other rejects, the market is transformed into a duopoly and the rejecting firm will get the duopoly profit. To arrange a merger to monopoly, the acquirer has to offer a bid b that both targets accept. Both firms accepting the bid would indeed be an equilibrium in case $b \geq \pi_2$. And since the acquiring firm wishes to maximize its own payoff, if would bid $b = \pi_2$ in case it would bid at all.

The question is now if the acquiring firm would wish to make a bid for the two competitors in period 1. By buying the competitors the acquiring firm would become a monopolist earning π^1 but the price would be to pay $2 \cdot \pi^2$. By not buying the competitors the triopoly remains and the firm would earn π^3. Clearly, then, the acquiring firm would make the bid if, and only if,

$$\pi^1 - 2 \cdot \pi^2 \geq \pi^3.$$

This condition may be rewritten as

$$\pi^1 \geq \pi^3 + 2 \cdot \pi^2.$$

Sine the duopoly profit is larger than the triopoly profit, $\pi^2 > \pi^3$, it follows that also some profitable mergers to monopoly may not

be possible to arrange. The reason is that it would be profitable for each target to stand outside the merger and let the other firms merge instead. Therefore the targets demand a large premium to participate in the merger. And, in the end, there may be no merger at all.

Formal Analysis of Merger Holdup (Sequential Bids) Consider now a slightly more realistic model of the acquisition process, requiring the acquiring firm to buy the targets one by one rather than at the same time. In period one the acquirer can bid some amount b_1 for target 1 and in period two the acquirer can bid some other amount b_2 for target 2. Again we need to analyze the game backwards to find the subgame perfect equilibrium.

Period 2 Assume first that the acquirer successfully bought the first target in period 1. Then the second firm is willing to sell if offered $b_2 = \pi^2$ or more. The acquirer will offer exactly this amount, assuming that $\pi^1 > 2 \cdot \pi^2$.

If the acquiring firm did not buy the first target in period 1, there will be no acquisition in period 2 assuming that $\pi^2 < 2 \cdot \pi^3$. In this case there is no price that would make both the acquiring firm and the target better off. Thus, if the acquiring firm doesn't buy the first competitor in the first period, the triopoly will also remain in the second period.

Period 1 In the first period the first target is willing to sell in case it is offered $b_1 = \pi^3$ or 1 more. The reason why the first target does not demand more than the triopoly profit is that if it rejects the offer, the triopoly will remain since, in this case, there will be no merger in the second period either. The acquirer will make the bid, assuming that $\pi^1 - b_1 - b_2 \geq \pi^3$ *i.e.* assuming that

$$\pi^1 \geq 2 \cdot \pi^3 + \pi^2.$$

There is thus less frictions in the merger process if acquisition are made sequentially. Still, some mergers for monopoly will not occur, even if they are profitable.

21.3.2. What Should the Goals of Merger Control Be?

The previous section indicates that even if the market probably discourages many anticompetitive mergers, since some are unprofitable as a result of the outsider response and since some may not be possible to arrange as a result of the holdup problem, there may still be room for a merger policy. The question is then how to design such a policy.

We have already touched on one design issue: the question if the competition authority should use an efficiency defense (i.e. to estimate the likely efficiencies on a case-by-case basis) or if they should only take efficiencies into account indirectly by only blocking mergers with so adverse anticompetitive effects that these effects can be assumed to dominate the efficiency gains.

Another design issue is what exactly the goals of the merger policy should be. Most competition authorities around the world use a so-called consumer welfare standard. This means that they only allow mergers that benefit the consumers. The consumer welfare standard implies very high demands on efficiency gains in order to allow a merger. Many economists have criticized the consumer surplus standard arguing that a policy which does not give any weight to firms' profits will be inefficient. Big companies are to a large extent actually owned by pension funds. Thus, blocking mergers may benefit the citizens in their capacities of consumers, but harm the same people when their pensions are paid out. According to this argument it would be better with a total welfare standard. If a merger increase profits (i.e. pensions and capital incomes) more than it hurts the people in the form of price increases, then the

"income effect" dominates the "price effect" and the merger should be allowed. In addition, any concerns for the distribution of income should be solved using taxes and subsidies.

Oddly enough, there is actually a reason why competition authorities should pursue a consumer surplus standard, even in case they wish to maximize total welfare. The reason is that competition authorities only have the right to assess notified mergers, but cannot suggest other mergers. Moreover, when they evaluate a notified merger, they can only compare the merger with the situation before the merger. They cannot compare the suggested merger with some alternative merger. (The reason for the second limitation is that it would be very difficult to gather information about other hypothetical mergers.) As a result of these limitations, it may be useful to be tough. The consumer welfare standard implies that a merger has to produce very high efficiency gains in order to be approved. With the total surplus standard more modest efficiency gains may be sufficient. If the firms in the industry are choosing between different mergers, some of which are profitable as a result of their anticompetitive effects, and some of which are profitable as a result of efficiency gains, then the tougher policy may force the firms to choose the mergers that are profitable for the right reasons.

Formal Analysis Consider a market where two mergers are possible. (Assume *e.g.* firm A either merges with B or with C, but that a merger between B and C for some reason would not make sense.) For simplicity we assume that the industry will select the merger maximizing aggregate profit.

Assume that before any mergers take place, in status quo, the aggregate profit is zero and the consumer surplus is €100 implying a total surplus of €100.

One of the possible mergers is anticompetitive. It reduces competition and only generates efficiency gains in the form of fixed cost savings that do not benefit the consumers. This merger would

increase aggregate profit to €100 but reduce consumer surplus to €50. The total surplus would be €150.

The second possible merger is pro-competitive. It generates large reduction in marginal cost and benefits also the consumers. With this merger profits are only increased to €50 but the consumer surplus is increased to €150. The total surplus would thus be €200. The situation is summarized in the following table.

Table 21.2.: Merger policy

	Industry profit	Consumer surplus	Total surplus
Status Quo	0	100	100
Merger 1	100	50	150
Merger 2	50	150	200

With a total surplus standard, the competition authority would approve both mergers. The industry, which proposes the merger maximizing total industry profits, subject to approval, will then go for the anticompetitive merger. In the end the total surplus would be €150 with the total surplus standard.

With a consumer surplus standard, the competition authority would only approve the pro-competitive merger. The industry, which proposes the merger maximizing profit, subject to approval, will in this case go for the second merger. In the end the total surplus would be €200 with the consumer surplus standard.

Thus, even if the ultimate goal for merger control is to maximize total welfare, it may actually be better that the competition authority uses a consumer welfare standard when making its decisions. By being tough, the competition authority forces the firms to select a merger that is profitable for the right reasons. This is an example of so-called *strategic delegation.* Strategic delegation means that it is sometimes better for a principal (here, the government) to select an agent (here, the competition authority) who has a goal that differs from the principal's own goal. Another example of strategic delegation may be that labor unions appoint radical

leaders, who care more about increasing wages and less about the unemployment that higher wages may cause.

21.3.3. Is Merger Policy Successful?

Is Merger Policy Successful? To evaluate if merger policy is successful or not, one would need to answer two questions: (i) are allowed mergers pro-competitive and (ii) are blocked mergers anticompetitive? The first question might be possible to answer by studying how prices are changed as a result of the merger. Clearly such studies are plagued by difficult methodological problems, such as eliminating the effects of other changes in the market that might affect prices and that occur during the study period. But the second question appears to be completely impossible to answer. How should one go about estimating the effect of something that did not happen?

A clever trick Actually there is a clever trick that might be used to answer this question. This trick relies on two ideas. First, recall that a merger that hurts the consumers, *e.g.* as a result of increased price, typically benefits the outsiders, *i.e.* the competitors to the merging firms. Likewise, a merger that benefits the consumers typically hurts the outsiders. Second, recall that a firm's stock market value is equal to the expected discounted value of its future profits. When a merger is announced the stock market values of the firms in the industry quickly adjust to incorporate the expected changes in the firms' profits. If the outsiders' stock market values are increased it must mean that the market expects the merger to increase the outsiders' profits. But if the outsiders' values are reduced it must mean that the market expects the merger to reduce their profits. In the first case we may conclude that the merger is anticompetitive and would reduce consumer welfare. In the second case the merger is pro-competitive.

Results Actually several studies using this event-study methodology have been undertaken. The result is that competition authorities appear to make many mistakes of both types, both blocking pro-competitive mergers and allowing anticompetitive mergers.

Critique of event studies But event studies can be criticized on several grounds. One critique is that the event studies underestimate the value of merger control, by not taking into account that many anticompetitive mergers are probably never even proposed, since the firms know that they would be blocked.[3]

A second critique is that a pro-competitive merger may actually increase the stock-market value of outsiders. The reason is that the first merger may inform the market that firms may realize large efficiency gains through mergers and that therefore also the outsider of the current merger will soon merge and reap similar benefits.

A third critique argues that anticompetitive mergers may actually reduce the stock market value of outsiders. To see this point, consider an industry with three firms. Assume that a merger allows the insiders substantial savings on their fixed costs. Then, even if the merger is anticompetitive, it may still be better to be an insider than an outsider. In this case all firms will compete to buy one of the rivals, to become an insider rather than an outsider. If the stock market is informed about the take-over battle, all firms will have an intermediate stock market value reflecting both the probability of being an insider, earning a high profit in the future, and the probability of being an outsider, earning a somewhat lower profit in the future. When a merger between two firms is announced, the merger is informed that the outsider lost the takeover battle. The stock market value of the outsider falls, since the probability of becoming an insider has been eliminated.

[3]This point is made in the paper by Lars Sörgard which will be discussed later during this course.

Vertical Restraints

To be added.

Exclusion

To be added.

Sector Regulation

Ramsey pricing

Network externalities

Part VIII.

Game Theory

This is a course in microeconomics, focusing on how markets work. We will not talk much about perfectly competitive markets, as those are rare in reality. We want to get closer to real-world markets, markets with limited competition and where both buyers and sellers have limited information. In such markets firms have the ability to influence the prices at which they trade – an ability often referred to as market power. Broadly speaking our aim is to understand why firms have market power and what the consequences of market power are.

The most important source of market power is probably that many markets are highly concentrated. There are still a few monopolies left, for examples the Swedish train operator SJ and until recently the pharmacy retailer Apoteket. But oligopolies are much more common. Examples include food retailing with three main actors (ICA, Coop and Axfood/Hemköp), telecommunications (TeliaSonera, Telenor, Tele2 and Three) and television (SVT, TV4 and couple of big others). The prices charged by these firms have a large effect on most peoples' real income. The safety of the drugs you buy at Apoteket, the taste and variety of the food you find in the ICA store, the quality of the TV-programs you spend 2 hours a day watching, all have a substantial effect on your well-being.

It is therefore important to understand how these firms make their choices, what prices they set and what qualities they provide. It is also important to understand if it is possible to improve the market outcome by public intervention. Until recently, Apoteket was the sole provider of medicines in Sweden. But the Swedish Government opened the market for competition. Monopoly was replaced by oligopoly. Will this reform improve welfare in Sweden?

One of the key complications in analyzing concentrated markets is that the firms' behaviors are interdependent. The prices and qualities chosen by one firm depend on the prices and qualities chosen by their competitors. If Coop reduces its prices, ICA will have to lower theirs, not to loose their customers. To predict the

firms' pricing, we cannot simply assume that firms maximize their profits. To predict ICA's prices we first need to know how Coop will behave. But that we cannot know before we know how ICA will behave. This may seem as a paradox or to require circular reasoning. Not so, fortunately! This is where game theory comes in. Game theory provides us with the tools we need to analyze interdependent behavior. The "theory of interdependent decision-making" would even be a better name for what now is known as "game theory." We will start our discussion with static games. Static games depict situations where all decision-makers make all their choices at the same time.

22. Static Games

In presenting game theory we will go through a series of famous "toy games." The appeal of these games is that they capture some important aspect of human interaction that arises in many different situations. They are also sufficiently simple to learn basic game theoretic concepts and analysis. Toy games are too stylized, however, to teach us everything about any particular real-world situations, such as the interaction between buyers and sellers in markets. We start with the most famous of all toy games, the prisoners' dilemma.

