Dynamic Product Differentiation and Economies of Scale

A Simplified Approach to Irreversible Investment

Doctoral dissertation submitted to the
Norwegian School of Economics and Business Administration,
Bergen, Norway

May 1998

Sigbjørn Sødal

Agder College, Norway

to Helje

– *in whom I made my most irreversible investment*
when the markup was more than large enough

Acknowledgements

One of the characteristic features of the models in this thesis is the influence of accidental events and history. As a matter of fact, the thesis is itself a good example. For various reasons I ran into some of the people who have helped me into the position to write it. Not all the events and all these people can be mentioned, but some of them deserve special attention.

On the academic arena, I got to know Victor Norman mainly because some bureaucrats had made a number of mistakes in a much debated political issue a few years ago. At that time, economics was nowhere near my main interest, but nonetheless, he encouraged me to become a full-time economist. He also offered to be my adviser when I entered a doctoral programme at Agder College. Without his suggestions on which topics and literature to focus on, his patience when discussing and criticizing preliminary results, and his inspiring enthusiasm in economics, this thesis would never have come about. I am deeply grateful to him for all this.

I would also like to thank Anthony Venables and Terje Lensberg, who have been members of the dissertation committee. In a friendly and clever manner they have criticized preliminary versions of the thesis, making suggestions that have improved it significantly.

Hans Jarle Kind called me one Saturday a couple of years ago, informing me that Avinash Dixit, whom I had never met, were to hold a postgraduate course in real options theory in Helsinki. However, the deadline for applications was the following Monday. I ended up as a participant, and, as a matter of fact, the main research idea in the thesis was clarified during this course. Avinash Dixit and Robert Pindyck most generously invited me to exploit the idea together with them, and the first chapter is a result of this. I am grateful to them for giving me the opportunity to cooperate with such distinguished researchers, as well as for their friendly and respectful attitude. They have both also commented on preliminary texts that have ended up in various other parts of the thesis. In particular, Avinash Dixit objected to an argument that could have increased the number and seriousness of errors.

Several colleagues at Agder College and Centre for International Economics and Shipping have contributed by making up an academic environment which has been very important to

me. The numerous discussions with Trond Randøy, and the feedback from Steen Koekebakker on preliminary texts must be emphasized. I would also like to thank Anne Liv Scrase for her check of the manuscript, which has reduced the number of misprints and improved my English in many ways.

On the private arena, my wife Helje deserves respect for bearing much of the cost, and I am most grateful to her for this. Ellen, Karen and Gjermund have also suffered, since their father has not been as available as he ought to. I am also indebted to my parents, who have taught me a lot about investment under uncertainty over the years by their ways of spending time with me. In many cases, their practical approach is far more useful than the theoretical approach that follows on the next 180 pages.

Kristiansand, May 6, 1998

Sigbjørn Sødal

Summary

It is often assumed that dynamic modelling of investment under uncertainty is too complex to gain wide acceptance as a valuable tool in mainstream economics. For example, as far as important research areas like international trade and economic growth are concerned, there is not an overwhelming literature on irreversibility and dynamic uncertainty. The main objective with this dissertation is to show that stochastic analysis can be simplified, and applied more extensively in such fields.

By use of a new approach to irreversible investment, a large number of firm-level and equilibrium models are developed. The approach is based on analogy to static modelling, and it simplifies the discussion of several dynamic parameters. In some cases, the analogy to static modelling is so close that well-known models can be used directly in a dynamic context by reinterpreting crucial variables. The approach is developed stepwise.

Part I presents the methodology at the firm-level. In *Chapter 1*, it is argued that there is a close relationship between the markup pricing rule of a static monopolist facing a downward-sloping demand curve conditional on fixed marginal costs, and the optimal timing decision for a firm with the option to invest a fixed amount to obtain a fluctuating benefit. This chapter has been co-authored with Avinash Dixit and Robert Pindyck, and I have attached two notes to it. The first one supplements previous examples by deriving the analogy between the static and dynamic approach in case of linear demand; the second one discusses a "dual" set of models based on a fixed benefit and a fluctuating cost.

In *Chapter 2*, the "smooth pasting" condition is derived for the most basic investment problem. It is well-known that this is a first-order condition for optimum, but nevertheless, it has remained somewhat mysterious. That is probably because the standard options approach assumes the smooth pasting condition instead of deriving it. By the new approach, it appears as any other first-order condition for maximum by setting the derivative of expected profits equal to zero.

Part II extends the approach to equilibrium based on homogeneous goods. It is noted, however, that many of the results obtained in Part II carry over to the more advanced setting with product differentation, which is discussed later.

Chapter 3 presents the equilibrium framework for a large number of general and specific assumptions, showing that the analogy to static modelling still holds. *Chapter 4* uses this framework for a discussion on the relationship between irreversible investment and endogenous growth, emphasizing implications of uncertainty. Several results obtained by other models are also confirmed. *Chapter 5* contains a similar discussion on the relationship between irreversible investment, trade and agglomeration. This chapter also confirms several results obtained by more common, and mainly static, models. However, it sheds new light on the influence of a number of parameters related to irreversibility and dynamic economies of scale.

Part III extends to product differentation. The main result is the dynamic interpretation of the Dixit-Stiglitz model that is presented in *Chapter 6*. It is noted that all major properties of the static model, like gains from scale and trade, also apply in the dynamic setting. In *Chapter 7*, the analogy is explored in some more detail by a dynamic interpretation of Paul Krugman's "core-periphery" model. It seems likely that a number of other static Dixit-Stiglitz models can be interpreted similarly.

Part IV is more suggestive than the rest. The objective is to show how some of the results can be put together, and to point at issues that need further research. *Chapter 8* contains a discussion on growth and agglomeration by combining the growth results from *Chapter 4* with the geography results from *Chapter 7*. In *Chapter 9* it is discussed how intra-industry trade and gains from scale can arise even with a mixture of homogeneous goods, a large number of firms, and free entry. *Chapter 10* contains a firm-level model that focuses on geographical entry and exit decisions under uncertainty of demand.

- - -

Finally, it should be noted that *Chapter 1* and *Chapter 2* are self-contained articles, as opposed to the remaining chapters; notation is also somewhat different. A slightly revised version of *Chapter 1* has been submitted to *The Economic Journal*, and the editor has indicated that it is likely to be published. *Chapter 2* has been published in *Economics Letters*.

Table of Contents

Acknowledgements i
Summary iii
Table of Contents v

PART I - Firm-Level Models 1

1. A Markup Interpretation of Optimal Rules for Irreversible Investment. (Co-authors: Avinash Dixit and Robert S. Pindyck) 3

   Notes: A. Linear Demand. 20

   B. Fluctuating Costs. 23

2. A Simplified Exposition of Smooth Pasting. 28

PART II - Equilibrium Models with Homogeneous Goods 37

3. A Framework for Equilibrium Modelling. 39

4. Endogenous Growth. 84

5. Agglomeration with True Externalities. 103

PART III - Equilibrium Models with Product Differentiation 121

6. Product Differentiation. 123

7. Agglomeration with Pecuniary Externalities. 138

PART IV - Miscellaneous 157

8. Agglomeration and Growth. 159

9. Time Differentiation. 165

10. Geographical Entry-Exit. 171

PART I

Firm-Level Models

CHAPTER 1

A Markup Interpretation of Optimal Rules for Irreversible Investment*

Co-authors: Avinash Dixit, Princeton University
Robert S. Pindyck, Massachusetts Institute of Technology

Abstract: We re-examine the basic investment problem of deciding when to incur a sunk cost to obtain a stochastically fluctuating benefit. The optimal investment rule satisfies a trade-off between a larger versus a later net benefit; we show that this trade-off is closely analogous to the standard trade-off for the pricing decision of a firm that faces a downward sloping demand curve. We reinterpret the optimal investment rule as a markup formula involving an elasticity that has exactly the same form as the formula for a firm's optimal markup of price over marginal cost. This is illustrated with several examples.

