Chapter 3

Production and Competition:
An Analytical Thermodynamic Theory

Economic and biological systems need to extract low entropy from the environment to compensate for continuous dissipation (Schrodinger, 1944; Georgescu-Roegen, 1971; Prigogine, 1980). This process can be represented by lognormal processes, which contain an extraction term and a dissipation term. The lognormal processes in turn can be mapped into a thermodynamic equation. From the entropy law, the thermodynamic diffusion of an organic or economic system is spontaneous. The extraction of low entropy from the environment, however, depends on specific biological or institutional structures that incur fixed or maintenance costs. The fixed costs help reduce variable costs in extracting low entropy from the environment. In this chapter, we solve the thermodynamic equation to derive an analytic formula that explicitly represents the relation among fixed costs, variable costs, uncertainty of the environment and the duration of a project, which is the core concern in most economic decisions. This analytical framework directly based on thermodynamic foundation greatly simplifies the understanding of economic activities and their relation with environmental changes.

3.1 Some Historical Background

Neoclassical economics was founded around 1870 by Jevons, Walras and others, who believed that economics should be built on a sound physical foundation. Since the dominant platform of physics in Jevons and Walras' time was rational mechanics, it was natural for them to adopt this platform.

About the same time when Jevons and Walras tried to establish economic theory on rational mechanics, physics had experienced a revolution whose impact is just to be felt in social sciences recently. Around 1870, Maxwell discussed the famous intelligent demon problem and Boltzmann established the theory of statistical mechanics. Boltzmann's theory met strong opposition because it seemed to contradict Newtonian mechanics. In despair, he wrote: "I am conscious of being only one individual struggling weakly against the current of time. But it still remains in my power to make a contribution in such a way that, when the theory of gases is again revived, not too much will have to be rediscovered." (Quotes from Isihara, 1971, p. 18) Boltzmann's growing pessimism led him to commit suicide in 1906, shortly before his theories were accepted.

Rational mechanics studies the movement of single particles. Statistical mechanics studies the dynamic properties of the systems of many particles. It is a generalization from rational mechanics. However, in its early days, people thought it was inconsistent with the rational mechanics. In rational mechanics, time is reversible. In statistical mechanics, time is irreversible and systems are evolutionary.

Gradually, the theory of statistical mechanics gained more influences. The first chapter of Wiener's (1948) *Cybernetics* was titled Newtonian Time and Bergosonian Time, in which he stated that most systems we encounter can be more precisely described by statistical mechanics instead of Newtonian mechanics. In Schrodinger's (1944) *What is Life*, he stated that the most fundamental property of life is their ability to extract negative entropy from the environment to compensate continuous dissipation. Since then some analytical theories related to thermodynamics have been developed, such as Lorenz' chaos theory and Prigorgine's far from equilibrium thermodynamic theory. Lorenz, a meteorologist, simplified weather equations, which are thermodynamic equations, into ordinary differential equations. He found chaos properties from these equations. Prigorigine developed the theory from some chemical reactions. These theories greatly influenced the thinking in biology and social sciences. However, they do not model life process or social activities directly. Chaos theory and Prigorgine's theory, while providing good insights to the research in social sciences, are mostly analogies.

As it is often the case, something from a totally different area turns out to be the key to the new development. In 1973, Black and Scholes developed an analytical theory of option pricing on financial assets. Initially it was a very technical subject. But gradually, it has been applied to many different areas. We will have a look at the basic properties of Black-Scholes theory to understand why it can be applied to so many different areas.