22.1. Game 1: Prisoners' Dilemma

A crime takes place and the police arrest two suspects. The evidence is sufficient to get the suspects convicted for a minor part of the offense, which would result in short prison sentences for both, say one month each. But the police need more evidence to get the suspects convicted for the more important parts of the crime, which would result in much longer sentences, say ten months.

One possibility would be to make the two suspects confess and tell all they know. But, why would they voluntarily reveal information that would increase their punishments?

A smart prosecutor comes up with an idea. He asks the police to place the two suspects in separate cells. Then the prosecutor visits each of them, offering them the same deal. If one criminal testifies against the other ("rats") and the other remains silent ("clams"), the betrayer goes free and the other criminal receives the full ten-month sentence. If both suspects betray the other, each receives

a five-month sentence. If both remain silent, both are sentenced for the minor charge to one month in prison. Each suspect must choose to rat or clam, ignorant about the other suspect's choice.

The suspects are not friends; they just did this job together. They don't care for one another and they will never meet again after court. Each suspect just wants to minimize his own sentence.

Please answer the exercise before reading any further.

Exercise 66. Put yourself in the shoes of one of the suspects. How would you act in this situation?

To help make up your mind, the whole situation is summarized Table 22.1 . The two rows indicate prisoner 1's two choices, to clam

Table 22.1.: Prisoners' Dilemma (months in jail)

		Prisoner 2	
		Clam	Rat
Prisoner 1	Clam	1, 1	10, 0
	Rat	0, 10	5, 5

and to rat. The two columns indicate prisoner 2's two choices. The first entry in each cell indicates prisoner 1's sentence in case the corresponding row and column where chosen. The second entry is prisoner 2's sentence. For instance, the lower left cell shows that prisoner 1 would go free and prisoner 2 would get 10 month in jail in case prisoner 1 rats and prisoner 2 clams.

The "common good," from the point of view of the prisoners, would be that they both clam. Then they will both spend only a month in jail. That is probably also what they would agree to do, if they could. But, is this the outcome we should expect? Will they both remain silent?

Let's assume that the two prisoners indeed had agreed, before being caught, not ever to cooperate with the police. If prisoner 1 expects his accomplice to honor their agreement, prisoner 1 faces

the choice between a one-month sentence and being set free. Freedom comes at the cost of sending the companion to nine extra months in prison, but since prisoner 1 doesn't really care about his accomplice, he will rat.

At this stage of his analysis, prisoner 1 probably realizes that his accomplice may have exactly the same thoughts, and that the accomplice might actually rat, despite their agreement. But then prisoner 1 also realizes that his own analysis would be flawed, as it is based on the presumption that the accomplice will clam. What should he do if the companion rats? Looking in the matrix again, prisoner 1 realizes that he faces the choice between ten and five months in prison, assuming that the accomplice would rat. Again prisoner 1 concludes that it would be better for him to rat.

Prisoner 1 is thus better off ratting, independent of how the other prisoner behaves. We should consequently expect prisoner 1 to rat. And the same goes for prisoner 2.

The only reasonable prediction is that both prisoners will rat and provide the police with all the necessary information for charging them with the full crime and that they will both receive the five-month sentence. It doesn't matter what they agreed before being caught. Sitting in the cell, they will both have an incentive to rat. It doesn't matter that they both end up with five months in prison rather than just one. They both minimize their own sentence by ratting.

Each individual suspect has an incentive to cheat on the other, to receive a personal rebate on the full sentence, but thereby increasing the sentences for the other *by more* than the own rebate. Taken together, the plus and the minus, adds time to the suspects' sentences.

22.1.1. Payoff functions and bi-matrixes

There is an alternative, but equivalent, way to summarize the situation for the two prisoners. Instead of specifying the sentences, we may specify the prisoners' payoffs (or utilities). Since they only care about their own sentence, the payoffs can for example be represented by: $u = 10 - months$. This utility function is of course arbitrary to some extent. We could e.g. have chosen some other constant (say 11 instead of 10), or we could have measured the sentence in days or minutes in stead of months. The only important property is that the payoff is higher the fewer months the suspect receives and that the payoff is unaffected by the number of months that the accomplice receives.

It is common to describe two-person games with only a few possible actions in a matrix (also called bi-matrix). By convention, the first player is the row-player, so we don't need to specify this. Table 22.2 thus describes the Prisoners' dilemma game with payoffs

Table 22.2.: Prisoners' Dilemma (utilities)

		Prisoner 2	
		Clam	Rat
Prisoner 1	Clam	9, 9	0, 10
	Rat	10, 0	5, 5

instead of months in jail. Please do the Exercise before reading any further.

Exercise 67. Explain what the second 9 in the upper left cell of the Prisoners' dilemma game signifies!

From now on, we will assume that the two prisoners want to maximize their payoffs (or utilities). Notice, however, that this is simply another way to say that they wish to minimize their sentences.

22.1.2. Best replies and best-reply functions

A simple *procedure* to predict the prisoners' behavior is to find their best-reply functions. This procedure can be used to analyze most games in this course.

We start with player 1's best-reply function.

1. Consider the possibility that prisoner 2 will clam. Which is the highest payoff player 1 could achieve in this case? The clam column shows that it is 10 (going free). Underline this highest payoff for player 1 in the clam-column. To achieve this payoff, player 1 has to rat. We therefore say that ratting is *a best reply* to clamming.

2. Consider the possibility that prisoner 2 will rat. Which is the highest payoff player 1 could achieve in this case? The clam column shows that it is 5 (spending 5 months in jail). Underline this highest payoff for player 1 in the rat-column. To achieve this payoff, player 1 has to rat. We therefore say that ratting is *a best reply* to ratting.

Table 22.3 shows player 1's best reply to *every* possible choice by

Table 22.3.: Player 1's best-reply function

		Prisoner 2	
		Clam	Rat
Prisoner 1	Clam	9, 9	0, 10
	Rat	<u>10</u>, 0	<u>5</u>, 5

player 2. This is called player 1's *best-reply function*. The best-reply function can be viewed as a list describing the player's best choice for every possible choice that the other player(s) can make. Please note the difference between a best reply and a best-reply function.

In most interesting situations a person's best choice depends on what the others do. Not so in the prisoners' dilemma game. In this

game the best reply function is constant – ratting is the favored alternative independent of what the other suspect does, i.e. both if the other suspect clams and if he rats.

Next, we find the best-reply function for player 2. The procedure is the same as for player 1, so try to do it yourself, in the matrix above. Please do the exercise before reading any further.

Exercise 68. What is player 2's best reply-function in the Prisoners' dilemma game?

Answer: Consider first the possibility that prisoner 1 will clam and underline the highest payoff for prisoner 2 in the clam-row. Consider next the possibility that prisoner 1 will rat and underline the highest payoff for prisoner 2 in the rat-row. The result is displayed in Table 22.4, together with player 1's best-reply function. ◀

22.1.3. Equilibrium

The game-theoretic prediction of how the two prisoners will behave is simply the cell where both prisoners' payoffs are underlined, i.e. the lower right cell in Table 22.4. This cell corresponds to both

Table 22.4.: Equilibrium

		Prisoner 2	
		Clam	Rat
Prisoner 1	Clam	9, 9	0, <u>10</u>
	Rat	<u>10</u>, 0	<u>5</u>, <u>5</u>

prisoners ratting. In this cell both prisoners maximize their own payoff, given the behavior of the other prisoner. We say that this is the equilibrium of the game. The basic idea of game theory is that people will behave according to such an equilibrium.

Also note that no other cell has this property, that the players' choices are mutual best replies. Thus, if we believe that people do

behave according the equilibrium of the situation we will be able to predict their behavior.

22.1.4. Insights

Social efficiency vs private incentives The Prisoners' dilemma game serves to illustrate two important insights. The first insight is that there is often a conflict between private incentives and social efficiency. That is, in many situations, the players' incentives will not lead them to do the common good. A possible outcome of the situation is that both prisoners serve only one month in jail. But, by following their self-interest, the prisoners will actually end up both serving five months in jail.

Again, each individual suspect has an incentive to cheat on the other, to receive a personal rebate on the full sentence, but thereby increasing the sentences for the other *by more* than the own rebate. Taken together, the plus and the minus, adds time to the suspects' sentences.

Dominance The second point is that even though the outcome depends on the actions taken by several people, it is sometimes possible for those involved to find out what best to do without knowing what the others are going to do. Each prisoner has a so-called dominant strategy – a strategy that is better than all other options, independent of what the other prisoner will do. Ratting is better independent of whether the other suspect rats or clams.[1]

[1] It is quite rare that people have a dominating strategy in interesting social situations. What is common however is that they have dominated strategies. In the prisoners' dilemma game clamming is dominated by ratting since clamming gives a lower payoff than ratting independent of what the other suspect does.

22.2. Normal Forms

A game is a highly stylized description of real some world situation. The rules of the game should capture the essentials of the situation, i.e. the key actors and the key determinants of their behavior. In all other respects, game theoretic models must be made as simple as possible. Otherwise we cannot solve them.

Ingredients of a normal form game A game in normal form specifies only three things, namely (i) who the players are, (ii) what their strategies are and (iii) their payoffs for all possible combinations of strategy choices.

In the Prisoners' Dilemma there are two ***players***, namely the two prisoners.

Both players have two strategies, namely to rat and to clam. The payoffs were summarized in the payoff matrix above. More formally, let S_i denote the set of ***strategies*** available to player i. In the Prisoners' Dilemma, the prisoners can choose between ratting and clamming, i.e. $S_i = \{clam, rat\}$. A particular strategy is denoted by s_i.

If we combine a strategy for player 1 with a strategy for player 2 we have a pair of strategies, denoted (s_1, s_2). Such a pair is called a ***strategy profile***. In the Prisoners' Dilemma there are in total four such strategy profiles, namely

$$\{(clam, clam), (clam, rat), (rat, clam), (rat, rat)\}.$$

This set of strategy profiles is essentially a list of all possible outcomes of the game. For instance, the profile (clam, rat) means that player 1 clammed while player 2 ratted.

Let u_i denote player i's ***payoff*** (or utility). Clearly each player's utility depends on what both players do, which is to say that the utility is a function of the two players strategies. To indicate this

dependence, we write: $u_i(s_1, s_2)$. For example, in the Prisoners' dilemma, player 1's utility function is the following (the first argument is player 1's strategy): $u_1(rat, rat) = 5$, $u_1(rat, clam) = 10$, $u_1(clam, rat) = 0$, $u_1(clam, clam) = 9$.

Information In normal form games, it is assumed that all players choose their strategy without knowing the other players' choices. It doesn't matter if the choices are truly simultaneously or if one player acts after another, as long as the late-mover cannot observe the early-movers' choice before choosing himself. In game-theoretic language, both possibilities are called simultaneous.

For now, we assume all players to know not only their own strategy sets (available options) and payoffs, but also all other players' strategy sets and all other players' payoffs. Such games are called games of complete information.

Nash equilibrium The most common method to predict the players' behavior is to find the so-called Nash equilibria of the game. A Nash equilibrium is defined at a strategy profile, specifying a strategy for each player, such that every player maximizes his utility by following his strategy, given that the other players follow their strategies.[2] In a two-player game, the strategy profile (s_1^*, s_2^*) is an equilibrium if,

- Player 1 maximizes his payoff given player 2's behavior. Using the notation introduced above this means that $u_1(s_1^*, s_2^*) \geq u_1(s_1, s_2^*)$ for all strategies s_1 in the strategy set S_1.
- Player 2 maximizes his payoff given player 1's behavior. Using the notation introduced above this means that $u_2(s_1^*, s_2^*) \geq u_2(s_1^*, s_2)$ for all strategies s_2 in the strategy set S_2.

[2]The name Nash equilibrium is in honor of John Nash, one of the most important contributors to game theory.

Notice that a Nash equilibrium only requires the players to choose a strategy that gives them a payoff that is *at least as good* as all other strategies, given what the others do. There may exist other strategies that are equally good.

In case a game has a strategy profile (s_1^*, s_2^*) such that $u_1(s_1^*, s_2^*) > u_1(s_1, s_2^*)$ for all strategies s_1 in the strategy set S_1 and $u_2(s_1^*, s_2^*) > u_2(s_1^*, s_2)$ for all strategies s_2 in the strategy set S_2, then (s_1^*, s_2^*) is called a *strict* Nash equilibrium.