Keywords: Investment, sunk costs, pricing decisions, optimal markups

JEL classification: D92, D81, E22

---

* Reprint (with minor corrections) of *NBER Working Paper* 5971, March 1997, with permission from Avinash Dixit and Robert S. Pindyck.

1. Introduction

Consider what is probably the most basic irreversible investment problem: a project can be undertaken that requires a sunk cost C and yields a benefit V. The cost is known and certain, but the benefit (measured as the present value at the time the cost is incurred) fluctuates as an autonomous Markov process $\{V_t\}$ with continuous sample paths.[1] Time is continuous, and at each point the firm must decide whether to invest or to wait and reconsider later. The firm's objective is to maximize the expected present value of net benefits, with a discount rate that is constant and equal to ρ.

At time t, all of the information about the future evolution of V is summarized in the current value V_t. Therefore the optimal decision rule must be of the form: invest now if V_t is in a certain subset of possible values, otherwise wait. Also, because the process is autonomous and the discount rate is constant, the optimal rule will be independent of time. So long as the process has positive persistence - i.e., a higher current value V_t shifts the distribution of the random value V_s at any future time s to the right in the sense of first-order stochastic dominance - the rule will be of the form: invest now if V_t is at or above a critical threshold V^*, otherwise wait.[2] The problem therefore boils down to determining the optimal choice for the threshold V^*.

As first shown by McDonald and Siegel (1986), the optimal V^* exceeds C by a "markup", or premium, that reflects the value of waiting for new information. One can think of the firm as having an option to invest that is akin to a financial call option, and, like the call option, is optimally exercised only when "deep in the money", i.e., when the stock price is at a premium over the exercise price. Thus one can solve the firm's investment problem (and determine the optimal markup) by finding the value of the firm's option to invest and the optimal exercise

[1] V may itself be explained in terms of other more basic economic variables like prices of output and/or inputs; we work simply with the end result.

[2] See Dixit and Pindyck (1996), pp. 104, 128-9.

rule.[3] Indeed, identifying and valuing the firm's option to invest has become the standard approach to solving irreversible investment problems.

However, as Baily (1995) has pointed out, an alternative way to find the optimal V^* is to examine the trade-off between larger versus later net benefits. Specifically, choosing a larger value for V^* implies that the net benefit, $V^* - C$, will be larger, but will be received at a more distant (but unknown) time in the future, and thus will be discounted more heavily. The optimal choice of V^* is that for which the additional net benefit from making V^* larger just balances the additional cost of discounting.

In this paper, we take this alternative perspective further by developing an intuitively appealing analogy with the trade-off involved in the pricing decision of a firm facing a downward-sloping demand curve - i.e., the trade-off between a higher profit margin and a lower volume of sales. We show that V can be regarded like a price, $(V - C)$ like a profit margin or markup, and the discount factor like a demand curve. The optimal V^* is then given by a markup formula involving the elasticity of the discount factor with respect to V, which has exactly the same form as the formula for a firm's optimal markup of price over marginal cost. This suggests extensions of the basic investment problem by analogy with the corresponding extensions of the monopolist's pricing problem. Here we develop one, namely the optimal choice of an ancillary expenditure in advertising or R&D which can speed up the (stochastic) passage to the threshold. The result is analogous to the formula for a monopolist's optimal advertising-to-sales ratio.

2. The optimal markup

Suppose the initial level of the benefit is V_0, and consider an arbitrary threshold $V > V_0$. Thus the firm will wait until the first time T at which the benefit V_T has reached V, and will then invest. (In technical terms, T is the first-passage time or hitting time from V_0 to V.) This time T is a random variable, and its distribution can be determined from the known probability law

[3] The option is valued assuming it is exercised optimally, so the valuation of the option yields the optimal exercise rule. See Dixit and Pindyck (1996).

of the evolution of V_t. Taking expectations using this distribution, the net present value of this policy is

$$E\left[e^{-\rho T}\right](V - C).$$

Note that the expectation of the discount factor in this expression depends on both the initial value V_0 and on the threshold value V of our decision rule. We therefore denote this discount factor as:

(1) $$D(V_0, V) \equiv E\left[e^{-\rho T}\right].$$

The *optimal* threshold, V^*, is the value of V which maximizes

(2) $$D(V_0, V)(V - C).$$

The first-order condition for the optimal V^* is

(3) $$D(V_0, V^*) + D_V(V_0, V^*)V^* = D_V(V_0, V^*)C,$$

where D_V is the partial derivative of the discount factor D with respect to its second argument, namely the threshold value V, and we are evaluating this at $V = V^*$. This condition simply says that if the investment opportunity is to be optimally exercised, the expected marginal discounted benefit from the investment should just equal the expected marginal discounted cost.

We can rewrite eqn. (3) in the following equivalent form:

(4) $$\frac{V^* - C}{V^*} = \left[-\frac{V^* D_V(V_0, V^*)}{D(V_0, V^*)}\right]^{-1} = 1/\varepsilon_D \ ,$$

where ε_D denotes the elasticity of the discount factor D with respect to V^*, i.e., $\varepsilon_D \equiv - V^* D_V / D$. The form of this expression should be very familiar: it is just like the markup pricing rule that follows from equating marginal revenue with marginal cost:

$$\frac{p - c}{p} = 1/\varepsilon_P \ ,$$

where p is the price, c is the marginal cost, and ε_p is the magnitude of the price elasticity of the firm's demand.

There is indeed a close connection between equation (4) for the investment markup and the markup pricing rule. To see this, compare the expression for the present value, (2), to that for the firm's profit in the usual pricing problem when marginal cost is constant, namely $(p-c)q(p)$. A higher p implies a higher profit margin (p - c), but a lower volume of sales $q(p)$. The trade-off that determines the optimal price is governed by the rate at which $q(p)$ declines as p is increased, i.e., by the price elasticity of demand. In our investment problem, a higher threshold value V^* yields a higher margin (V^* - C) of benefits over costs, but a smaller discount factor $D(V_0,V^*)$ because the process is expected to take longer to reach the higher threshold. The investment trade-off depends on the elasticity of the discount factor with respect to the threshold.