In Black-Scholes option theory, the price movement of financial assets is modeled with lognormal processes

$$\frac{dS}{S} = rdt + \sigma dz\,.$$

where r is the rate of expected return and σ is the rate of uncertainty. Option prices, as functions of prices of its underlying assets, satisfy the following Black-Scholes equation.

$$\frac{\partial C}{\partial t} + rS\frac{\partial C}{\partial S} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 C}{\partial S^2} = rC$$

I had been thinking about an analytical thermodynamic theory of life systems for many years when I learned about the Black-Scholes option theory. The most fundamental property of life is their ability to extract low entropy from the environment to compensate continuous dissipation. Soon I realized this property can be represented by lognormal processes, where r is the rate of extraction of low entropy and σ is the rate of diffusion. Every stochastic process can be mapped into a deterministic thermodynamic equation, which is often easier to handle and yields more results. So I hope Black-Scholes equation and option theory may offer some insight for an analytical thermodynamic theory of life systems. After several years, I first developed such a theory based an analogy between option theory and living systems (Chen, 2000). Later I was able to derive the theory directly without depending on its analogy with option theory. (Chen, 2002b) While my original interest was in life sciences, I now make a living teaching financial economics. That is why currently, this theory is mostly applied to social sciences. In the next section, we will describe this theory.

3.2.Basic Theory

All biological systems, human or non-human, need to extract low entropy from the environment to compensate for continuous dissipation. In human societies, most human activities are measured by economic value. Suppose S is the amount of low entropy of a biological system, r, the rate of extracting low entropy from the environment and σ, the rate of diffusion of the low entropy into the environment. Similarly in an economic system, S represents economic value of a commodity, r, the rate of change of the value of this commodity and σ, the rate of uncertainty. Then the process of S can be represented by the lognormal process

$$\frac{dS}{S} = rdt + \sigma dz \tag{3.1}$$

A production system is parallel to a biological system. A firm, which has blueprint to produce a product, such as cars, is similar to a biological entity, which has genes to produce offspring. The production of a good involves fixed cost, K, and variable cost, C, which are functions of the S, the value of the product. If the discount rate of a firm is q, from Feymann-Kac formula, (Øksendal, 1998) the variable cost, as a function of S, satisfies the following equation

$$\frac{\partial C}{\partial t} = rS\frac{\partial C}{\partial S} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 C}{\partial S^2} - qC \tag{3.2}$$

To solve for variable cost from this equation, we need to determine the initial condition that the variable cost has to satisfy at time zero. We perform a thought experiment about a project with a duration that is infinitesimally small. When the duration of fixed cost is infinitesimal small, the project has only enough time to produce one piece of product. If the fixed cost is lower than the value of the product, the variable cost should be the difference between the value of the product and the fixed cost to avoid arbitrage opportunity. If the fixed cost is higher than the value of the product, there should be no extra variable cost needed for this product. Mathematically, the initial condition for variable cost is the following

$$C(S,0) = \max(S - K,0) \qquad (3.3)$$

where S is the value of the product and K is the fixed cost. Suppose the duration of a project is T. Solving the equation (3.2) with the initial condition (3.3) yields the following solution

$$C = Se^{(r-q)T}N(d_1) - Ke^{-qT}N(d_2) \qquad (3.4)$$

where

$$d_1 = \frac{\ln(S/K) + (r + \sigma^2/2)T}{\sigma\sqrt{T}}$$

$$d_2 = \frac{\ln(S/K) + (r - \sigma^2/2)T}{\sigma\sqrt{T}} = d_1 - \sigma\sqrt{T}$$

The function $N(x)$ is the cumulative probability distribution function for a standardized normal random variable. When q, the discount rate, is equal to r, the rate of change, Formula (3.4) becomes

$$C = SN(d_1) - Ke^{-rT}N(d_2) \qquad (3.5)$$

It takes the same form as the well-known Black-Scholes (1973) formula for European call options. A more technical discussion of this theory can be found in Chen (2002c).

This theory, for the first time in economic literature, provides an analytic theory that explicitly represents the relation among fixed costs, variable costs, uncertainty of the environment and the duration of a project, which is the core concern in most economic decisions. “The progress of science is marked by the transformation of the qualitative into the quantitative. In this way not only do notions become turned into theories and lay themselves open to precise investigation, but the logical development of the notion becomes, in a sense, automated. Once a notion has been assembled mathematically, then its implications can be teased out in a rational, systematic way." (Atkins, 1994, p. 29)

A new theory is ultimately justified by its implications. We will look at the properties and implications of this theory. For simplicity, we will only explore the special case when the discount rate is equal to the rate of change, that is, Formula (3.5). Several properties can be derived from (3.5). First, when the fixed cost investment, K, is higher, the variable cost, C, is lower. Second, for the same amount of fixed investment, when the duration, T, is longer, the variable cost is higher. Third, when uncertainty, σ, is higher, the variable cost increases. Fourth, when the fixed cost approaches zero, the variable cost will approach to the value of the product. Fifth, when the value of the product approaches zero, the variable cost will approach zero as well. All these properties are consistent with our intuitive understanding with production processes.