Interpretation of equilibrium The normal (but not only) interpretation of the Nash equilibrium is that two conditions must be satisfied.

1. **Rationality**: Rationality means that (i) players form expectations about what the other players will do and that (ii) they choose the strategy giving them the highest utility, given their expectations. For example, let $E_1 s_2$ be the action that player 1 believes that player 2 will take. Then, rationality requires player 1 to chose an action s_1^* that satisfies $u_1(s_1^*, E_1 s_2) \geq u_1(s_1, E_1 s_2)$ for all strategies s_1 in S_1.

2. **Coordination**: Coordination means that the players have correct expectations of what the other players will do. That is, $E_1 s_2 = s_2^*$.

On coordination It is important to understand that we have not answered the question why the players would have correct expectations in any situation. Such coordination does not follow from the rationality of the players. The Nash equilibrium simply presumes that the players for some reason or another do have correct expectations.

But why should we expect e.g. the prisoners to act according to the equilibrium? Ultimately this is of course an empirical question. But there are some reasons to use equilibrium analysis. If we believe

that it is possible for us, as outside observers, to predict the outcome in some social situation, then it seems likely that also the players themselves should be able to predict the outcome of that situation. But if that is the case, predictions must be *self-fulfilling* to make sense. That is, all players must have an incentive to actually behave according to the prediction. If not, they will do something else and the prediction will be refuted. Now, the only way a prediction could be self-fulfilling is that it is an equilibrium. This is easy to see in the Prisoners' Dilemma Game. For example, if the prediction would be that prisoner 1 rats and prisoner 2 clams, the prediction would be refuted since prisoner 2 would not want to follow the prediction.

But there are also some more substantial reasons why coordination may actually occur. As will be discussed below, coordination may be achieved through various different means. In certain types of situations people may learn over time how to behave. They learn what to expect form others and they learn what others expect from them. In other situations people may have the opportunity to agree in advance how to behave. If it is clear that nobody has an incentive to deviate from such an agreement, they may all trust that they will all honor the agreement. (But if they do not have the incentive to follow the agreement, the agreement is worthless, as demonstrated in the Prisoners' dilemma.)

These coordination mechanisms (learning and pre-play communication) are not part of standard game theoretic analysis, however. But, when we solve a game, using the idea of a Nash equilibrium, we should ask whether it is likely that the players really would be able to coordinate their expectations some way or the other.

Prisoners' dilemma, revisited Consider again the prisoners' dilemma, but now from the point of view of the prosecutor. The prosecutor wants to convict the suspects to the longest possible sentences. The problem is that he does't have sufficient information to prove the more serious parts of their crime. Given the information that he

has, the two suspects will both get one month in prison. But if at least one of them would rat, they could both be sentenced to ten month. That is the maximum punishment, absent any rebates.

Table 22.5.: Prisoners' dilemma without rebates (months in prison)

	Clam	Rat
Clam	1, 1	10, 10
Rat	10, 10	10, 10

Exercise 69. Find the Nash equilibria of the prisoners' dilemma without rebates.

Answer: First find the two suspects best-reply functions. Notice that if the other rats, then both ratting and clamming are best replies. Thus there are two Nash equilibria of this game, namely (rat, rat) and (clam, clam).◀

The prosecutor is not satisfied with this situation. Therefore, he starts to think about rebates. Let r_1 denote the rebate given to a suspect who rats on the other, if the other clams. Let r_2 denote the rebate given to a suspect who rats on the other, if also the other clams.

Table 22.6.: Prisoners' dilemma with rebates (months in prison)

	Clam	Rat
Clam	$1, 1$	$10, 10 - r_1$
Rat	$10 - r_1, 10$	$10 - r_2, 10 - r_2$

Exercise 70. What rebates r_1 and r_2 does the prosecutor need to give in order to:

1. Guarantee that (Rat, Rat) is a Nash equilibrium?
2. Guarantee that (Rat, Rat) is a strict Nash equilibrium?
3. Guarantee that (Rat, Rat) is the only Nash equilibrium?

Answer: (1.) The only requirement for (Rat, Rat) to be a Nash equilibrium is that $10 - r_2 \leq 10$. Thus, no rebates are required. (2.) The only requirement for (Rat, Rat) to be a strict Nash equilibrium is that $10 - r_2 < 10$. Thus, the prosecutor must give some minimal rebate when both rat, i.e. $r_2 > 0$. (3.) To guarantee that $(Clam, Clam)$ is not a Nash equilibrium, it is required that $10 - r_1 < 1$. That is, $r_1 > 9$. Then, in order for $(Clam, Rat)$ not to be an equilibrium (when $r_1 > 9$), it is required that $10 - r_2 < 10$, i.e. $r_2 > 0$.◀

Exercise 71. What rebates would you recommend the prosecutor to use?

Answer: It seems more reasonable to assume that the prisoners would rat if (Rat, Rat) would be a strict Nash equilibrium and if there is no other Nash equilibrium. Thus, the prosecutor should select some rebates such that $r_1 > 9$ and $r_2 > 0$. Setting r_1 high is costless for the prosecutor, if the prisoners actually behave according to game theory. It may then make sense to set this rebate as high as possible, i.e. $r_1 = 10$. Doing so should make it clear to the prisoners' that they have an incentive to rat if they believe that the other would clam. Thinking about r_2 is different. If the prisoners actually behave according to game theory, which we assume here, this is the rebate they will actually get. To maximize the punishment, the rebate should therefore be as small as possible. ◀

22.3. Game 2: Coordination game

Two cars meet on a road. If both drivers keep to the right, they will pass one another without any problems. Also if both drivers keep to the left, they will pass one another without any problems. If one drives on his right and the other on his left, they collide. The drivers only care about passing each other safely; the side does

not matter. The drivers must make their choices simultaneously without communicating with each other.

This traffic situation is clearly a situation of interdependent decision-making. Each driver's utility depends not only on what he himself does, but also on what the other driver does. In fact, each driver's best choice depends on which side of the road the other driver goes. So, lets specify a normal form game to represent this situation.

Recall that a normal form game has three ingredients, (i) the set of players, (ii) a set of strategies for each player and (iii) a utility function for each player. In the traffic situation, there are two players, driver 1 and driver 2. Each player has two strategies, left and right. We also know the physical outcome, which is summarized in Table 22.7 . But to analyze this situation, we need to come up

Table 22.7.: Traffic situation

	Left	Right
Left	no crash	crash
Right	crash	no crash

with some utility functions that represent the drivers' preferences. Since there are only two types of outcomes "crash" and "no crash" we only need to work with two levels of utility, say 0 (to represent a crash) and 0 (to represent that the cars do not crash). The normal form is summarized by the game-matrix in Table 22.8 .

Table 22.8.: Coordination game

	Left	Right
Left	1, 1	0, 0
Right	0, 0	1, 1

Following the procedure of underlining each players best response to every possible choice by the opponent, leads to the conclusion that this game has two Nash equilibria, namely $(left, left)$ and $(right, right)$. If both drivers believe that the other will keep to

the left, it is indeed rational for both drivers to keep to the left. If both drivers believe that the other will keep to the right, it is indeed rational for both drivers to keep to the right.

This game serves to illustrate the importance of coordination. It is not sufficient that the two drivers are rational. They need to have correct expectations of the other driver's behavior to avoid a crash. There is not even any conflict of interest in this situation. But it is still a non-trivial game, since both drivers need to adapt their behavior to what the other driver does.

In the real world, this particular coordination problem is solved by the convention to drive on a particular side of the road. In continental Europe and Scandinavia it is the right hand side; in the UK it is the left hand side. In any case, the convention helps the drivers to correctly anticipate others' behavior.

How do such conventions arise? Elementary game theory does not have an answer to that question. But it may for example be through learning. If for some reason more people drive on the left side, people will soon start to learn that they on average are involved in fewer crashes if they also drive on the left side. People will then start to switch from right to left. And with more and more people driving on the left side, it becomes easier and easier for the people still driving on the right to see that it is better for them to switch. In the end expectations are coordinated and equilibrium is reached.

But conventions are not the only solution. Another possibility might be to make driving on the left side illegal and to punish those who do. That, however, changes the payoffs of the game. If those who drive on the left hand side are always detected and punished severely enough, driving on the right may even become a dominating strategy. In Table 22.9 , it is assumed that anyone driving on the left side is punished and that their utility is reduced by $-p$ as a result.

Table 22.9.: Coordination game with outside punishment

	Left	Right
Left	$1-p$, $1-p$	$-p$, 0
Right	0, $-p$	1, 1

Exercise 72. How severe must the punishment be (i.e. how high must p be) in order for driving on the right to be a dominant strategy?

Answer: Driving on the right is a dominant strategy if it gives a higher payoff no matter which side the other car drives. In order for driving on the right to give a higher payoff than driving on the left, when the other car drives on the left, the punishment must satisfy $p \geq 1$.◀

Punishments for driving on the "wrong" side must probably be sever to make driving on the "right" side a dominant strategy. It would probably take years of imprisonment for people to prefer driving on the right side when an approaching car drives on the left.

Exercise 73. Please represent the following situation as a normal form game and summarize it using a game-matrix: Two Parisians want to meet for dinner. They can either meet in their favorite restaurant on the Left Bank of the river Seine or in their favorite restaurant on the Right Bank. Both restaurants are equally pleasant to both parties.

I guess that your game-matrix looked very much like the one in Table 22.8. Clearly, also this situation is a coordination game. And, clearly, there are two Nash equilibria also in this situation.

Exercise 74. Is there any reason to believe that the two Parisians will go to the same restaurant? (That is, is there any reason to believe that the two Parisians will behave according to the predictions of game theory?)

This time, the learning story seems less plausible. It is quite likely, however, that they have the time to discuss their choices before the game is played. We call this pre-play communication. They may then agree beforehand to both choose "left". Even if such an agreement is not legally binding, it is clear to both players that they will both have an incentive to follow the agreement. We say that such an agreement is self-enforcing.

Recall that, in contrast, an agreement between the two prisoners to clam in the prisoners' dilemma game would not be self-enforcing. Such an agreement would not change anything. The value of pre-play communication thus varies with the situation.

22.4. Game 3: Chicken

Two drivers head for a single lane bridge from opposite directions. The first to swerve away yields the bridge to the other. If neither player swerves, the result is a fatal head-on collision. If a crash is avoided, the driver who stays straight will be in time for his meeting while the other driver will come late.

Exercise 75. Please represent the situation above as a normal form game and summarize it using a game-matrix!

Answer: There are two players, driver 1 and driver 2. Each player has two strategies, swerve and straight. Finally we need to come up with some utility functions that represent the drivers' preferences. Since there are only three consequences for an individual player "coming in time" and "coming late" and "crashing" we only need to work with three levels of utility, say 2, 0 and -10. The normal form is summarized by Table 22.10 .◀

Now let's try to predict the outcome of this situation. Please answer the exercise before reading any further.

Exercise 76. Would a convention to always swerve be an equilibrium? Explain why or why not!

Table 22.10.: Chicken

	Swerve	Straight
Swerve	0, 0	0, 2
Straight	2, 0	−10, −10

Following the procedure of underlining each players best response to every possible choice by the opponent, leads to the conclusion that this game has two Nash equilibria, namely (swerve, straight) and (straight, swerve).

This game serves to illustrate additional problems of coordination. The first problem is that even if the two players are in exactly the same situation, they cannot behave the same way if they want to reach their final destinations quickly and avoid the crash. A convention like "always swerve" would not work as well as the convention "always keep to the right" in the game in the previous section.

The reason why a convention to always swerve would not be an equilibrium is that if both players expect the other to follow the convention, both players would have an incentive to do something else.

But if there is a difference in (say) age between the players and if the players can observe each others' age before making their choices there is a better chance of a workable convention. Please answer the following exercise before reading any further.

Exercise 77. Explain intuitively why the convention requiring the young to allow the elderly to pass first through a door may be thought of as an equilibrium.

The reason why this convention may be thought of as an equilibrium is that if two people of different ages meet at a door and if both of them expect the other to follow the convention, then both people will indeed have an incentive to do so.