We can put this analogy in graphical terms by considering an arbitrary threshold V, and re-writing eqn. (3) as

$$(5) \qquad V + D(V_0,V) / D_V(V_0,V) = C.$$

We can think of the first term in this equation, $V = V(D,V_0)$, as the inverse of the discount factor: it is analogous to the inverse demand, or average revenue function, $p(q)$, for the price-setting firm. Likewise, the discount factor $D(V_0,V)$ is analogous to quantity for the price-setting firm, so the left-hand side of eqn. (5) - the marginal benefit from an increase in D - is analogous to the marginal revenue function.[4] These two functions of the discount factor D are plotted in Figure 1. The optimal threshold V^*, and the corresponding optimal discount factor $D^*(V_0,V^*)$, are found at the point where the marginal benefit $D(V_0,V) + D(V_0,V)/D_V(V_0,V)$ is

---

[4] To see this, obtain the first-order condition for the investment problem by the discounted net payoff (2) with respect to D instead of V, recognizing that $V = V(D,V_0)$:

$$V + D\frac{dV}{dD} - C = 0.$$

This can be re-written as eqn. (5) above.

equal to the cost, C. Note that $V^* > C$; this is the markup that incorporates the option premium, or value of waiting. If the firm instead used a simple Net Present Value rule to decide when to invest, it would invest sooner, when $V = C$, so its discount factor, denoted by D^N, would be larger. (Note that in Figure 1, the current value of the benefit, V_0, happens to be below the cost of the investment, C, so the firm would not invest immediately even if it followed a simple NPV rule, and $D^N < 1$.)

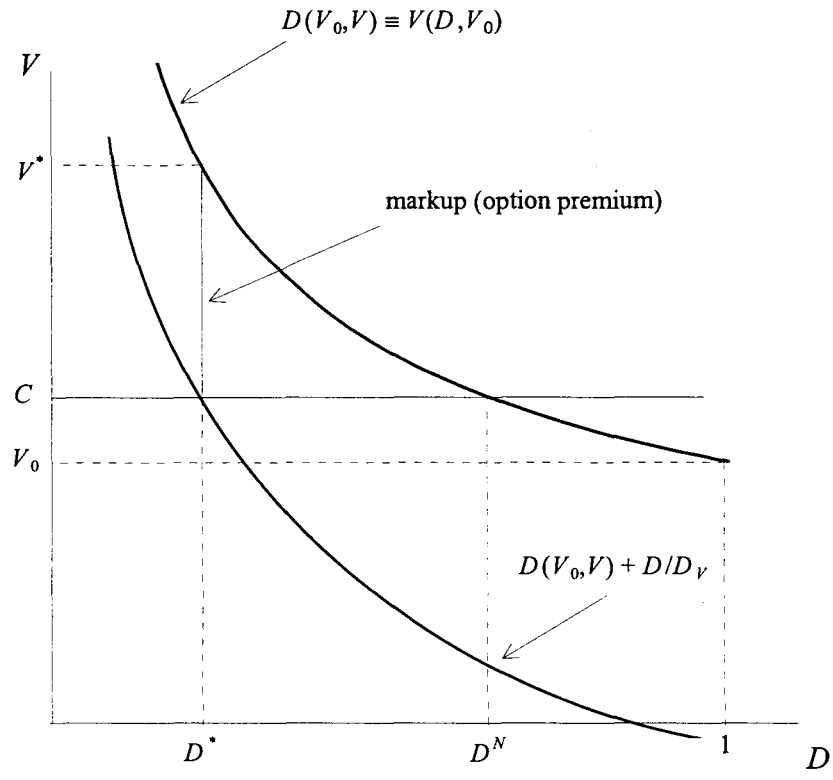

Figure 1. The Optimal Investment Markup.

It remains to sort out one potential difficulty. It would be unfortunate if the elasticity ε_D depended on the initial value V_0, as that would imply that if we reconsidered the choice after som intermediate value V_1 had been reached, we would get a different answer for the optimal V^*. To examine this, consider any three values $V_0 < V_1 < V$. Suppose that, along any path of the process $\{V_t\}$, starting at V_0 the first time the value reaches V_1 is T_1, and starting at V_1 the first time it reaches V^* is T_2. Then the first time the value reaches V^* starting at V_0 is just $T = T_1 + T_2$. (In the second interval of time T_2 we have already supposed that the process does

not reach V^*, and in the first interval of time T_1 the process could not have reached V^* without hitting V_1 earlier, which would contradict our definition of T_1 as the first time to V_1.) Now

$$e^{-\rho T} = e^{-\rho T_1} e^{-\rho T_2},$$

and because of the Markov property of the process $\{V_t\}$, the random variables T_1 and T_2 are independent. Therefore we can take expectations of the two factors on the right-hand side to get

$$D(V_0, V) = D(V_0, V_1) D(V_1, V).$$

Then

$$D_V(V_0, V) = D(V_0, V_1) D_V(V_1, V)$$

and

$$(6) \qquad \frac{V D_V(V_0, V)}{D(V_0, V)} = \frac{V D_V(V_1, V)}{D(V_1, V)}.$$

A similar argument can be constructed for $V_2 < V_0$, by considering paths where the process starting at V_0 first falls to V_2 before rising again and eventually reaching V^*.

This proves that the elasticity is independent of the starting value. In particular, using eqn. (6) we can write the elasticity as

$$(7) \qquad \varepsilon_D = -\frac{V^* D_V(V^*, V^*)}{D(V^*, V^*)} = -V^* D_V(V^*, V^*)$$

since $D(V^*, V^*) = 1$. Hence the optimal markup rule given by eqn. (4) is independent of the starting value V_0. This can also be seen in Figure 1; although the discount factors D^* and D^N depend on V_0, the optimal markup $V^* - C$ does not.

Finally, note that the elasticity of the discount factor, ε_D, can be equivalently expressed in terms of the value of the firm's option to invest. Let $F(V)$ denote the value of the firm's

investment option. At the optimal exercise point, $F(V^*)$ must satisfy the value matching condition

$$F(V^*) = V^* - C,$$

and the smooth pasting condition,

$$F_V(V^*) = 1.$$

Combining these two conditions, we have:

$$(8) \qquad \frac{V^*}{V^* - C} = \frac{V^* F_V(V^*)}{F(V^*)} \equiv \varepsilon_F .$$

The right-hand side of (8), denoted by ε_F, is the elasticity of the value of the investment option with respect to the value of the underlying project. Since $V^*/(V^* - C) = \varepsilon_D$, at an optimum the elasticity of the discount factor coincides with the elasticity of the value of the investment option.

3. Examples

To use this approach to finding optimal investment rules, one must find the discount factor D, given the stochastic process for V_t. This can be done as follows.

Suppose that V_t follows a general Ito process of the form

$$(9) \qquad dV = f(V)dt + g(V)dz .$$

We want to find $D(V,V^*) = E[e^{-\rho T}]$, where T is the hitting time to V^*, starting at $V < V^*$. Over a small time interval dt, V will change by a small random amount, dV. Hence (suppressing V^* for simplicity):

$$D(V) = e^{-\rho dt} E\big[D(V + dV)\big].$$

Expanding $D(V+dV)$ using Ito's Lemma, noting that $e^{-\rho dt} = 1-\rho dt$ for small dt, and substituting eqn. (9) for dV, we obtain the following differential equation for the discount factor:

(10) $\frac{1}{2}g^2(V)D_{VV} + f(V)D_V - \rho D = 0.$

This equation must be solved subject to the two boundary conditions: (1) $D(V^*,V^*) = 1$, and (2) $D(V,V^*) \to 0$ as $V^* - V$ becomes large.

To illustrate, we will obtain solutions using this approach for several different stochastic processes, and draw further analogies to the profit-maximizing decisions of a price-setting firm.