Unlike a conceptual framework, this analytical theory enables us to make precise calculation of returns of different projects under different kinds of environments. First, we examine the relation between fixed cost and variable cost at different levels of uncertainty. For example, a product can be manufactured with two different technologies. One needs ten million dollars of fixed cost and the other needs one hundred million fixed cost. Assume the other parameters are unit value of the product, to be one million, discount rate, to be 10% and duration of the project, to be twenty-five years. When uncertainty of the environment is 30% per year, variable cost for the low fixed cost project is 0.59 million and variable cost for the high fixed cost project is 0.14 million, calculated from (3.5). When uncertainty of the environment is 90% per year, variable cost for the low fixed cost project is 0.98 million and variable cost for the high fixed cost project is 0.94 million. In general, as fixed costs are increased, variable costs decrease rapidly in a low uncertainty environment and decreases slowly in a high uncertainty environment. This is illustrated in Figure 3.1. In the extreme environment when the uncertainty reaches infinity, the variable cost is equal to the value of the product, regardless of the level of the fixed cost investment. In this environment, the value of any fixed asset is zero. For example, the elaborate institutional structures of the Roman Empire, which were once of great value, became worthless during the empire's chaotic collapse. In general, the value of any physical capital, institutional capital and human capital depends highly on the environment.

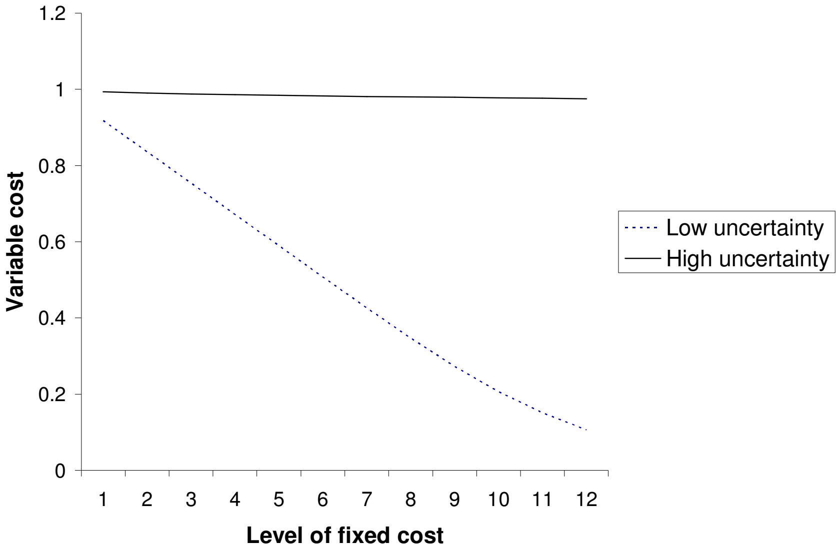

Figure 3.1 Uncertainty and variable cost

Next we discuss the returns of investment on different projects with respect to the volume of output. K is the fixed cost of production and C is the variable cost. Suppose the volume of output during the project life is Q, which is bound by production capacity or market size. We assume the present value of the product to be S and variable cost to be C during the project life. Then the total present value of the product and the total cost of production are

$$SQ \text{ and } CQ + K \tag{3.6}$$

respectively. The return of this project can be represented by

$$\ln\left(\frac{SQ}{CQ + K}\right) \tag{3.7}$$

Continuing the example on two technologies with different fixed costs, we now discuss how the expected market sizes affect rates of return. Suppose the level of uncertainty is 30% per year and other parameters are the same. If the market size is 100, the return of the low fixed cost project, calculated from (3.7), is 37% and the return of the high fixed cost project is -12%. When the market size is 400, the return of the low fixed cost project is 48% and the return of the high fixed cost project is 97%. Figure 3.2 is the graphic representation of (3.7) for different levels of fixed costs. In general, higher fixed cost projects need higher output volume to breakeven. At the same time, higher fixed cost projects, which have lower variable costs in production, earn higher rates of return in large markets.