Also pre-play communication would be much more difficult in the

Chicken game than in the Coordination game. An agreement that player 1 swerves would be self-enforcing. But player 1 may see no reason why he and not the rival should yield. But tossing a coin together may resolve the conflict. The parties may agree that player 1 swerves in case of head and that player 2 swerves in case of tail. The expected payoff to each driver would then be $0.5 \cdot 2 + 0.5 \cdot 0 = 1$. But, then again, the looser may wish to renegotiate.

22.5. Game 4: Stag hunt

Two hunters live close to a forest with only two animals. There is a stag (an adult deer) and a rabbit. If both hunters would bring their stag hunting gear, they would join forces and hunt the stag together. Success is guaranteed and they will both bring home 10 kilos of meet. But stag-hunting really requires cooperation - a single hunter has no chance to bring down the stag alone. Hunting the one rabbit in the forest is easy. But only one will be successful and the meal will be much less filling (2 kilos of meat). If both hunt hare, they have the same chance of spotting the hare first, meaning that they can expect to get 1 kilo of meat (i.e. 2 kilos with probability one half). Unfortunately the two hunters have to decide if they wish to bring their stag hunting or hare hunting gear before they meet.

Exercise 78. Please represent the situation above as a normal form game and summarize it using a game-matrix!

Answer: There are two players, hunter 1 and hunter 2. Each player has two strategies, to bring the stag hunting or hare hunting gear. Finally we need to come up with some utility functions that represent the hunters' preferences. In this case, the expected kilos of meat is probably a good representation. The normal form is summarized by Table 22.11 .◀

Now let's try to predict the outcome of this situation. Please

Table 22.11.: Stag hunt

	Stag	Hare
Stag	10, 10	0, 2
Hare	2, 0	1, 1

answer the exercise before reading any further.

Exercise 79. Find the Nash equilibria in the Stag hunt game!

Answer: Following the procedure of underlining each players best response to every possible choice by the opponent, leads to the conclusion that this game has two Nash equilibria, namely (stag, stag) and (hare, hare).◀

Please answer the following exercise before reading any further:

Exercise 80. Which Nash equilibrium do you think is the more reasonable prediction? Why?

Answer: There is no obvious answer to this question. On the one hand, both players are better off if the stag equilibrium is played than if the hare equilibrium is played. We say that the stag equilibrium **Pareto dominates** the hare equilibrium. Since both hunters are aware of the Pareto ranking, it may be natural to expect them to bring their stag hunting equipment. On the other hand, bringing the stag hunting equipment is **riskier**. If the other hunter goes for the hare, the cooperative hunter will have no dinner. Bringing the hare hunting equipment may feel more safe. Then, each hunter can expect a dinner independent of what the other does. ◀

The hunters may also pick up the phone and agree on stag hunting, removing the fear that the other party may have any misunderstood the situation. The agreement would be self-enforcing, assuming both hunters to trust their partner's promise. But, trust is actually not guaranteed in this situation.

Exercise 81. Can you see why?

Answer: A hunter who intends to go for the stag has an incentive to reveal his intentions. But also a hunter who intends to go for the hare has an incentive to convince the other hunter to go for the stag, thereby increasing his own chance of successfully shooting the hare. So, in this game, if you don't trust your partner beforehand, pre-play communication is actually not informative about the players' intentions.◀

22.6. Game 5: Football penalty game

A penalty shot in football (soccer) is a one-on-one game, the shooter against the goalkeeper. The shooter must decide to shoot either left or right. The goalkeeper must decide which side to defend. The ball only takes 0.2 - 0.3 seconds from the kick to reach the net, so the goalie must make up his mind already before the kick; otherwise it would be too late.

Exercise 82. Please represent the situation above as a normal form game and summarize it using a game-matrix!

Answer: There are two players, the shooter and goalie. The shooter has two strategies, to go left and to go right. The goalie has two strategies, to defend the left side or to defend the right side. Finally we need to come up with some utility functions that represent the players' preferences. In this case, there are two possible outcomes, namely goal and no goal. The shooter prefers goal to no goal, which may be represented by +1 and -1. The goalie has the opposite preferences, which may be represented by -1 and +1. Letting the shooter be player 1, the normal form is summarized by Table 22.12 .◀

Now let's try to predict the outcome of this situation. Please answer the exercise before reading any further.

Exercise 83. Find the Nash equilibria of the football penalty game!

Table 22.12.: Football penalty game

	Left	Right
Left	$-1, 1$	$1, -1$
Right	$1, -1$	$-1, 1$

Answer: Following the procedure of underlining each players best response to every possible choice by the opponent, suggests that this game has no Nash equilibria. If the players would have correct expectations about the outcome of this game, one player would have an incentive to switch strategy. If both players expect both to go left, the shooter would actually have an incentive to deviate and shot right. If both players expect the shooter and the goalie to go different directions, they both expect a goal, and the goalie has an incentive to deviate.◀

So what is the key difference between this game and the previous games that we have studied? The key characteristic of this game is that it portrays an extremely competitive situation. One player's gain is the other player's loss. It is a zero-sum game.

In a zero-sum game, players don't want to be predictable. If the shooter is predictable, the goalie wins, and vice versa. This means that there cannot be a common understanding of how people behave in zero-sum games. If it would be commonly known that goalies always defend their right, the shooters would always go left, and the goalies would actually not do what they are known to do - a contradiction.

The players may instead consider behaving in a random way to avoid being predictable.

22.6.1. Tossing coins to be unpredictable

Let's say that we allow each player an additional strategy, namely to *secretly* toss a coin before the kick. Head means left; tail means

right. Each player then has three strategies: "go left", "go right" and "let the coin decide".

What are the payoffs if a player lets the coin decide? If the goalkeeper lets the coin decide, the shooter will score a goal 50 percent of the time, independent of the whether the shooter goes left, right or lets the coin decide. The *expected* payoff to both players is thus zero if the goalkeeper lets the coin decide. Similarly, the expected payoff to both players is zero also if the shooter lets the coin decide.

Exercise 84. Find the Nash equilibria of the football penalty game when both players are allowed to "toss a coin"!

Answer: To start out, we represent the new situation with the normal form summarized by the bi-matrix in Table 22.13 . The

Table 22.13.: Football penalty game with coin tossing

	Left	Coin	Right
Left	$-1, \underline{1}$	$\underline{0}, 0$	$\underline{1}, -1$
Coin	$0, \underline{0}$	$\underline{0}, \underline{0}$	$0, \underline{0}$
Right	$\underline{1}, -1$	$\underline{0}, 0$	$-1, \underline{1}$

players' best reply functions are also indicated in the matrix. Note in particular that if one player tosses the coin, then it is a best reply for the other player to also toss the coin. Tossing the coin is not a unique best reply, but it is a best reply. In fact, there exists a unique equilibrium in the game, requiring both players to toss the coin, i.e. to behave in a random way.◀

But, surely, you may object, sensible people don't toss coins when making important decisions. (Scoring goals is the business of professional athletes and their incorporated teams.) Oddly enough, it seems that they do. Or, at least, it appears that people sometimes behave *as if* they toss coins. Empirical studies of athletes in high stake games such as tennis players' serves and football players'

penalty kicks suggest that they do use mixed strategies and that their behavior is close to equilibrium predictions.

22.6.2. Existence of equilibrium

In the football penalty game an equilibrium exists, at least if we allow the players to toss a coin. This is, an example of a much more general result.

To take full advantage of the idea of randomization, we cannot stick to coin-tossing. Tossing coins requires the players to choose the two strategies with equal probabilities. This worked well in the football penalty game that we just looked at. But 50:50 probabilities would not work in all games. We should allow the players to choose any probability distribution they want, e.g. to go right with a probability of ten or perhaps thirty-two percent.

To illustrate why it is important to let the players choose their probabilities freely, consider the game in Table 22.14 . Player 1

Table 22.14.: Generalized football penalty game

	Left	Right
Up	A, $\underline{a}$	$\underline{C}$, c
Down	$\underline{B}$, b	D, $\underline{d}$

can choose between two *pure strategies*, Up and Down. Player 2 can also choose between two pure strategies, Left and Right. These two strategies are said to be pure to emphasize that they do not involve any randomization.

Player 1's payoffs are indicated by capital letters and it is assumed that $B > A$ and $C > D$. That is, player 1 prefers to go Down if player 2 goes Left and Up otherwise. Player 2's pay-offs are indicated by lower-case letters and it is assumed that $a > c$ and $d > b$. That is, player 2 prefers to go Left if player 1 goes Up and Right otherwise. If $B = C = a = d = 1$ and $A = D = c = b = -1$ we would obtain the same football penalty game as above.

Independent of the exact values of the payoffs, it is easy to see that the game has no equilibrium in pure strategies. For instance, if both players would believe that Down and Left will be chosen, player 2 would actually have an incentive to deviate, refuting their theory.

22.6.2.1. Games in mixed strategies

Now let's transform the original game into a *game in mixed strategies*. To do so, we need to change the original game in two ways.

First, we need to redefine the players' strategy sets to allow them to randomize their behavior. Let's say player 2 chooses Left with probability p and Right with the remaining probability $1-p$. Player 2 is allowed to chose whatever probability p he prefers, as long as it is a real number between zero and one. That is, the interval between zero and one is player 2's strategy set. Of course, $p = 1$ corresponds to the pure strategy Left and $p = 0$ corresponds to the pure strategy Right. All other probabilities imply that player 1 chooses both Left and Right with positive probability. Similarly, player 1 chooses Up with probability q and Down with probability $1 - q$. That is, also player 1's strategy set is the interval between zero and one.

Second, we need to redefine a player's payoff in terms of his *expected utility*. Given that player 1 goes Up with probability q and that player 2 goes Left with probability p, player 1 will get a payoff of A with probability $p \cdot q$. He will get a payoff of C with probability $(1 - p) \cdot q$ and so on. That is, player 1's payoff function is given by

$$EU_1(p, q) = [A \cdot p + C \cdot (1 - p)] \cdot q + [B \cdot p + D \cdot (1 - p)] \cdot (1 - q),$$

which associates an expected utility to all possible strategy profiles (p, q) that the players might choose.

Now, we have a well-defined game in normal form. We know the

two players, the shooter and the goalie. We know that the shooters strategy set is to choose some probability between zero and one and that the goalies strategy set is to choose some probability between zero and one. We know that the players' payoff functions are their expected utilities when they play their chosen strategies.

22.6.2.2. Best reply functions

To find out how the players will play, we start out finding their best reply functions. For player 1 we need to know what probability q that maximizes his expected payoff, for every probability p that player 2 might choose. And vice versa.

Player 1 has to alternatives, Up and Down. Clearly, player 1's expected utility of going Up and Down depends on what player 2 is doing. If player 2 is going Left with high probability, then Down is better than Up. Then, player 1 should choose Down with certainty. There is no reason to choose Up with any positive probability. Similarly, if player 2 is going Right with high probability, then Up is better than Down. Then player 1 should choose Up with certainty. Only if Up and Down would give player 1 the same expected utility would it be reasonable for player 1 to choose both these alternatives with positive probability. Now, let's make this more precise.

To investigate how player 1's expected utility depends on his strategy q, it is convenient to define player 1's expected utility of choosing Up and Down with certainty. They can be computed as:

$$EU_1(Up) = EU_1(p, 1) = A \cdot p + C \cdot (1 - p)$$

and

$$EU_1(Down) = EU_1(p, 0) = B \cdot p + D \cdot (1 - p)$$

Using these expressions, we can now rewrite expected utility as

$$EU_1\left(p, q\right) = EU_1(Up) \cdot q + EU_1(Down) \cdot (1 - q).$$

Notice that player 1's expected utility is strictly increasing in q if $EU_1(Up) > EU_1(Down)$ and strictly decreasing if the inequality is reversed. If $EU_1(Up) = EU_1(Down)$ all q give the same expected utility.