Geometric Brownian motion

First, suppose that V_t follows the geometric Brownian motion:

(11) $dV = \mu V dt + \sigma V dz,$

with $\mu < \rho$.[*] Then $f(V) = \mu V$ and $g(V) = \sigma V$, and it is easily seen that the solution to eqn. (10) is

(12) $D(V_0,V) = (V_0/V)^{\beta_1},$

where β_1 is the positive root (exceeding unity) of the following quadratic equation in β:

(13) $\frac{1}{2}\sigma^2\beta(\beta-1) + \mu\beta - \rho = 0;$

see Dixit and Pindyck (1996, p. 316).

In this case the elasticity of the discount factor is constant and equal to β_1. The markup formula (4) thus implies a constant proportional markup,

$$(V^* - C)/V^* = 1/\beta_1,$$

or

---

[*] The original text denotes the drift of the geometric Brownian motion by the letter α. To avoid confusion with notation in the remaining chapters, I have replaced this with μ. A similar change applies to the arithmetic Brownian motion below.

$$V^* = \frac{\beta_1}{\beta_1 - 1} C.$$

This well-known result is analogous to the price-cost markup rule for a firm facing an isoelastic demand curve. A geometric Brownian motion for V_t implies an isoelastic discount factor because the probability distribution for percentage changes in V is independent of V, so changes in the discount factor resulting from a percentage change in V will also be independent of V.

Arithmetic Brownian motion

Next, suppose V_t follows the arithmetic Brownian motion

(14) $dV = \mu dt + \sigma dz$.

Then the solution to eqn. (10) is

$$D(V_0, V) = \exp[-\gamma_1 (V - V_0)],$$

where γ_1 is the positive root of the quadratic

$$\tfrac{1}{2}\sigma^2\gamma^2 + \mu\gamma - \rho = 0;$$

see Harrison (1985, p. 42). In this case, the elasticity of the discount factor is $\gamma_1 V$. Hence $(V^* - C)/V^* = 1/\gamma_1 V^*$, and we get a constant additive markup:

$$V^* = C + (1/\gamma_1).$$

This is analogous to the markup formula for a firm facing an exponential demand curve. (For the demand curve $q(p) = a\exp[-bp]$, the elasticity of demand is bp, and the profit-maximizing price is $p^* = c + 1/b$.)

Mean-reverting process

Finally, suppose that V_t follows the mean-reverting process:

(15) $dV = \eta(\overline{V} - V)Vdt + \sigma V dz$.

The value, V, might follow such a process if, for example, the firm's output was a durable good so that its demand was subject to a stock-adjustment process.

Then eqn. (10) for $D(V,V_1)$ becomes:

$$\tfrac{1}{2}\sigma^2 V^2 D_{VV}(V,V_1) + \eta(\bar{V} - V)VD_V(V,V_1) - \rho D(V,V_1) = 0.$$

This equation has the following solution (see Dixit and Pindyck, pp. 162-163):

$$D(V,V_1) = AV^{\theta} H(\frac{2\eta}{\sigma^2}V,\theta,b),$$

where A depends on V_1, θ is the positive solution to the quadratic equation

$$\tfrac{1}{2}\sigma^2\theta(\theta-1) + \eta\bar{V}\theta - \rho = 0,$$

and

$$b = 2\left(\theta + \frac{\eta\bar{V}}{\sigma^2}\right).$$

Here $H(x,\theta,b)$ is the confluent hypergeometric function, which has the following series representation:

$$H(x,\theta,b) = 1 + \frac{\theta}{b}x + \frac{\theta(\theta+1)x^2}{b(b+1)2!} + \frac{\theta(\theta+1)(\theta+2)x^3}{b(b+1)(b+2)3!} + \ldots.$$

The first boundary condition is used to determine A, yielding:

$$A = \frac{1}{V_1^{\theta} H\left(\frac{2\eta}{\sigma^2}V_1,\theta,b\right)}.$$

Hence the discount factor becomes:

$$(16) \qquad D(V_0,V) = \left(\frac{V_0}{V}\right)^{\theta} \frac{H\left(\frac{2\eta}{\sigma^2}V_0,\theta,b\right)}{H\left(\frac{2\eta}{\sigma^2}V,\theta,b\right)}.$$

From the series representation, we obtain the following relationship between H and its derivative with respect to the first argument:

$$H_x(x,\theta,b) = \frac{\theta}{b} H(x,\theta+1,b+1).$$

Using this, we can determine that the elasticity of the discount factor is:

$$(17) \qquad \varepsilon_D = \theta \left[1 + \frac{2\eta V}{\sigma^2 b} \frac{H\left(\frac{2\eta}{\sigma^2}V,\theta+1,b+1\right)}{H\left(\frac{2\eta}{\sigma^2}V,\theta,b\right)} \right].$$

Thus, ε_D is equal to a constant θ - which represents pure geometric growth - plus a term which corrects for the mean reversion effect. As the mean reversion speed η approaches zero, the second term also goes to zero, and θ approaches β_1, as in the case of geometric Brownian motion. As η increases, mean reversion dominates.

The implications of mean reversion are easiest to see from some numerical calculations. Mean reversion implies that V is expected to stay close to $\overline{V}$. Hence when $V - \overline{V}$ is small, the discount factor must be larger for the mean-reverting process than for the corresponding geometric Brownian motion. Likewise if $V - \overline{V}$ is large, it can be expected to decline, so that the discount factor will be relatively small. Figure 2 illustrates this; it shows the discount factor as a function of V for a mean-reverting process (η = 0.2) and a geometric Brownian motion (η = 0). (In both cases, ρ = 0.05, σ = 0.2, $\overline{V}$ = 1, and V_0 = 1.) This effect of mean reversion is also reflected in the elasticity of the discount factor, which is increasing in V. For example, $\varepsilon_D(V = 1) = 1.4$, and $\varepsilon_D(V = 2) = 8.54$; while the corresponding constant elasticity for the geometric Brownian motion (η = 0) is $\beta_1 = 2.16$. Figure 3 shows how the elasticity depends on the speed of mean reversion, η. When $V - \overline{V}$ is small (V = 1.0), ε_D decreases with η, but when it is large (V = 2.0), it increases with η.