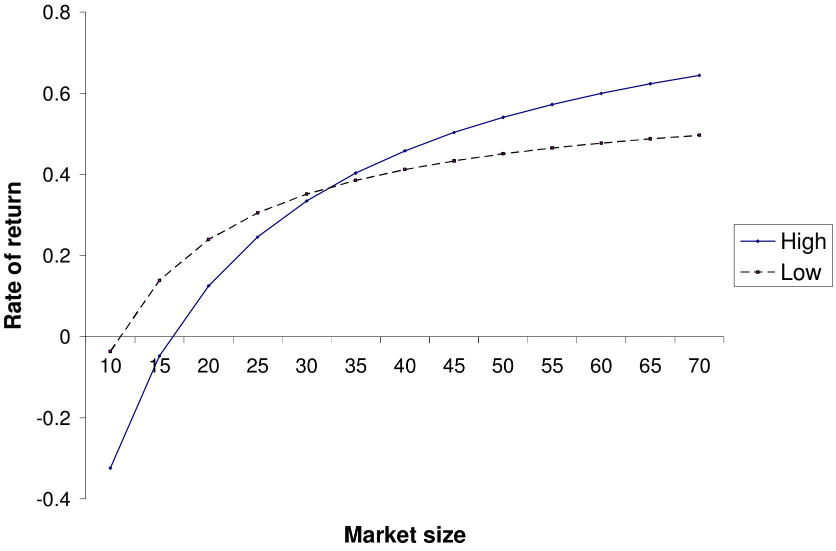

Figure 3.2 Market size and return with different levels of fixed costs

From the above discussion the level of fixed investment in a production system depends on the expectation of the level of uncertainty of production technology and the size of the market. When the outlook is stable and market size is large, production systems with high fixed investment earn higher rates of return. When the outlook is uncertain or market size is small, production systems with low fixed cost breakeven easier.

In the following, we will use the example of industry life cycles to illustrate the evolution of production systems. When an industry is new, there is a lot of uncertainty about future development. This environment offers opportunities for small companies with low fixed costs as they are more flexible. When uncertainty is high, increases in fixed costs will not reduce variable costs very much. (Figure 3.1) Economy of scale is not significant under high risk condition, which permits many low fixed cost companies to compete. When an industry becomes mature and uncertainty decreases, increases in fixed costs, K, (capital investments and accumulated human capital), drive down variable costs rapidly, which permits leading companies to lower product prices and drive out small high variable cost companies. So, in a mature industry only very few big companies can stay in business. In fact, without anti-trust legislation, many industries with high capital intensity, or high fixed assets would probably end with monopolies or regional monopolies (Acs and Audretsch, 1988; Mazzucato, 2000).

While large firms with high fixed cost are highly competitive in their own fields, they are less effective in entering new markets with high uncertainty. First, from Figure 3.2, high fixed cost systems need a large market size to break even. So they tend to hesitate in moving into new markets, which are typically of small size in their early stages. Second, the high concentration of wealth in large companies often attracts litigation and other attempts to extract wealth from them. So large companies often incur high legal costs and are cautious in pursuing new opportunities. Third, internal coordination in large companies is much more complex and difficult than in small companies. While it is relatively easy to incorporate innovative ideas by adjusting firm structures in simple and small companies, innovation is often very disruptive to a highly coordinated and efficient complex structure. Large companies often develop highly optimized structures to reduce uncertainty and bring down variable costs in producing particular products. This however often stifles the innovative spirit inside the companies and makes it difficult for them to adjust in a changing environment.