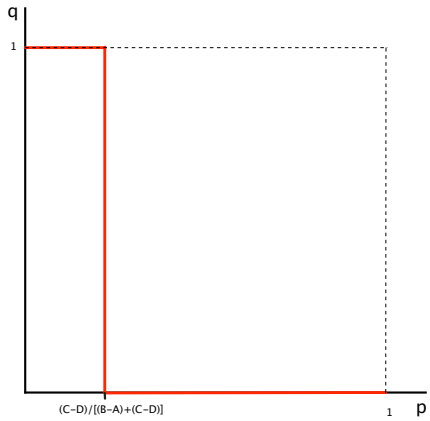
This means that player 1 strictly prefers to go Up with certainty, i.e. probability one ($q = 1$), over all other mixed strategies if $EU_1(Up) > EU_1(Down)$. Expressed differently, $q = 1$ is the unique best reply if $EU_1(Up) > EU_1(Down)$. Recall that $(B - A) > 0$ and $(C - D) > 0$ so that we can rewrite the condition as

$$p < p^* \equiv \frac{(C - D)}{(B - A) + (C - D)}.$$

Similarly, player 1 strictly prefers to go Down with certainty ($q = 0$), over all other mixed strategies if $p > p^*$. Finally, player 1 is indifferent between all strategies q if $p = p^*$. In this case, all probabilities q are best replies for player 1.

We can now display player 1's best reply function in a Figure 22.1. The thick red line shows player 1's optimal choice of q for all

Figure 22.1.: Player 1's best-reply function



possible values of p. If $p < p^*$, player 1's best reply is $q = 1$. If $p > p^*$, player 1's best reply is $q = 0$. If $p = p^*$, every q between

zero and one is a best reply.

Exercise 85. Find the player 2's best-reply function!

Answer: Use the same procedure as for player 1, defining $q^* = \frac{(d-b)}{(a-c)+(d-b)}$. ◀

Player 2's best-reply function is also displayed in Figure 22.2.

22.6.2.3. Equilibrium

The Nash equilibrium is where both players are on their best replies, i.e. the intersection of the red and the green curves in Figure 22.2 . Notice that this equilibrium requires both players to randomize between their two choices. Also note that the equilibrium probabilities depend on the payoffs, A, B, C, D as well as a, b, c and d.

22.6.2.4. Existence of equilibrium

A fundamental result in game theory, due to John Nash, is that any game with (i) a finite number of players and (ii) where each player has a finite number of strategies has at least one Nash equilibrium. Sometimes, as in the football penalty game, there doesn't exist an equilibrium in pure strategies. But then there is always at least one equilibrium in mixed strategies.

22.6.3. A rationale for mixed strategy equilibrium (Difficult)

Many people are skeptical about mixed strategy equilibria. It appears unlikely that mixed equilibria are good predictions of how people behave in the real world. Thus, a substantial part of economic analysis is put into question. One problem with mixed strategy equilibria is that they seem to be bad descriptions of real-world decision making processes. "Players" including governments, firms

and regular people acting on their own behalf don't usually throw dice when making important decisions. And why should they? Why would such players follow the particular probability distribution, prescribed by game theory, when there are several pure strategies that give the same expected payoffs? People may rather be inclined to just choose one or the other pure strategy, without much thought. And there is no obvious reason why an individual's tendency to select one equally good strategy over the other would coincide with the probabilities prescribed by game theory. The answer that game theory would have us believe is that players must follow their mixed strategies, not because it is necessary for maximizing their own utilities, but because it is necessary for making the other players indifferent between the pure strategies they are supposed to use with positive probability. That is clearly not a very convincing argument.

Fortunately, there are better arguments. But then we need to go a bit outside of the narrow game we are studying. Remember that the formal games we analyze are intended to be simplified descriptions of the world, removing some of the complexity that makes real-world situations so difficult to understand. For instance, the payoffs we attach to different outcomes are only meant to describe the "typical" payoff associated with that outcome. In the real world people differ; some people get a slightly higher and some get a slightly lower payoff than the "typical" payoff. Thus, put in a situation where all people would consider two alternatives almost equally good, some of them will actually slightly prefer one alternative and others will slightly prefer the other.

The big question is then if there is any reason to believe that mixed strategies can be interpreted as the *frequency distribution of choices in the population*? The so-called purification theorem demonstrates that the answer is "yes." To illustrate the purification theorem, we will consider the so-called matching pennies game (a.k.a the football penalty game).

Matching Pennies The matching pennies game is described by the payoff matrix in Table 22.15. This game has two players, called 1 and 2. Both players have two strategies, called H (head) and T (tails). Player 1 wins if the players choose the same strategies, i.e. H and H or T and T, and player 2 wins otherwise. Both players get payoff 1 when winning and -1 when losing.

Table 22.15.: Matching Pennies

	H	T
H	1, -1	-1,1
T	-1,1	1,-1

This game can for example be viewed as a stylized representation of a penalty kick in football. The goalie (player 1) wins if he defends the same side that the shooter kicks.

In this game there is no pure strategy equilibrium. There is however a (unique) mixed strategy equilibrium, requiring both players to select their two strategies with equal probabilities. But, again, why would they?

Matching Pennies with incomplete information Remember that Table 22.15 is a simplified description of the situation. Different people acting in the *role* of player 1 have slightly different payoffs. It may for example be that they get some extra utility or disutility from selecting H. Let us say they get $\varepsilon \cdot x_1$. The first factor, $\varepsilon < 1$, which is a small positive number, is the same for all people who act in role 1. It is also known by all people, including those who act in role 2. The second factor, x_1, differ between different individuals. And, while every person acting in the role of player 1 knows his own type (his own value x_1) the opponent does not. We say that there is incomplete information. Let us for simplify assume that x_1 is uniformly distributed on $[-1,1]$.[3] Similarly, every person acting

[3]This means that the probabiltiy that $x_1 \geq y$ is given by $\frac{1-y}{1-(-1)} = \frac{1}{2} - \frac{1}{2} \cdot y$.

in the role of player 2 gets the extra payoff $\varepsilon \cdot x_2$ when selecting H. These payoffs are illustrated in Table 22.16.

Table 22.16.: Matching Pennies with incomplete information

	H	T
H	$1+\varepsilon \cdot x_1$, $-1+\varepsilon \cdot x_2$	$-1+\varepsilon \cdot x_1$,1
T	-1,$1+\varepsilon \cdot x_2$	1,-1

Assume now that every person acting on one of the two roles will choose one of the two actions with certainty. Nobody uses a mixed strategy. Still, however, different people acting in the role of player 1 may prefer different actions, depending on their types (i.e. depending on their x_1). Similarly, different people acting in the role of player 2 will make different decisions, depending on their types. To describe these different choices we let p_1 denote the share of people acting as player 1 who select H. Similarly, p_2 denotes the share of people acting as player 2 who select H.

Since a person in the role of player 1 does not know what type his opponent is, the opponent's choice will appear as a random variable, with the probability of H being p_2. Then, the expected utility of choosing H, for a person in the role of player 1, characterized by x_1, is given by

$$EU_1(H) = (1+\varepsilon \cdot x_1) \cdot p_2 + (-1+\varepsilon \cdot x_1) \cdot (1-p_2).$$

Similarly, the expected utility of choosing T, is given by

$$EU_1(T) = (-1) \cdot p_2 + (1) \cdot (1-p_2).$$

Thus, a person acting in the role of player 1 will select H if and only if

$$x_1 \geq \underline{x}_1 = 2 \cdot \frac{1-2 \cdot p_2}{\varepsilon}.$$

Since x_1 is uniformly distributed over $[-1,1]$, the probability that a person in the role of player 1 selects H is given by $p_1 = \frac{1-\underline{x}_1}{1-(-1)}$.

Thus, substituting for $\underline{x}_1$, we can compute the population share of players 1 selecting H as a function of the population share of players 2 selecting H as:

$$p_1 = \frac{1}{2} - \frac{1 - 2 \cdot p_2}{\varepsilon}.$$

Exercise 86. Show that the probability that a person in the role of player 2 selects H is given by

$$p_2 = \frac{1}{2} + \frac{1 - 2 \cdot p_1}{\varepsilon}.$$

We are thus left with a system of two linear equations in two unknowns and it is easy to see that the solution is $p_1 = p_2 = \frac{1}{2}$. The point is, of course, that this is the same distribution as prescribed by the mixed strategy equilibrium of the original game.

Interpretations Notice that people acting in the role of player 1 need to be informed about p_2 to determine their own best choice. There are two alternative assumptions to motivate why they may possess this knowledge. First, they may observe that half of players 2 have selected H historically and therefore assume that their current opponent will select H with these probabilities. An alternative interpretation is that player i is informed about the distribution of the unobservable x_j in the population and that player i therefore can compute the equilibrium value of p_j.

An alternative version of the story told here is that we are studying the same two people playing the game over and over, but that their utilities vary slightly over time. The formal analysis would be the same.

Purification theorem We started out with a normal form game with a Nash equilibrium in mixed strategies. We constructed a new "perturbed" game of incomplete information that is very similar to the original game. The perturbed game has the same set of players

and the same set of strategies. But, the players' payoffs are stochastic and vary slightly around their means. We assumed that the players only know their own payoffs. We demonstrated that there is a Bayesian-Nash pure-strategy equilibrium of the incomplete information game that implies the same probability distribution over outcomes as the mixed strategy equilibrium of the original game.

The purification theorem by Harsanyi[4] states that the same procedure can be used for any mixed strategy Nash equilibrium for almost all (finite) normal form games. [5] It follows that we may interpret the mixed strategy equilibrium of a game as an approximation of the population distribution of strategies in a slightly more complicated game with incomplete information.

22.7. Basic idea of an equilibrium

An equilibrium can be described as a *common understanding* of how people will behave in a certain situation. Sometimes there is such an understanding. An example is that we all know that all drivers in Sweden keep to the right. It is important to note that in order for such a common understanding to arise, every person must have an incentive to indeed do what is expected from him. Otherwise it would be foolish to expect him to do it.

It could not be an equilibrium that 90% of all drivers keep to the right and 10% keep to the left, since then the people who are presumed to drive on the left side would be better off switching lane. They would be involved in fewer accidents.

That there is a common understanding of how people behave

[4]Harsanyi, J. 1973: Games with randomly disturbed payoffs: a new rationale for mixed-strategy equilibium points. *International Journal of Game Theory* 2(1): 1-23.

[5]The fact that the probability distribution over outcomes is exactly the same is not typical. In most games, the equilibrium distribution of the incomplete game will depend on ε and be slightly different from the mixed strategy equilibrium of the original game. But, as $\varepsilon \to 0$ the two distributions will converge.

does not imply that we all behave the same. There is room for differences. In England it is commonly understood that everyone keeps to the left. Both "everyone keeps to the right" and "everyone keeps to the left" can be an equilibrium – a common understanding of how people behave. In other words, there are *multiple equilibria* in certain situations. Still, there is only room for one equilibrium to prevail at any particular time and place. Within a community everyone will keep either to the right or to the left. In driving, there is only room for variation across different communities.

An equilibrium can also be described as a self-fulfilling prophecy. Given that everyone expects a certain outcome, everyone will indeed behave accordingly and, in the end, be correct in their expectations. For example, everyone in Sweden expects everyone else to drive on the right side, therefore everyone has an incentive to do so and, in the end, our presumptions about other peoples' behavior is proved correct.

In other types of situations, different people must behave differently also within the same community. Since two people cannot walk through the same door at the same time, one has to yield to the other. It may then be commonly understood that men let women pass first, or that the young give way for the old.

In both these examples the fact that there is a common understanding of how people behave help them avoid crashes or long periods of waiting in the door. A common understanding may thus improve the welfare of all parties. A common understanding may help produce a Pareto efficient outcome. In other situations a common understanding will not help people improve on the situation. Think about fishing. All countries surrounding a lake understand that all countries will catch all the fish that happen to enter that country's territorial water. Therefore you either catch the fish or let someone else have it. No one has an incentive to save some of the fish for next year. There may even be a common understanding that the fish will be extinct in the end. A joint declaration to

reduce fishing would not work unless it is combined with measures to ensure that all parties honor the agreement. But that is a very useful insight in itself: "You can't trust Melanie but you can trust Melanie to be Melanie."

What this means is that:

Conclusion 4. If we have reasons to believe that *the people involved in some situation have a common understanding of how they will all behave*, then we can predict their behavior.