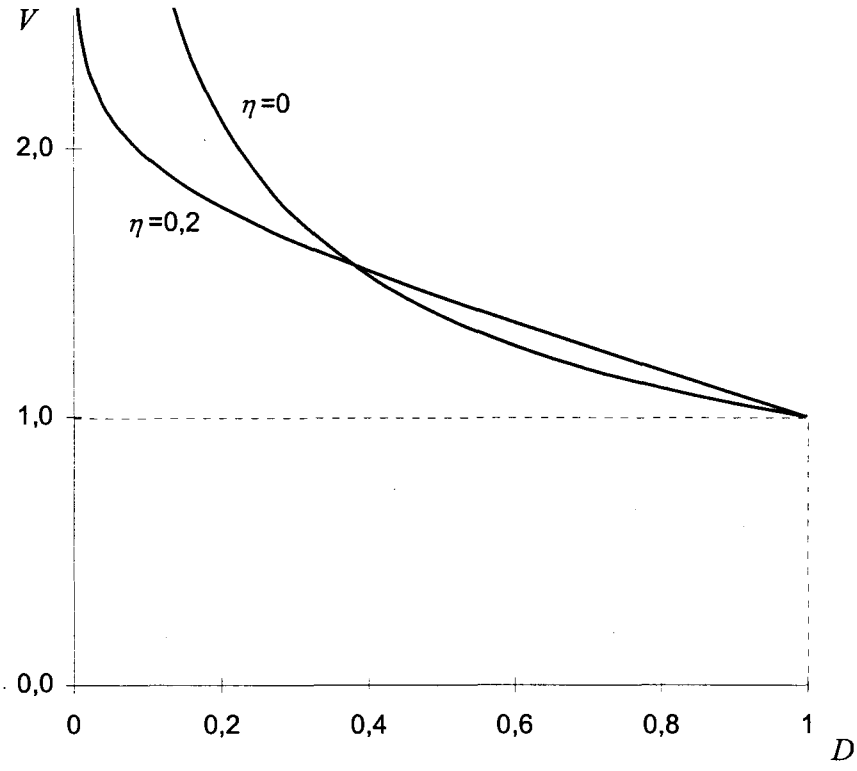

Figure 2. Discount Factor for Mean-Reverting Process and Geometric Brownian Motion ($\rho = 0.05$, $\sigma = 0.2$, $\overline{V} = 1$, $V_0 = 1$).

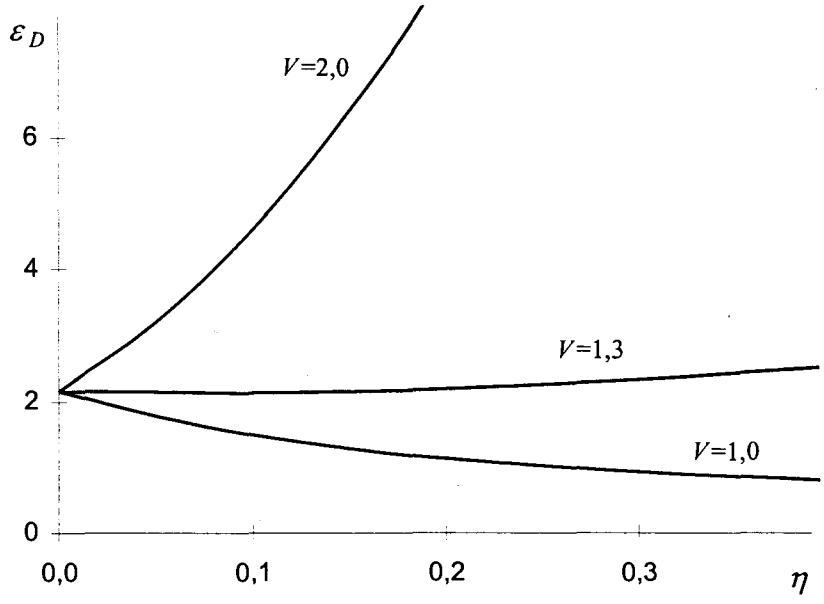

Figure 3. Elasticity of Discount Factor as a Function of the Speed of Mean Reversion.

4. Ancillary investments in advertising or R&D

The close connection between investment decisions and pricing decisions has pedagogical value, but also provides insight into investment-related decisions more broadly. As an example, consider a price-setting firm that must also decide how much money, A, to spend on advertising, given its demand $q = q(p,A)$, with $\partial q / \partial A > 0$. As students are taught in intermediate microeconomic courses, the profit-maximizing advertising-to-sales ratio is given by:

$$(18) \qquad \frac{A}{pq} = \varepsilon_A / \varepsilon_P ,$$

where $\varepsilon_A = (A/q)\partial q / \partial A$ is the firm's advertising elasticity of demand, and ε_P is the price elasticity of demand.[5]

Now let us return to our investment problem. Suppose that the firm, prior to making the sunk expenditure C in return for the benefit V, can make an ancillary investment, costing A, in advertising, marketing, or R&D activities. The exact nature of these activities is unimportant; what matters is that they lead to more rapid increases in V, and hence to an increase in the discount factor $D(V_0,V)$. We can then re-state our investment problem as:

$$(19) \qquad \max_{V,A}\left[(V-C)D(V_0,V,A) - A\right].$$

The two first-order conditions for this problem are

$$(20) \qquad D(V_0,V,A) + (V-C)D_V(V_0,V,A) = 0,$$

and

$$(21) \qquad (V-C)D_A(V_0,V,A) - 1 = 0.$$

[5] Eqn. (18) follows from maximizing profit with respect to p and A, and is sometimes referred to as the Dorfman-Steiner (1954) theorem.

Now define the elasticities of the discount factor with respect to V and A, respectively, as $\varepsilon_V^D \equiv -VD_V / D$ and $\varepsilon_A^D \equiv AD_A / D$. Then by combining the first-order conditions (20) and (21), it is easy to see that

$$(22) \qquad \frac{A}{DV} = \varepsilon_A^D / \varepsilon_V^D .$$

Eqn. (22) is a condition for the optimal ratio of expenditures on advertising (or marketing, or R&D) to the expected discounted value of the benefit. (Remember that the *actual* discounted value of the benefit is unknown because the time until V reaches the threshold V^* is stochastic; DV is the *expected* discounted value of the threshold V^*.) It is exactly analogous to condition (18) for the advertising-to-sales ratio of a price-setting firm.

As an example, suppose that a pharmaceutical firm is deciding whether to invest in a plant to produce a new drug. Suppose the benefit from this investment, V_t, follows the geometric Brownian motion of eqn. (11), and will grow over time (at the expected rate μ) even before the plant is built as doctors and patients learn about the drug. However, the expected growth rate μ can be increased via expenditure A on advertising and marketing.[6]

To determine the optimal level of A for this example, note that the discount factor is again given by eqn. (12), with β_1 again the solution to the quadratic eqn. (13). Hence the elasticity ε_V^D is again equal to β_1. But now β_1 is a function of A, since μ depends on A. Differentiating the quadratic eqn. (13) with respect to A and rearranging yields the following expression for the elasticity ε_A^D:

$$(23) \qquad \varepsilon_A^D = -\frac{A \cdot (d\mu / dA) \cdot \log D}{\sigma^2 \beta_1 + \mu - \frac{1}{2}\sigma^2} .$$

Defining the elasticity $\varepsilon_A^\mu \equiv (A / \mu) d\mu / dA$, the optimal ratio of A to the discounted benefit is thus given by:

---

[6] We treat A as a lump-sum expenditure. If the advertising and marketing expenses must be spread out over time, then A is just the present value of those expenses.

$$(24) \qquad \frac{A}{DV} = \frac{\mu \varepsilon_A^{\mu} \log(V^* / V_0)}{\sigma^2 \beta_1 + \mu - \frac{1}{2}\sigma^2}.$$

This ratio will be larger the larger ε_A^{μ} - the more productive is advertising and marketing, the more that should be done. But note that this ratio is also larger the larger is the threshold V^*. A larger V^* implies that the option to invest is more valuable (the expected net payoff (V^* - C) is larger), which increases the expected return from advertising and marketing expenditures. Hence this ratio is larger if there is greater uncertainty over the evolution of V; an increase in σ increases V^*, and (with some algebra) can be shown to reduce the denominator of (24). Finally, note that $A \to 0$ as $V^*/V_0 \to 1$; when $V^* = V_0$ there is no option premium, and thus no benefit to increasing μ.