In general, large companies, which have invested a great amount on existing technologies, may be unwilling or unable in the short term to switch to new and potentially better technologies. This opens opportunities for new and typically small companies when new industries emerge. For example, the champions of the IT revolution, such as Microsoft, Intel, CISCO, Oracle, AOL, are all relatively new companies that reaped tremendous profits by being able to quickly respond to and take advantage of newly emerging markets. Microsoft is now a mature company in a rapidly evolving industry. Its success and the resulting high concentration of wealth within Microsoft has attracted litigation and other attempts to extract from it wealth and market share.

3.3. A Comparison with Neoclassical Economic Theory

Since its birth, the foundation or "assumptions" of neoclassical economic theory has been criticized for its lack of relevance to reality. To this, Friedman replied:

> In so far as a theory can be said to have "assumptions" at all, and in so far as their "realism" can be judged independently of the validity of predictions, the relation between the significance of a theory and the "realism" of its "assumptions" is almost the opposite of that suggested by the view under criticism. Truly important and significant hypotheses will be found to have "assumptions" that are wildly inaccurate descriptive representations of reality, and in general, the more significant the theory, the more unrealistic the assumptions (in this sense). The reason is simple. A hypothesis is important if it "explains" much by little, that is, if it abstracts the common and crucial elements from the mass of complex and detailed circumstances surrounding the phenomena to be explained and permits valid predictions on the basis of them alone. To be important, therefore, a hypothesis must be descriptively false in its assumptions; it takes account of, and accounts for, none of the many other attendant circumstances, since its very success shows them to be irrelevant for the phenomena to be explained. (Friedman, 1953, p. 16)

He further challenged:

> As we have seen, criticism of this type is largely beside the point unless supplemented by evidence that a hypothesis in one or another of these respects from the theory being criticized yield better

> predictions for as wide a range of phenomena. (Friedman, 1953, p. 31)

In the next several chapters, we will offer more detailed discussion on how the new theory, which is based on more realistic assumptions, does "yield better predictions for as wide a range of phenomena". In the following, we will compare the new theory briefly with neoclassical economic theory.

Consistency with physical and biological theories

Neoclassical economics was founded around 1870 by Jevons, Walras and others, who believed that economics should be built on a sound physical foundation. Since the dominant platform of physics in Jevons and Walras' time was Newtonian mechanics, it was natural for them to adopt this platform. However, theories derived from rational mechanics often do not offer good explanation to economic behaviors. Gradually, explicit identification with physics disappears while analogies between physics and economics are frequently mentioned. The following quote from Samuelson's Nobel lecture is quite representative:

> There is really nothing more pathetic than to have an economist or a retired engineer try to force analogies between the concepts of economics. How many dreary papers have I had to referee in which the author is looking for something that corresponds to entropy or to one or another form of energy.

In the very next paragraph, however, Samuelson found some analogy himself.

> However, if you look upon the monopolistic firm hiring ninety-nine inputs as an example of a maximum system, you can connect up its structural relations with those that prevail for an entropy-maximizing thermodynamic system. Pressure and volume, and for that matter absolute temperature and entropy, have to each other the same conjugate or dualistic relation that the wage rate has to labor or the land rent has to acres of land.

Mirowski observed, "The key to the comprehension of Samulson's meteoric rise in the economics profession was his knack for evoking all the outward trapping and ornament of science without ever coming to grips with the actual content or implications of physical theory for his neoclassical economics" (Mirowski, 1989, p. 383).

Life systems are non-equilibrium thermodynamic systems. The current dominant economic theory is general equilibrium theorem. Social system is a special case of living systems. When a theory about a special case is inconsistent with general foundation, either the general foundation or the special theory is wrong. So far, economists have not challenged the validity of the non-equilibrium thermodynamic theory of life systems. This theory shows that an analytical theory of economics can be directly derived from basic physical and biological laws. By this, it establishes social sciences as an integral part of physical and biological sciences.