The reason is that certain types of behavior are not viable as expected behaviors. If they would be expected, people would not have an incentive to actually do what is expected from them. Thus, that behavior could not be commonly understood.

Game theory has been developed in an attempt to define exactly what we can know about behavior in situations where a common understanding is likely to arise. The examples given here may appear trivial but we will soon discuss more useful example, once we have learned to understand the basic ideas in the context of well-known examples.

But, is there any reason to believe that a common understanding of how everyone will behave should arise? In many situations there is. In traffic, anyone belonging to the minority, driving on the "wrong side," will quickly *learn* that they can avoid more crashes by switching lane. The minority will thus become smaller, increasing the gains even further for the rest of the minority to also switch side. (Or they don't learn, in which case the minority also becomes smaller.) *Traffic lights* help drivers coordinate in crossings. Cars in different direction often take turns, using a commonly observed *history* of events help coordinate behavior. And if a Swede and an Englishman happen to meet on a road in a land without conventions, they might have the possibility to talk to each other and agree to both drive on the right side, knowing that they will both have an incentive to honor the agreement. That is, the *agreement*

is self-enforcing, without the help of any police or court.

Some situations, however, are such that we cannot expect a common understanding of how everyone will behave to arise, at least not if we interpret behavior in a narrow sense. It could not be commonly known that goalkeepers in football (soccer) always defend their right, since then the outfield players would always kick the ball on the left side, always scoring a goal. The goalies would then have an incentive to change their behavior. For similar reasons, it could not be commonly known that penalty kickers always kick the ball on the right side. A football match is a situation of pure conflict: one team's gain is the other team's loss. In such a zero-sum situation, it will be in no one's interest to be predictable. Both the goalie and the kicker will do their best not be predictable. They will vary their behavior. This does not, of course, mean that they will simply alternate, going left and right every second time. The opponent would soon detect such a pattern and exploit it to his advantage. The players will rather try to be as irregular as they possibly can. Which is to say, they will do their best to replicate the choice of a dice. But then, in the end, a common understanding may arise, albeit of a slightly different kind, namely that both the goalie and the kicker will behave as if their choices where made by rolling a dice.

A social scientist who wishes to predict how people behave in a certain situation can thus use the idea of an equilibrium. If the social scientist is convinced that the situation is such that a common understanding will somehow arise, then he can eliminate certain outcomes as implausible. Without having visited all countries around the world, the social scientist can safely predict that there doesn't exist any country where 90% of the drivers keep to the right and the rest keep to the left.

In this course we will use the same theory of games but to make predictions about how people behave in markets. One of the first predictions we will derive is that firms set lower prices the more

competitors they have.

22.8. Summary

- Game theory is a tool for predicting how people behave in situations where the outcome depends on several peoples' behavior and where each person's preferred behavior depends on what the others do. Game theory may also be called the theory of interdependent decision-making.
- A game in normal form specifies (i) who the players are, (ii) the players' available strategies and (iii) the players' pay-offs for every possible strategy profile (i.e. combination of strategies chosen by the players).
- A game is intended to be a highly stylized representation of some real world situation, focusing on the essential features of the situation, but also disregarding many features that are believed to be less important for understanding people's behavior.
- A Nash equilibrium is a strategy profile, i.e. one strategy for each player, such that every players' strategy is the optimal choice for that player, given that all other players follow their strategies.
- A Nash equilibrium requires people to maximize their utility given their beliefs about what others will do (rationality) and to have correct beliefs about the other players behavior (coordination).
- Coordination may be achieved in many different ways, e.g. through learning, social conventions, pre-play communication or by being suggested by some authority (e.g. the law).

- In any finite game, there exists at least one Nash equilibrium. The equilibrium may require the players choose their strategy randomly. In some situations such random behavior may be interpreted as a wish to not be predictable.
- A commonly used procedure to identify the Nash equilibrium starts out by computing every player's best-reply function. The best-reply function describes the player's best strategy for every strategy that the other player(s) could choose.

22.9. Additional Exercises

22.9.1. Battle of the sexes

The morning newspaper lists two Friday night events: a football match at the stadium and a play at the theater. The Svenssons is a conventional couple. Mr. Svensson prefers the match while his wife prefers the play. But both spouses also enjoy each other's company very much: "shared joy is double joy," they often say. In fact, both prefer being together with their spouse at their least preferred event to go alone to their favorite event. Unfortunately the two spouses where unable to plan their evening activities before leaving home in the morning. And when the evening comes, one of their mobile phones is out of battery. They must make up their minds independent of one another. Since the stadium is located in the suburbs and the theater is in the city center, they will not have the time to go to more than one place.

1. What three objects make up a game in normal form?
2. Represent the situation above as a game in normal form! Write down the game matrix!
3. What is a best reply? What is a best reply function?

4. Derive the best reply functions of this game!
5. What is a Nash equilibrium?
6. Compute the Nash equilibria of this game! Does an equilibrium exist? Are there multiple equilibria?
7. What is the conclusion of the game theoretic analysis of this situation?
8. Are the equilibria reasonable predictions of how people would behave in a situation as this? What could possibly coordinate the two spouses?
9. What is a mixed strategy? Define the set of mixed strategies for both players.
10. How is the payoff functions defined in a game with mixed strategies? Write down the two payoff functions.
11. Derive the best reply functions in the game with mixed strategies! Display them in a figure with Mr. Svensson's strategy set on the x-axis and Mrs. Svensson's strategy set on the y-axis!
12. Compute the Nash equilibria in mixed strategies of this game!

22.9.2. Prisoner's dilemma, revisited

Consider the prisoners' dilemma above, but assume that both prisoners care about their partners' sentence. For example, prisoner i's payoff may be given by $u_i = 20 - y_i - \alpha \cdot y_j$ where y_i is the number of months prisoner i has to serve.

1. How altruistic must the prisoners be in order for an agreement to clam to be self-enforcing? (I.e. how large must α be?)

2. How altruistic must the prisoners be in order for clam to be a dominant strategy?

3. But would it not be irrational for a player to spend more time in jail than necessary? Can a rational person care about others?

22.9.3. Football penalties, revisited

Consider the generalized football penalty game in Table 22.14 again.

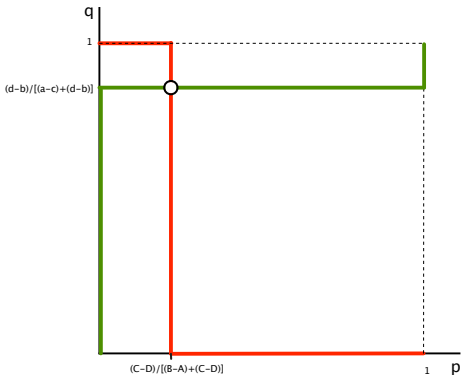
1. What is $u_1(up, left)$?

2. Transform player 1's payoff to become $v_1 = a + b \cdot u_1$ where u_i is the payoff of the original game and b is a strictly positive number! (This means that $v_1(up, left) = a + b \cdot u_1(up, left)$ and so on.)

3. Find the Nash equilibria in the transformed game!

4. How is predicted behavior changed by the transformation of the payoff function?

Figure 22.2.: Equilibrium



23. Dynamic Games

The time dimension is crucial for understanding many important issues in economics. Firms sometimes lower their prices below their competitors' costs (sometimes even below their own costs) to force weaker rivals to exit the market. If such *predatory pricing* is successful, competition is reduced and the predator can subsequently increase its prices to a much higher level. Another example is *entry deterrence*. Incumbent firms can undertake various activities to discourage new firms from entering the market. Some incumbents may for example invest in low cost technologies to be able to set very low prices in case of entry. Also for *cartel agreements* the time dimension is important. A cartel agreement is an agreement between firms not to compete with one another. The firms may for instance agree to all charge the monopoly price and to share the market equally. Such agreements are "unstable." The reason is that all firms have an incentive to undercut their rivals' prices and to sell more than their agreed share of the market. Therefore a cartel agreement will only work if the firms are able and willing to punish any rivals who cheat on the agreements. A final example is *price negotiations* between buyers and sellers in intermediate goods markets. An important determinant of the price is then how patient and willing to continue haggling the parties are. The more patient you are, the better deal you get.

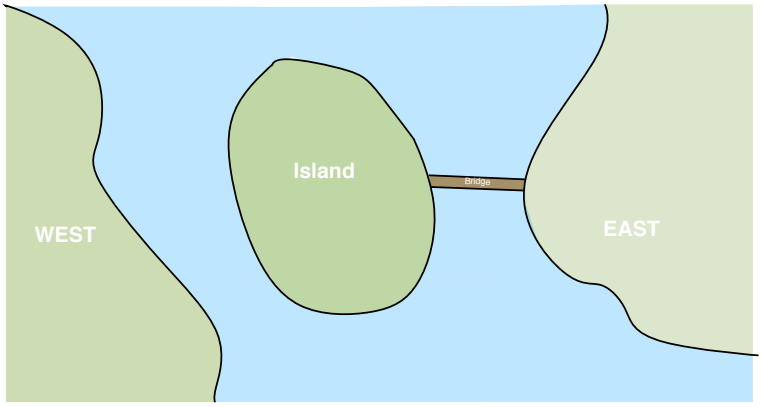
Before we study predation, cartels and bargaining we need to acquaint ourselves with the necessary game theoretic tools, however. To study situations with an important time dimension, it is better to use so-called games in "extensive form" rather than games in

"normal form." To introduce extensive form games we will consider a toy game called War & Peace. This game also illustrates two key ideas of dynamic analysis. The first idea is *credibility*. If you wish to influence somebody else's behavior e.g. by promising to reward kindness or to retaliate against wrong-doers, then you need to have an incentive to actually provide the reward or to carry out the threat. Threats and promises must be credible. Otherwise your opponents will just ignore them. The second idea is *commitment*. Sometimes you can influence your own future behavior. By investing in a low cost technology today, you commit to charge a low price and produce more tomorrow.

23.1. Game 1: War & Peace

Consider a territorial conflict over a fertile island located in the channel between two countries, called East and West. East is currently holding the island and has built the Bridge of National Unity between the island and the mainland. West claims the island his-

Figure 23.1.: War & Peace



torically belongs to them and its navy is standing ready to cross the channel to land the National Libration Army. Once the army has disembarked at the island, the ships would have to return to base

to protect the mainland from other potential intruders. To deter an invasion, East has posted an army on the island. If there is a war, both countries have a 50% chance of winning and the victor's value of holding the island is V. The cost of warfare is $C > V/2$ to both parties. The cost is thus very high compared to the value of the island. East's army could also choose not to fight an invasion by retreating over the bridge back to mainland East. Then, there is no war and consequently no cost of war. But then West takes the island with certainty.

23.1.1. Naive analysis

To study the conflict, we can write down a normal form game. In this game, there are two players, East and West. West has two strategies, to Attack or Not Attack. Also East has two strategies, to Defend or Not Defend. The payoffs are summarized in the game matrix.

Table 23.1.: Naive game

	Defend	Not defend
Attack	$\frac{1}{2} \cdot V - C, \frac{1}{2} \cdot V - C$	$\underline{V}, \underline{0}$
Not attack	$\underline{0}, \underline{V}$	$0, \underline{V}$

The best-reply functions are indicated by underlining the appropriate payoff. Let us start with West: If East will defend the island it is better for West not to attack. If East would not defend, West would prefer to attack. Now consider East's best- reply function: In case of an attack, it would be better for East not to defend the island. In case of no attack, East's choice is irrelevant. This best-reply analysis reveals that there are two Nash equilibria: one equilibrium prescribes West to attack and East not to defend; the other equilibrium prescribes East to defend the island and West to refrain from attacking.

This analysis suggests that East's threat to defend the island might be successful; at least this is one of two possible Nash equilibrium outcomes. West is scared off and does not invade. And East does not even have to fight the war. It seems that it is the mere threat to defend the island that does the job.