5. Conclusions

Framing the optimal investment decision as a trade-off between larger versus later net benefits has allowed us to interpret the investment rule as a simple markup formula involving an elasticity. We have seen that the markup is exactly analogous to a firm's optimal markup of price over marginal cost. For economists, this may be more intuitively appealing than the standard approach to irreversible investment problems in which one values the firm's option to invest and finds the optimal exercise rule.

If the benefit, V, follows a geometric Brownian motion - as is typically assumed in applications - then the markup formula is particularly simple, since the elasticity of the discount factor is constant and equal to β_1, the solution to the fundamental quadratic equation (13). In this case the discount factor is isoelastic with respect to V, so the investment problem is analogous to the pricing problem for a firm facing an isoelastic demand curve.

Even if V does not follow a geometric Brownian motion, this markup formulation provides a rule of thumb that can be of value to practitioners. Compared to equating marginal cost with marginal revenue, it can be easier for a manager to think about pricing in terms of a markup based on the elasticity of demand, estimates of which can be based on statistics or on judgment. Likewise, it can be easier to think about investment hurdles as a markup based on the elasticity of the discount factor, "estimates" of which can be found analytically or judgmentally.

References

Baily, Walter Toshi, 1995, *Essays in Finance*, unpublished Ph.D dissertation, M.I.T.

Dixit, Avinash and Robert Pindyck, 1996, *Investment Under Uncertainty*, Second printing, NJ: Princeton University Press

Dorfman, Robert and Peter O. Steiner, 1954, "Optimal Advertising and Optimal Quality", *American Economic Review* 44, 826-836

Harrison, J. Michael, 1985, *Brownian Motion and Stochastic Flow Systems*, New York: John Wiley

McDonald, Robert and Daniel Siegel, 1986, "The Value of Waiting to Invest", *Quarterly Journal of Economics* 101, 707-728

CHAPTER 1 - Note A

Linear Demand

The analogy between expected discount factors and demand functions raises a question that may be of some interest: Is there a set of stochastic processes yielding a discount factor which is analogous to the perhaps most common demand function in economic theory - linear demand? Yes, indeed, there is, but instead of just presenting the processes, I will also show how to find them. This will illustrate an approach that may be useful when searching for the analogous dynamic representation of other static demand functions as well.

Start by considering the deterministic case. Then we must have $D(V_0,V)=e^{-\rho T}$, where T is the deterministic time from V_0 to V. To find the process $V(T)$ we can calculate backwards: From the definition of the elasticity we can determine the discount factor, and from the discount factor the process is easily found. The initial value V_0 arises as a constant of integration. For a linear demand curve, the elasticity is

$$(1) \qquad \varepsilon_D = \frac{V}{\phi - V},$$

where ϕ is a positive constant. Using the definition of the elasticity, this can be rewritten as:

$$(2) \qquad \frac{dV}{V-\phi} = \frac{dD}{D}.$$

By integration, and using the first boundary condition (which says that the discount factor is unity when its two arguments coincide), we find the following linear discount factor:

$$(3) \qquad D(V_0,V) = \frac{\phi - V}{\phi - V_0}.$$

This discount factor is plotted as a downward sloping demand function in Figure 1.

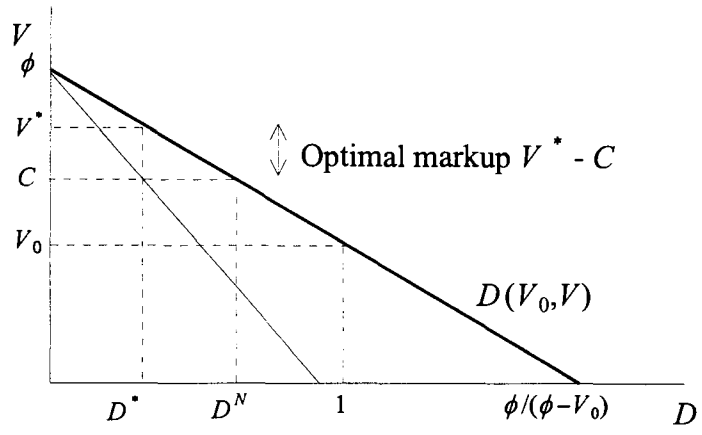

Figure 1. A Linear Discount Factor.

The inner curve represents the dynamic analogy of a static marginal revenue function. The optimal rule is to invest when the marginal revenue equals the marginal cost, with a corresponding optimal price V^* and a markup $V^* - C$. Here C is the constant investment cost, and

$$(4) \qquad V^* = \frac{\phi + C}{2}.$$

No solution exists unless $C < \phi$. The process is

$$(5) \qquad V(T) = \phi - (\phi - V_0)e^{-\rho T},$$

which is an increasing function in T that approaches ϕ as long as $V_0 < \phi$. Hence, linear demand in a static model corresponds to a dynamic model with a process where the value of the project approaches a constant level asymptotically. The increment is $dV = \rho(\phi - V)dt$, so the obvious candidate for a stochastic process is

$$(6) \qquad dV = \mu(\phi - V)dt + \sigma(\phi - V)dz,$$

where σ (>0) and μ are constants. Notice the character of this process: It has an upper barrier at $V = \phi$. Further, the trend and the volatility are proportional to the distance from the barrier. The geometric Brownian motion has the same properties with respect to a lower barrier at zero. Thus, the process (6) is simply an inverted geometric Brownian motion. By the same approach as in *Chapter 1*, the expected discount factor becomes

(7) $$D(V_0,V)=\left(\frac{\phi-V}{\phi-V_0}\right)^{\psi},$$

where ψ is the positive root of the following characteristic equation in x:

(8) $$\tfrac{1}{2}\sigma^2 x(x+1)+\mu x-\rho=0.$$

The discount factors for various ψ make up a set of curves with fixed endpoints, as shown in Figure 2. If $\sigma=0$ and $\mu=\rho$, the deterministic case with $\psi=1$ is obtained. However, there is an infinite number of parameter combinations with $\sigma>0$ (and $\mu<\rho$) that also have $\psi=1$, so a stochastic process that yields a linear expected discount factor can easily be found from the characteristic equation.

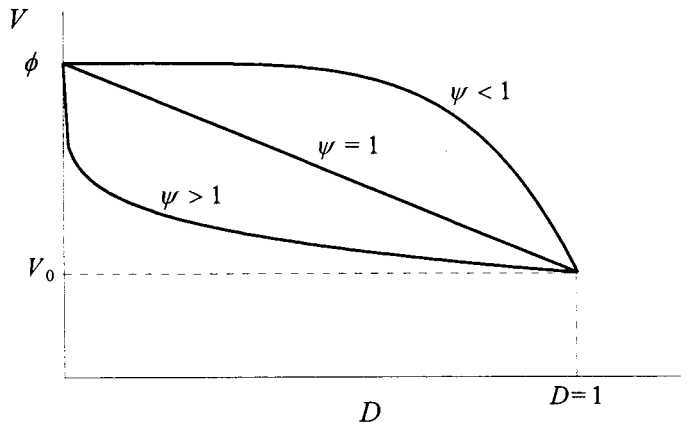

Figure 2. Three Expected Discount Factors.