A comparison with production functions

Production functions, such as Cobb-Douglas production function, form the fundamental blocks in general equilibrium production theory. Cobb-Douglas function takes the form

$$Y = AL^{\alpha} K^{\beta}$$

where Y, L and K denote output, labor (variable cost) and capital (fixed cost) respectively. Solow had made following comment about the production function:

> I have never thought of the macroeconomic production function as a rigorously justifiable concept. In my mind it is either an illuminating parable, or else a mere device for handling data, to be used so long as it gives good empirical results, and to be discarded as soon as it doesn't, or as something better comes along. (Solow, 1966, p. 1259)

By contrast, the analytical production theory developed here is derived rigorously from the fundamental property of life systems. It gives simple and clear results of returns to investment under different market conditions. The form and parameters of Cobb-Douglas function are given without rigorous justification. *A*, the coefficient in Cobb-Douglas function, "has been called, among other things, 'technical change', 'total factor productivity', 'the residual' and 'the measure of our ignorance'" (Blaug, 1980, p. 465).

Since production functions are widely used in economic literature in constructing economic models, even a small improvement on this topic should have a big impact in understanding economics.

Optimality vs. trade-off

Optimization theory holds the central position in neoclassical economics. Paul Samuelson's Nobel Lecture is titled *Maximum Principles in Analytical Economics*. Alchian (1950) and Friedman (1953) tried to reconcile the maximization principle with evolutionary theory. Friedman stated:

> Confidence in the maximization-of-return hypothesis is justified by evidence of a very different character. … unless the behavior of businessmen in some way or other approximated behavior consistent with the maximization of returns, it seems unlikely that they would remain in business for long. Let the apparent immediate determinant of business behavior be anything at all --- habitual reaction, random chance, or whatnot. Whenever this determinant happens to lead to behavior consistent with rational and maximization of returns, the business will prosper and acquire resources with which to expand; whenever it does not, the business tend to lose resources and can be kept in existence only with addition of resources from outside. The process of "natural selection" thus helps to validate the hypothesis --- or rather, given natural selection, acceptance of the hypothesis can be based largely on the judgment that it summarized appropriately the conditions for survival. (Friedman, 1953, p. 22)

We will use an example of project investment to illustrate the problem of Friedman's argument. Assume the relevant parameters are unit value of the product to be one million, discount rate to be 4%, diffusion to be 40%, duration of the project, to be thirty years and market size to be 150 over the project life. It can be calculated that a project with a fixed cost of 25 million dollar will be optimal. However, if any parameter changes, the optimal value of fixed cost investment will change as well. For example, if diffusion increases to 60%, the optimal value of fixed investment will become 11 million. Since fixed cost is spent or committed at the beginning of the project while other parameters may change over the course of project life, it is impossible to determine optimality in advance. Furthermore, higher fixed cost systems, which are often the winners of earlier market competition, suffer more from the increase of uncertainty. This means that long term survival is not necessarily consistent with short term optimization.

Earlier, we have shown that systems with higher fixed costs earn higher rates of return in large markets and stable environments than those with lower fixed costs. These systems may appear superior. However, the performance of high fixed cost systems deteriorates in high volatile environments. In Chapter 4, we will show that the main theme of economic and biological evolution is the tradeoff between competitiveness of high fixed cost systems in a stable environment and flexibility of low fixed cost systems in a volatile environment. Biologists haven't found a universally applicable measure of fitness. (Stearns, 1992, p. 33) Our theory shows that there does not exist such a measure. For the same reason, there will not exist a universally applicable measure of optimality.

On imperfection

In neoclassical economics, many phenomena that are not consistent with theories are considered "imperfect". For example, from Modigliani and Miller (1958) theory, in a perfect market, capital structure is irrelevant. Since capital structure is relevant in reality, the real market is imperfect. There are many similar terms, such as "imperfect information", "imperfect contract", "imperfect competition", "inefficient property right", "market failure", "government failure", "externality". Before discussing these imperfections, we briefly review the idea of imperfection in old astronomy.