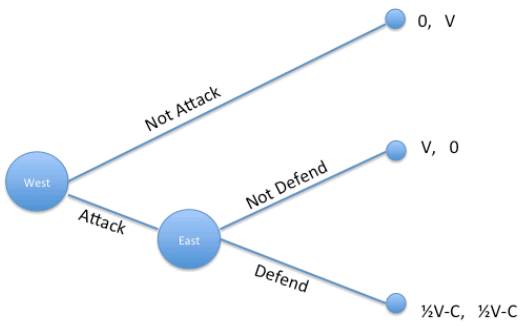
There is, however, a severe problem with this normal form game and the Nash equilibrium analysis. The time-structure of the real situation is not properly reflected in the analysis. West's choice to attack by landing the National Liberation Army on the island comes *before* East's choice to defend the island or to retreat over the bridge. Assume that West for some reason invades the island. Then, it is actually better for East not to defend the island but to retreat over the bridge back to the mainland. East does not have an incentive to really defend the island in case of an invasion. That is to say, East's threat to defend the island is not credible. Naturally, West's generals understand the situation and will therefore invade the island and take it over. When the time-structure of the game is taken into account, it is clear that one of the two Nash equilibria of the normal form game above is not a reasonable prediction of how the two countries will actually behave. Not all Nash equilibria are equally appealing as predictions of how people will behave.

Fortunately, it is possible to make the idea of credibility part of the game theoretic analysis. It is possible to refine what we mean by an equilibrium, to only encompass the more reasonable prediction. As a first step, we need to enrich the description of the situation. It is not sufficient to only name the players, list their strategies and specify their payoffs. We have to somehow also include the time-structure "Who moves before the other?" in the description of the game. Using a so-called *extensive form game* is often a more convenient format for analyzing interdependent decision-making taking place over time.

23.1.2. Extensive Form Game: Backwards Induction

An extensive form game differs from a normal form game, primarily by explicitly specifying the *timing of decisions*. The extensive form can be described by a game tree. A game tree is a decision tree, but with several decision makers. The War & Peace Game is summarized in Figure 23.2. This game tree has two decision-nodes.

Figure 23.2.: War & Peace - The extensive form



The first decision-node belongs to West; West can choose between two available actions, to attack or to not attack. The two branches leading out from the node indicate the two actions that West can take. The second decision-node belongs to East; East can choose between two available actions, to defend or to not defend. The two branches indicate the two actions East can take. There are also three end-nodes, indicating the outcome of the game and, in particular, the expected payoffs to both players. West's payoff is displayed first.

The extensive form specifies the same information as a normal

form game (the set of players, the actions that they can take, and their payoff functions) and in addition the timing of the decisions.

Note that a player's different choices at a particular node are called actions, rather than strategies. In extensive form games, a *strategy* is a complete plan of action, i.e. a rule prescribing an action to every decision-node that a player has. Since in this simple game, both players only have one decision-node each, strategies and actions amount to the same thing.

The War & Peace Game is an example of a game with *perfect information*. A game is said to have perfect information if two conditions are satisfied, namely that (i) there are no simultaneous moves, and (ii) at each node the player knows which choices have been made at earlier nodes.

The crucial trick when analyzing an extensive form game of perfect information is to start the analysis with the last decision-nodes, to find out what the best actions are at these nodes, then to consider the second to last decision nodes, to find out what the best actions are at these nodes, and then to continue to work backwards until the first node of the game tree (the so-called root) is reached. This procedure is called *backwards induction* (or dynamic programming).

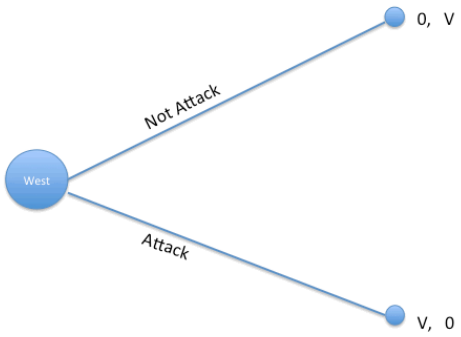
In the War & Peace Game, we consequently start with East's choice to defend or not defend. Note that the game tree makes it clear that East's choice follows after a (still hypothetical) attack by West. Anyway, after an attack, should it occur, East can choose between retreating, which gives zero payoff, or defending the island, which gives a lower payoff $\frac{1}{2} \cdot V - C < 0$. The prediction is clear: East would retreat, should an attack occur.

Working backwards, the time has now come to analyze West's decision to attack or not. West's choice is not trivial. It depends on what West believes East would do in case of an attack. West's best choice is to attack, if West believes that East would retreat; otherwise West's best choice would be to not attack. But West has

full information about the game, including East's available actions and East's payoff function. West can therefore predict that East's rational reaction to an attack is to retreat. The prediction is clear: Knowing that East will retreat in case of an attack, West will attack.

This analysis is simplified by truncating the game tree. Since both we (as outside observers) and West can predict East's decision, we can replace East's decision-node with the relevant end-node. Again,

Figure 23.3.: War & Peace - Truncated game



correctly predicting that East will retreat, West will clearly attack.

By (i) specifying the extensive form game and (ii) using backwards induction we have produced a unique prediction: Attack and Not defend. Note that this solution is also one of the two Nash equilibria. By using backwards induction, we have thus *selected* the more reasonable Nash equilibrium. We have selected the equilibrium not involving any empty threats.

23.1.3. Commitment and Credible Threats

Assume now that East's generals reconsider their strategic situation, already before West has had the opportunity to attack. They

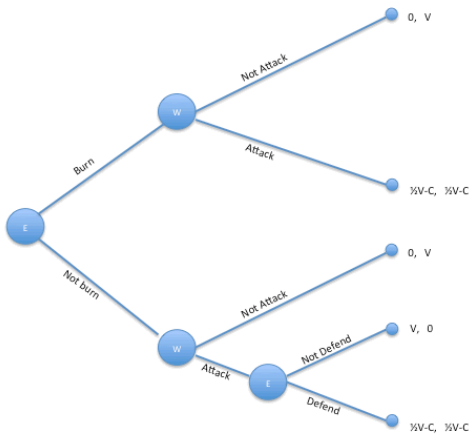
realize that just posting an army on the island will not deter an invasion by West. Also incentives matter.

One of the generals suggests that they might burn the bridge between the island and the mainland, to hinder the army from retreating. Another general objects that without the bridge, West's attack will lead to war, the worst possible outcome.

Exercise 87. What should East do? What would you recommend East to do? Decide before you read any further!

Let us apply game theoretic reasoning also to the new situation. To do so, we need to first represent the new situation by a new extensive form game. The new game must includes the possibility for East to burn the bridge before West decides whether to attack the island. The game tree in Figure 23.4 describes the new extensive

Figure 23.4.: Burning bridges

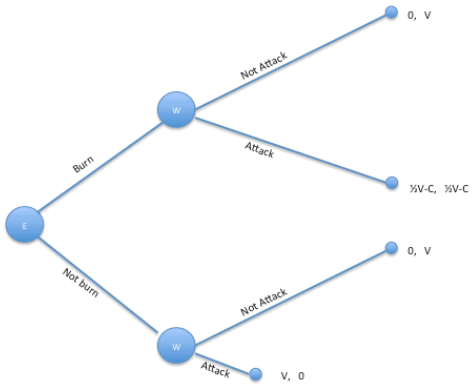


form game. East moves first and decides whether or not to burn the bridge. If East doesn't burn the bridge the continuation game is exactly the same game as before. If East burns the bridge, only West has a decision to make, either to attack or not. Absent the

possibility to retreat, East has to defend the island in case of an invasion. We continue to display West's payoff before East's payoff at each end-node.

Since also this game has perfect information, we solve it by using backwards induction. We start with East's decision following an attack from West, in case the bridge is still there. We already know the prediction: East would retreat. The following game tree is truncated to reflect this choice. Then next step is to study West's

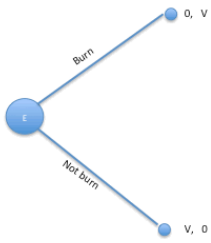
Figure 23.5.: Truncated game



choice. We need to study West's choice both for the case when East has burned the bridge and for the case when East has not burned the bridge. Clearly these are two different situations, since the payoffs following an attack are different. If East hasn't burnt the bridge, West predicts that East will retreat in case of an attack. West will therefore attack, just as before. On the other hand, if East has burnt the bridge, West knows that an attack would lead to war, in which case it is better for West not to attack. West will thus make different choices in period two, depending on East's choice in period one. To help analyze East's choice in the first period, the

game is further truncated by replacing West's decision nodes with the relevant end-nodes: At the first stage, East can predict that

Figure 23.6.: Further truncations



West will not attack in case the bridge is burned, but that West will attack otherwise. Burning the bridge is the better choice for East. East's generals realize that just enlisting an army, with lots of soldiers and equipment, is an empty threat in their situation. The incentive to fight matters. If retreating is the rational choice following an invasion, setting up an army doesn't deter an invasion. But by being committed to fight in case of an invasion, East can successfully deter West's aggression.

23.2. Game 2: Entry Deterrence

It is well known that monopoly firms sometimes charge prices that are much higher than their costs and that they therefore may enjoy substantial profits. High profits attract the attention from other firms, however. Firms in related markets and new start-ups may view the high profits as a signal of business opportunities. To deter

potential entrants from setting up new competing operations, an incumbent firm may threaten them with a price war, should they enter the market. The potential entrants may then prefer not to invest.

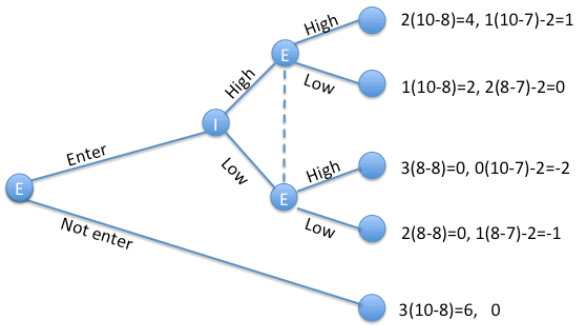
But such threats may not be effective. A price war is costly also to the incumbent. So, once entry has already occurred, the incumbent might not have an incentive to carry out the threat. Unfortunately, for the incumbent, only threats that would be carried out are credible and succeed to deter entry. In some markets, monopoly profits may therefore be short-lived. But in other markets monopolies persist for a long time. Can entry deterrence have anything to do with that?

23.2.1. Imperfect information

To analyze the issue of entry deterrence more formally, we describe the situation as a game with two time periods. In period one, the entrant decides whether to enter or not. In period two, the firms set prices simultaneously and then the consumers make their purchases. This game differs from the War & Peace Game in one important respect, namely that both players have to take simultaneous decisions following entry. When the entrant sets its price, it doesn't know the incumbent's price, and vice versa. It will soon be clear that this simultaneity calls for a slightly different type of analysis than the straight-forward backwards induction used in the War & Peace Game. To simplify the analysis we assume that the two firms' goods are perfect substitutes. Consumers are willing to pay $V = 10$ for one unit of the good but nothing for a second unit. There are three consumers, one of which is not aware of the entrant. This consumer will only buy from the incumbent. The two other consumers buy from the firm with the lowest price and, in case of equal prices, one buys from the incumbent and the other from the entrant. The marginal cost of producing the good is constant

to both firms. The entrant is more efficient and can produce at a lower marginal cost than the incumbent, $C_E = 7 < C_I = 8$. The cost of entry is $K = 2$. For simplicity, the firms can only choose between charging a high price, equal to the monopoly price, i.e. $p_H = V = 10$, or a low price equal to the marginal cost of the high-cost firm, i.e. $p_L = C_I = 8$. In case of no entry, the incumbent would charge the monopoly price, thus earn a mark-up of 2, and sell 3 units. In this case the incumbent makes a profit of 6 and the entrant makes a profit of zero. All other payoffs are summarized in the game tree. The first payoff is the incumbent's. Please make sure that you understand how the payoffs were computed.

Figure 23.7.: Entry deterrence



As already mentioned, this game differs from the War & Peace Game in one important respect, namely that both players have to take simultaneous decisions following entry. When the entrant sets its price, it doesn't know the incumbent's price, and vice versa. This means that the entrant doesn't know which of the two nodes it is at. We say that the two nodes belong to the same *information set* and illustrate this by connecting the entrant's two pricing setting nodes

by a dashed line in the game tree. Since the entrant is unaware of which node it is at, the entrant must make the same choice at both these nodes. (Of course we could equally well have chosen to describe the entrant's decision as a node to the left and the incumbent's decision as two connected nodes to the right.) When two or more players have to make decisions at the same time, we say that the game has *imperfect information*. (Actually the term imperfect information is broader. We say that a game has imperfect information also whenever a player has to take a decision without knowing what somebody else has done in the past.)