Furthermore, D is strictly convex if $\psi>1$, and strictly concave if $\psi<1$. Note also that $\psi\to 0$ as $\sigma\to\infty$. Thus, a vertical demand curve ($D=1$) arises in the limit with infinite uncertainty and a sufficiently large upper barrier. Finally, if $\mu=0$, it can be shown that $\psi\to\infty$ as $\sigma\to 0$. This removes all dynamics that could create value from waiting, and we get a horisontal curve $V=V_0$ for any finite time ($D>0$).

CHAPTER 1 - Note B

Fluctuating Costs

1. Introduction

Chapter 1 focused on the analogy between an irreversible investment based on fluctuating prices and fixed costs, and a static pricing decision with constant marginal costs and a downward-sloping demand curve. This note gives a brief introduction to the "dual" case, in which costs are fluctuating and prices are fixed. (The economic sense of this kind of investment problem is discussed more broadly in *Chapter 3*.)

While the investment problem that was studied in *Chapter 1* is akin to a financial call option, the one discussed here is akin to a put option. As we are about to show, there is also a close symmetry between the two investment problems in several other respects. The symmetry probably remains for almost any generalization of the approach.

2. The optimal investment rule

Consider the following investment problem: A firm has the option to obtain a revenue V by investing C, where V is constant while C fluctuates according to a continuous Markov process $\{C_t\}$. Assume that the current level of the process, C_0, is so high that immediate investment is not optimal. In particular, this is the case if C_0 exceeds V. By an argument similar to that applied in *Chapter 1*, the investment should not take place as soon as C gets lower than V, because there is value from waiting to see whether it decreases even further. The optimal rule is to invest the first time the cost reaches a specific level $C^* < V$.

Below, we shall find the optimal cost C^* by first assuming that investment takes place when some arbitrary $C < C_0$ is reached for the first time, and then optimizing expected and discounted profits with respect to C. The discount factor when going from C_0 to $C < C_0$ for the first time can be defined equivalently as a function of C_0 and C:

(1) $$S(C_0, C) \equiv E\left[e^{-\rho T}\right].$$

$S(C_0,C)$ is strictly increasing in C ($\leq C_0$), since the closer C is to C_0, the more quickly will C be reached. In the limit, we have $S(C_0, C_0) = 1$. The expected net present value to be maximized becomes:

(2) $$S(C_0,C)(V - C).$$

From (2) it is observed that the discount factor S is analogous to a quantity measure; in this case in the form of a supply function depending on C, since C_0 is constant. Maximizing (2) with respect to C gives the following first-order condition for optimum:

(3) $$\frac{V - C^*}{C^*} = \frac{1}{\varepsilon_S};$$

where

(4) $$\varepsilon_S \equiv \frac{S_C(C_0, C^*) \cdot C^*}{S(C_0, C^*)}$$

is the elasticity of the discount factor with respect to the investment threshold. Further, S_C is the derivative of S with respect to the second argument. Eqn. (3) is analogous to the optimal decision rule for a monopsonistic firm with an upward sloping supply curve. It is also recognized as the optimal tariff formula from trade theory. (V is analogous to the home market price, C^* to the foreign price, and ε_S to the elasticity of foreign supply.)

The condition (3) may be stated alternatively for an arbitrary C as follows:

(5) $$S(C,C_0) + S(C,C_0) / S_C(C,C_0) = V.$$

The left-hand side of eqn. (5) is analogous to a marginal cost function, and optimum is found where the marginal cost equals the price. This is illustrated in Figure 1.

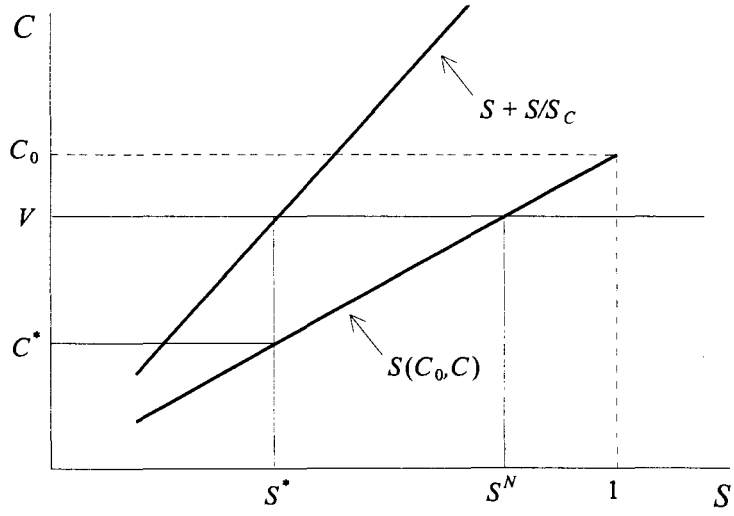

Figure 1. Optimal Investment with Fluctuating Costs.

The optimal discount factor S^* is smaller than S^N, that would apply by use of a simple net present value rule. As in *Chapter 1*, this is due to the option value from waiting. The relationship to the former approach is so close that a couple of other results do not have to be spelled out in detail:

First, it can be shown that the elasticity ε_S is independent of C_0; thus the optimal investment rule is not affected by changes in the initial cost. Second, for a general Ito process $dC = f(C)dt + g(C)dz$, a dynamic programming argument will show that the expected discount factor when going from a general C to a fixed $C^* < C$ for the first time, arises from the following differential equation (leaving out the constant C^*):

$$(6) \qquad \tfrac{1}{2} g^2(C) S_{CC}(C) + f(C) S_C(C) - \rho S(C) = 0 .$$

This is exactly like the equation that determines the discount factor with fluctuating prices. The first boundary condition is also identical, saying that the discount factor must equal unity when the two arguments coincide; i.e., $S(C^*, C^*) = 1$. The second one, which is the one that makes the difference, is that $S(C, C^*)$ must approach zero if $C - C^*$ becomes large.

3. Examples

Geometric Brownian motion

If the cost process is geometric Brownian

(7) $dC = -\mu C dt + \sigma C dz$,

then eqn. (6) is characterized by the following equation in x:

(8) $\frac{1}{2}\sigma^2 x(x-1) - \mu x - \rho = 0$.

Using the two boundary conditions, we find

(9) $S(C_0, C) = \left(\frac{C}{C_0}\right)^{\alpha}$,

where $\alpha > 0$ is the magnitude of the negative root of eqn (8). The cost at which it is optimal to invest becomes a fixed fraction of V:

(10) $C^* = \frac{\alpha}{\alpha+1} V$.

The elasticity of the discount factor is

(11) $\varepsilon_S = \alpha$,

which means that the discount factor for a downward moving geometric Brownian (cost) process is analogous to an isoelastic supply function. As shown in *Chapter 1*, a similar upward moving (price) process yields a discount factor that is analogous to an isoelastic demand function. The relationship becomes even clearer noting that α alternatively can be expressed as the positive root of the following quadratic equation:

(12) $\frac{1}{2}\sigma^2 x(x+1) + \mu x - \rho = 0$.

Except for a sign shift, this is like eqn. (13) in *Chapter 1*. However, it is exactly like eqn. (8) in *Chapter1, Note A*, as the process in that case was an inverted geometric Brownian motion.