Ancient people had long observed that stars moved in perfect harmony in the sky. Several planets, however, moved in irregular trajectories. It was thought that this was caused by the imperfectness of the planets. There were many elaborate theories why the planets were imperfect. Kepler, however, derived that all planets moved in perfect elliptic orbits. This story tells us that "imperfection of the world" often reflects imperfection of the theory that is used to understand the world. In later chapters, we will explain how this analytical thermodynamic theory offers a unified understanding of various “imperfection” or “externality”. In the following we will briefly discuss “imperfect competition”. Again, we quote Friedman as a reference point.

> The theory of monopolistic and imperfect competition is one example of the neglect in economic theory of these propositions. The development of this analysis was explicitly motivated, and its wide acceptance and approval largely explained, by the belief that the assumptions of “perfect competition” or “perfect monopoly” said to underlie neoclassical economic theory are a false image of reality. And this belief was itself based almost entirely on the directly perceived descriptive inaccuracy of the assumptions rather than on any recognized contradiction of predictions derived from neoclassical economic theory. The lengthy discussion on marginal analysis in the American Economic Review some years ago is an even clearer, though much less important, example. The articles on both sides of the controversy largely neglect what seems to me clearly the main issue --- the conformity to experience of the implications of the marginal analysis --- and concentrate on the largely irrelevant question whether businessmen do or do not in fact reach their decisions by consulting schedules, or curves, or multivariable functions showing marginal cost and marginal revenue. Perhaps these two examples, and the many others they readily suggest, will serve to justify a more extensive discussion of the methodological principles involved than might otherwise seem appropriate. (Friedman, 1953, p. 16)

From value theory developed in Chapter 2, the unit value of a product is

$$-\log_b P$$

where b is the number of producers and P is the abundance of the product. If the market is highly competitive, that is, b and P are very large, then the value of the product is very low. This shows that goods produced in nearly perfect competitive markets are of low economic values and models of perfect competition are of little use in describing most important and dynamic economic sectors. For example, companies of high market valuation such as Microsoft and Intel all have dominant positions in their own industries. So imperfectness in competition is the very source of economic value. To appreciate the magnitude of valuation, we calculate the values of two products, one in highly competitive market and other in less competitive market. Suppose the first product is produced by 10 firms and with a market saturation of 90% and the second product is produced by 2 firms with a market saturation of 20%. The values of two products are

$$-\log_{10} 0.9 = 0.046 \quad \text{and} \quad -\log_2 0.2 = 2.322$$

respectively. This means highly competitive markets are of extremely low economic value.

Strictly speaking, perfect competition can not even be properly defined. Form (3.5), variable cost is a function of fixed cost, uncertainty and other factors. Suppose one product is manufactured by several companies. If from today's perspective, the technologies or organizational structures of all competing companies are not "perfect", does it always induce some new firms or existing firms to adopt better technologies or organizational structures? This is not necessarily the case because a new technology always involves fixed cost. Empirical evidences and real option theory suggest that required rates of return can be substantial before new entry can occur. Therefore, perceived "imperfection" can persist indefinitely.

Is marginal cost equal to marginal revenue?

Traditional economic theory suggests that companies will keep increasing the output until the marginal cost of the product is equal to its marginal revenue (Friedman, 1953, p. 16). Empirical evidences show that companies generally charge a substantial price mark up on their products. This analytical theory offers a simple and clear understanding about price markup. For example, if a software is targeted to sophisticated users, its interface can be simple, which reduce development cost and its sales effort can be small, which reduce variable cost. If the software developer considers increasing the market size by targeting general users, the interface of the software needs to be very intuitive with many help facilities, which increase development cost and its sales effort and after sales service can be substantial for less sophisticated users, which increase variable cost. Since the increase of market size often involve both the increase of variable cost and fixed cost, most projects are designed that the marginal cost to be much lower than the product value to maximize potential profit.

To keep increasing the output until its marginal cost equal to marginal revenue often means that the company may have to enter difficult areas, which will have repercussion on its earlier units. For example, when employees in a WalMart store in Quebec decided to unionize, WalMart closed down that store although that store would remain profitable. To keep a unionized store open will affect the margin of other stores, whose staff will attempt to unionize as well. From (3.7), the rate of return not only depends on the market size, but also depends on other factors. If the increase of market size will increase the diffusion rate as well, companies have to consider the total effect on long term profitability.