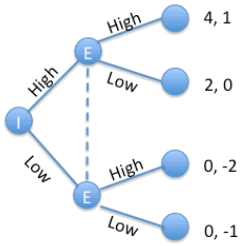
In a game with simultaneous decisions, the idea of backwards induction must be slightly modified. To see this, note that the entrant actually prefers different choices at his two price-setting nodes. If the incumbent charges a high price the entrant also prefers to charge a high price, but if the incumbent charges a low price the entrant also wants to charge a low price. It is therefore impossible to predict how the entrant will behave without predicting the incumbent's pricing decision at the same time. Straightforward backwards induction breaks down and we need to use a slightly more complicated procedure.

23.2.2. Sub-games and Sub-game perfect equilibrium

To explain this procedure, note that the price setting decisions in the second period, following entry, constitutes a game-tree in itself: Such a part of a game, which constitutes a game in itself, is called a *sub-game*. Here, we may refer to this sub-game as the duopoly pricing game. An key rule in forming such sub-games is that all nodes belonging to the same information set must belong to the same sub-game(s).

A Nash equilibrium in the complete entry game is a reasonable prediction only if it also prescribes a Nash equilibrium in the

Figure 23.8.: The duopoly-pricing subgame



duopoly pricing game. Otherwise at least one of the players would have an incentive to deviate from the equilibrium in the duopoly pricing game, should entry occur. A Nash equilibrium which induces a Nash equilibrium in every subgame is called a *subgame perfect equilibrium*. The essence of subgame perfection is the same as the essence of backwards induction. Consider again the War & Peace Game. A Nash equilibrium in the complete War & Peace Game is only a reasonable prediction if it prescribes a maximizing choice by East following entry. In fact, East's decision problem following entry may be viewed as a simple one-person subgame and its maximizing choice may be viewed as a Nash equilibrium of this subgame.

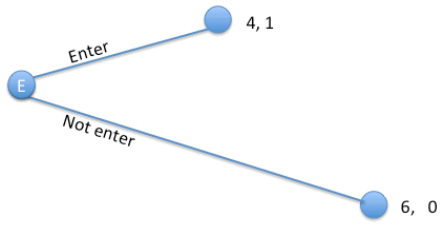
To determine the Nash equilibrium of the duopoly pricing subgame, it is convenient to write down the normal form of this game. Both players have two strategies, namely to charge a high or a low price. The game is represented by the following payoff matrix, with the incumbent as a row player and the entrant choosing column: Best replies are indicated by underlining the appropriate payoffs.

Table 23.2.: Normal form of the duopoly sub-game

	High	Low
High	$\underline{4}, \underline{1}$	$\underline{2}, 0$
Low	$0, -2$	$0, \underline{-1}$

The only equilibrium is for both to charge a high price. Knowing what equilibrium play must be in the duopoly pricing game, we can continue backwards to analyze the entry decision. To do so we can truncate the entry game by replacing the duopoly pricing subgame by an end-node and the equilibrium payoffs: Clearly, he entrant

Figure 23.9.: Truncated game



will choose to enter the market in the first period.

We therefore have a complete description of the subgame perfect equilibrium in the full entry game. In the first period the entrant enters and in the second period both firms set high prices. The new firm enters since it can predict that the outcome of the duopoly pricing sub-game will be more beneficial than staying outside the market.

23.2.3. Nash equilibrium

The equilibrium just described is not the only Nash equilibrium of the entry game, however. To see this we need to write down the normal form of the complete game. The incumbent has only one decision-node with two available actions, to set a high or a low price. Consequently, the incumbent has two strategies (two complete plans of actions) to choose from, namely to set a high or a low price. The entrant has two decision-nodes with two available actions at each node. The entrant therefore has four strategies (four complete plans of actions) to choose from, namely "entry & low price2, "entry & high price", "no entry & low price" and "no entry & high price".[1] The normal form is given by the following payoff matrix: Recall that incumbent's choice of high and low price

Table 23.3.: Nash equilibria

	Enter & High	Enter & Low	Not enter & High	Not enter & Low
High	$\underline{4}, \underline{1}$	$\underline{2}, 0$	$\underline{6}, 0$	$\underline{6}, 0$
Low	$0, -2$	$0, -1$	$\underline{6}, \underline{0}$	$\underline{6}, \underline{0}$

refers to the prices that the incumbent would set in following entry.

There are two types of equilibria. The first requires the entrant to enter and both firms to charge high prices. This is the sub-game perfect equilibrium discussed above. The second type of equilibrium requires the entrant to stay out of the market and the incumbent to charge a low price (following entry). This may be interpreted as a threat from the incumbent to wage a price war if the entrant enters the market. This threat is not credible, however, since the incumbent does not have an incentive to charge a low price, should entry occur. The entrant can foresee this and will therefore enter

[1] One might argue that it is superfluous to specify a pricing decision in case the new firm does not enter. We do so here, however, to clearly illustrate that strategies are complete plans of actions – they should tell the player what to do at all points in time and for all possible situations the player might find himself in.

the market.

23.2.4. Difference between Nash and Sub-game perfect equilibrium

So, what is the difference between a Nash equilibrium and a sub-game perfect equilibrium? The crucial difference is that a Nash equilibrium does not need to prescribe equilibrium play in sub-games that are not reached when the equilibrium is played. For example, an equilibrium prescribing the new firm not to enter does not need to prescribe an equilibrium of the duopoly pricing game, since that sub-game is not reached absent entry. A sub-game perfect equilibrium, on the other hand, requires equilibrium play in all sub-games. Expressed differently, a subgame perfect prediction requires realistic predictions also about how people will behave also in the situations that will not occur.

23.2.5. Commitment

Sometimes an incumbent can make its threat of low prices credible. One way would be for the incumbent to invest in a new technology to lower its marginal cost. The low marginal cost commits the incumbent to charge low prices. To analyze if an incumbent would have an incentive to do so, we would need to enrich the model by adding a new first period in which the incumbent is allowed to make the investment or not. This analysis is left as an exercise, see below.

23.3. Summary

- In many economic situations, the timing of decisions is important. Examples include entry deterrence, cartel formation and price negotiations. When the time dimension is important, it is often not convenient to describe the situation as a

normal form game; it is better to describe it as an extensive form game. An extensive form game describes who the players are, what actions they can choose from and the payoffs generated by the actions taken. But an extensive form game also describes the timing of the decisions: who moves first, second, third and so on.

- An extensive form game can be thought of as a game tree, which is a decision tree with several decision makers. The game tree describes the different decisions as nodes and the available actions as branches.

- In extensive form games it is important to distinguish between games with perfect information, which is games where players never move at the same time and where all previous choices are know to all players, and games with imperfect information, where there may be simultaneous choices and some early choices may be hidden from those who move late.

- It is always possible to rewrite an extensive form game as a game in normal form to find the Nash equilibria of the game. The key to such a transformation is to define the players' strategies. A player's strategy is a complete plan of action; it tells the player what action to take whenever it is this player's time to make a decision; it may prescribe different actions depending on what actions where taken in the past. The normal form describes who the players are, their strategies (as just defined) and their payoffs as functions of all possible combinations of strategies that the players can choose.

- Not all Nash equilibria are equally compelling as predictions of what the player will do, however. For example, while entry deterrence may be a Nash equilibrium, it may still be an unlikely outcome. An incumbent firm may threaten a new firm with a price war, should the new firm enter the market, and

the new firm may choose not to enter for this reason. The new firm's choice to stay out of the market is a best reply since a price war would make entry unprofitable. The incumbent's choice to be tough is a best reply when the new firm doesn't enter, since then the incumbent will not have to fight the war.

- But, if the new firm would actually enter the market, the incumbent might not have an incentive to start the price war. The incumbent is likely to lower its price from the monopoly level to the normal duopoly level. But reducing the price even further, to a really low level, perhaps below cost, is not only costly to the entrant, but also to the incumbent. The entrant, who can predict the incumbent's future rational reaction to entry, would then enter the market. So, if the incumbent does not have an incentive to fight the war, if entry actually occurs, the war is not a credible threat. We should therefore consider the Nash equilibrium prescribing the new firm to enter and the incumbent to charge normal duopoly prices as a more reasonable prediction.

- For this reason we define a new equilibrium concept called subgame perfect equilibrium, which picks out the more reasonable Nash equilibria. In games with perfect information, the subgame perfect equilibrium can be found using a simple technique called backwards induction. As the name suggests, the basic idea is to start the analysis with the last choices and then to analyze the game backwards. First, we study the incumbent's pricing choices given that entry has already occurred. Then, we study the new firm's choice to enter or not. The entry choice is informed by the new firm's ability to predict the incumbent's future behavior.

- The difference between a non-credible Nash equilibrium and a sub-game perfect equilibrium is that the Nash equilibrium

does not prescribe equilibrium play in sub-games that are not reached when the equilibrium is played. For example, an equilibrium prescribing a new firm not to enter does not need to prescribe an equilibrium of the duopoly pricing game, since that sub-game is not reached absent entry. A sub-game perfect equilibrium, on the other hand, requires equilibrium play in all sub- games.

23.4. Exercises

23.4.1. Short questions

1. What do you need to specify to define an extensive form game?
2. What is a game tree?
3. What is the difference between actions and strategies?
4. What is meant by perfect information?
5. What is meant by backwards induction?
6. What is meant by a sub-game?
7. What is meant by a sub-game perfect equilibrium?
8. What is meant by credible threat? Think about the Peace & War game.
9. What is meant by commitment? Think about the Peace & War game.

23.4.2. Divide & Choose

Two children are to split a cake. The first child gets to split the cake in two pieces and the second child gets to choose which piece he

wants. Both children want as much cake as possible for themselves. This is understood by both. For simplicity, you may assume that the first child cuts the cake into a left and a right piece and that the left piece must be 25%, 50% or 75% of the cake.

1. Write down the extensive form game tree!
2. Does this game have perfect or imperfect information?
3. Does this game have complete or incomplete information?
4. What is the appropriate solution concept?
5. Compute it!
6. Write down the normal form!
7. Find the Nash equilibria!
8. Is there any difference between the Nash equilibria and the equilibria you found question 5? Explain!

23.4.3. Entry deterrence

Consider a game with three time periods. In period one, the incumbent decides whether to invest in a cost reducing technology or not. In period two, the entrant decides whether to enter or not. In period three, the firms set prices simultaneously and then the consumers make their purchases. The two firms' goods are perfect substitutes. Consumers are willing to pay $V = 10$ for one unit of the good but nothing for a second unit. There are three consumers, one of which is not aware of the entrant. This consumer will only buy from the incumbent. The two other consumers buy from the firm with the lowest price and, in case of equal prices, one buys from the incumbent and one from the entrant. The marginal cost of producing the good is constant to both firms. Absent an

investment in the cost-reducing technology, the entrant is more efficient and can produce at a lower marginal cost than the incumbent, $C_E = 7 < C_I = 8$. If the incumbent makes the investment, the incumbent's cost is lowered to $C_I = 0$. The cost of the investment is $K_I = 25$. For simplicity, the firms can only choose between charging a high price, equal to the monopoly price, i.e. $p_H = V = 10$, or a low price equal to the marginal cost of the high-cost firm. Absent any investment in cost-reducing technology, the incumbent is the high-cost firm, in which case charging a low price means to set $p_L = 8$. In case of an investment in the cost-reducing technology, the entrant is the high-cost firm, in which case charging a low price means to set $p_L = 7$. In case of no entry, it is simply assumed that the incumbent charges the monopoly price. The cost of entry is $K = 2$.

1. Write down the game tree for this game.
2. Is it a game of perfect information or imperfect information?
3. What is the appropriate solution concept for this game (Nash equilibrium, Backwards induction or Subgame perfect equilibrium)?
4. Which subgames can you define?
5. Solve the game!
6. Will the incumbent invest in the cost-reducing technology?
7. Would the incumbent invest in the cost-reducing technology, absent the threat of entry?