From (2.5) and (2.6), the increase of production will decrease unit value and ultimately total revenue. Therefore, if possible, companies will not keep increasing the level of output until its marginal cost is equal to marginal revenue.

Market and regulation

In neoclassical economics, regulation is justified when there is a "market failure". From this theory, it can be shown that regulation is largely driven by industries themselves to keep high rate of returns. We will use an example to illustrate how the level of regulation is influenced by sizes of the markets.

From the value theory, the unit value of a product is

$$-\log_b P$$

where b is the number of producers and P is the abundance of the product. Suppose the market size is M. MP is the number of customers. Then on average, each producer's revenue is

$$\frac{MP}{b}(-\log_b P)$$

To produce the product, it takes fixed cost and variable cost. The value of a product will increase if the number of producers can be reduced. Since monopoly often brings a lot of legal actions, we assume companies will aim to achieve duopoly by raising fixed costs with regulatory or other means.

Suppose the cost structure of each company is the same. Then the total cost of each firm is

$$K + C\frac{MP}{b}(-\log_b P)$$

where *K* is the fixed cost and *C* is the variable cost, which is defined by the percentage of the product value. The rate of return for each company is

$$\ln\left(\frac{\frac{MP}{b}(-\log_b P)}{K + C\frac{MP}{b}(-\log_b P)}\right)$$

Table (3.1) displays the relation among fixed costs, market sizes and returns of companies where there are three or two companies producing the same product. Variable cost is computed from (3.5), assuming S =1, r= 0.03, T = 25, σ = 0.55. Here variable cost means the percentage of the value of the product. *P* is assumed to be equal to 0.5.

If possible, firms will raise fixed costs so that three competing firms will lose money on average. This will allow only two firms in the market. Then we calculate the average returns for these two firms. As can be shown from Table 3.1, as market sizes increase, the level of fixed costs needed to maintain duopoly increases as well, so are the rates of return. This indicates that leading companies in industries with large market size often have strong incentives to help introduce costly regulatory requirements for their products to reduce the number of competitors to gain higher rate of return. This is why industries of vital importance to most families, such as education, health care and pharmaceutical industry, are heavily regulated. In fact, primary and secondary educations in most countries have achieved total monopoly. In a later chapter, we will discuss why trade is mostly free, while migration, which of much larger economic impact, is highly regulated (Hamilton and Whalley, 1984; Moses, J. and Letnes, 2004).

Market size	200	400	800
Fixed cost	5.55	16.8	44.75
Variable cost	0.737	0.601	0.4676
Revenue to each firm when there are three firms	21.03	42.06	84.124
Rate of return	0.000	0.000	0.000
Revenue to each firm when there are two firms	50	100	200
Rate of return	0.165	0.263	0.3692

Table 3.1: Required levels of fixed cost to achieve duopoly with different market size

The above calculation also explains why biological and chemical weapons are banned by international treaties while nuclear weapons, which can cause much more destruction than chemical weapons, are not. Biological and chemical weapons, which are sometimes called poor men's nuclear weapons, are cheaper to make. If these weapons are not banned, many people can make them, which will reduce the value of weapons of mass destruction. To maintain the high value of such weapons, international treaties, which are generally initiated by leading political powers, banned weapons of mass destruction that are cheap to make.

3.4. What is Next?

North (1981) pointed out some missing parts in the current economic theory. They are (p. 68)

1. Theory of demographic change.
2. Theory of the development of military technology: Military technology and changes in military technology were crucial to the structure and size of the state in history.
3. Deficiency in the model of state, especially on modern pluralist state.

4. Neat supply function of new institutional arrangements.
5. A positive theory of the sociology of knowledge.

In the next two chapters, we will apply the theory developed in the first three chapters to major economic and social problems. In the process, five problems listed by North will also be resolved.