Arithmetic Brownian motion

If the cost process is arithmetic Brownian

(13) $dC = -\mu dt + \sigma dz$,

the discount factor becomes

(14) $S(C_0, C) = e^{-\xi(C_0 - C)}$,

where ξ is the positive root of the following quadratic:

(15) $\frac{1}{2}\sigma^2 x^2 + \mu x - \rho = 0$.

Hence, the discount factor with a downward moving arithmetic Brownian cost process is analogous to an exponential supply function (similar to an exponential demand function for an upward moving price process).

The linear case

It is not hard to find a combination of parameters in eqn. (12) for which $\alpha = 1$, and thus a linear discount factor that takes off from origo. A more general representation follows by using the process:

(16) $dC = -\mu(C - \phi) + (C - \phi)dz$.

This is like the geometric Brownian motion (7), except that the barrier has been placed at an arbitrary level ϕ ($< C_0$). The discount factor becomes

(17) $S(C_0, C) = \left(\frac{C - \phi}{C_0 - \phi}\right)^{\alpha}$,

where α is the positive root of eqn. (12). If $\alpha = 1$, this is analogous to a linear supply function intersecting the vertical axis at $C = \phi$ (see Figure 1), and having the elasticity $\varepsilon_S = C/(C - \phi)$.

CHAPTER 2

A Simplified Exposition of Smooth Pasting*

Abstract: The decision on when to make an irreversible investment is considered as a trade-off between the instantaneous size of the net benefit and the time at which it is obtained. The benefit can be larger by waiting longer, but then it will also have to be more discounted. Smooth pasting arises as a first-order condition for maximum expected profit. The relationship to the standard approach is illustrated by a geometric Brownian price process.

Keywords: Irreversible investment, optimization, value matching, smooth pasting

JEL classification: D92, D81, C61

1. Introduction

Consider the basic problem when to invest a constant C to obtain (once) a revenue V that is fluctuating according to a continuous Markov process $\{V_t\}$. McDonald and Siegel (1986) looked upon this investment problem as a perpetual call option, involving a right but no obligation to invest. The optimal rule, which will be constant as long as the discount rate is constant, and the process is characterized by first-order stochastic dominance in the sense that a higher current value shifts the distribution to the right, is to invest the first time V reaches a specific $V^* > C$. The markup from C to V^* reflects the value of the opportunity to wait.

The standard approach to solving irreversible investment problems of this kind consists of calculating option values, determining optimal decisions by the familiar value matching and

---

* Published in *Economics Letters*, 1998, vol. 58, pp. 217-223.

smooth pasting conditions. Denoting the option value by $\overline{F}(C,V)$ and its derivative with respect to the second argument by $\overline{F}_V(C,V)$, the conditions here are

(1) $\overline{F}(C,V^*) = V^* - C$,

and

(2) $\overline{F}_V(C,V^*) = 1$.

The value matching condition (1) reflects an intuitive requirement for continuity at the optimal exercise point V^*. Further, it is well-known that the smooth pasting condition (2) is a first-order condition for optimum, as already proposed by Samuelson (1965). Under weak conditions it can also be shown to be sufficient; see Brekke and Øksendal (1991). However, the general theory underlying such results is rigorous and hardly accessible to many economists. Even simplified results showing the necessity of smooth pasting, as in Dixit and Pindyck (1994:130-132), are based on rather technical arbitrage arguments considering what would happen if $\overline{F}(C,V)$ had a kink at V^*, not leaving much room for basic intuition.

The new approach to irreversible investment proposed by Dixit et al. (1997) can be used to simplify the treatment of smooth pasting. They regard the investment decision as a trade-off between the size of the net benefit V-C and the effect of discounting. If assuming that the current value of the process is some low V_0, and that investment takes place when it reaches an arbitrary $V > V_0$, the expected and discounted profit becomes:

(3) $E\left[e^{-\rho T}\right](V - C)$.

Here E is the expectations operator, ρ is the discount rate, and T is the first hitting time from V_0 to V. Since the process is continuous, the expected discount factor will be strictly increasing in V_0 and decreasing in V, so it can be described equivalently as a function $D = D(V_0, V)$. Thus the expected profit from a decision to invest when the price has increased to V for the first time, can be stated as

(4) $D(V_0,V)(V - C)$,

which is to be maximized with respect to V. The first order condition for optimum becomes

(5) $D_2(V_0,V^*)\cdot V^* + D(V_0,V^*) = D_2(V_0,V^*)\cdot C$,

where D_2 is the derivative of D with respect to the second argument. Alternatively, we have

$$(6) \qquad \frac{V^* - C}{V^*} = \frac{1}{\varepsilon_D},$$

where ε_D is the elasticity of the discount factor with respect to the investment threshold:

$$(7) \qquad \varepsilon_D \equiv -\frac{V^* \cdot D_2(V_0, V^*)}{D(V_0, V^*)}.$$

Eqn. (6) is analogous to the markup pricing rule in a static model with a downward sloping demand curve D depending on a price V, regarding V_0 as a constant. The elasticity of the discount factor is analogous to a price elasticity of demand, V^* to an optimal price, and C to a constant marginal cost. It can be shown that ε_D does not depend on V_0. This ensures that the optimal investment rule is not affected by changes in V_0.

As D is strictly decreasing in V, the inverse function $V = V(V_0, D)$ could alternatively be used to maximize (4) with respect to D. That would yield the following revised version of eqn. (5) for an arbitrary $V > V_0$:

$$(8) \qquad V + D(V_0, V) / D_2(V_0, V) = C.$$

The left-hand side of eqn. (8) is analogous to a static marginal revenue function. Optimum V is found by setting the marginal revenue equal to the investment cost.

2. Smooth pasting

Let us reconsider the investment problem above, denoting the net benefit from investing at a general $V > V_0$ by:

$$(9) \qquad F = V - C.$$

Since C is constant, the process for the benefit $\{F_t\}$ will share basic properties with $\{V_t\}$; and the optimal rule will be to exercise the option when a specific $F^* > 0$ is reached for the first time. The expected profit can be expressed as a function

$$(10) \qquad \Phi(F_0, F) = Q(F_0, F) \cdot F,$$

where Q is the expected discount factor similar to D of the previous section. Thus Φ is the expected and discounted profit from exercising the option when the benefit has increased to $F > F_0$, instead of doing it right away and obtaining F_0. As the benefit F obtained by waiting arises in the future, it must be discounted by the appropriate factor $Q(F_0,F)$. The first order condition for maximum can be stated as

$$(11) \quad \varepsilon_Q = 1,$$

where

$$(12) \quad \varepsilon_Q \equiv -\frac{Q_2(F_0,F^*)\cdot F^*}{Q(F_0,F^*)}.$$

Eqn. (11) establishes the smooth pasting condition for this problem. It simply says that in optimum, the marginal cost of discounting equals the marginal net benefit from further waiting. As ε_Q is independent of F_0, optimum arises at a unique F^*. Further, the initial slope of Φ when evaluated as a function of F, becomes:

$$(13) \quad \Phi_2(F_0,F_0) = Q_2(F_0,F_0)\cdot F_0 + 1.$$

Since $Q_2 < 0$, $\Phi_2(F_0,F_0)$ is larger or smaller than one, depending on whether F_0 is smaller or larger than zero. The reason is that if $F_0 < 0$, the marginal effect of waiting will be to discount a loss, while if $F_0 > 0$ it will be to discount a benefit. It is also observed that if $F_0 \leq 0$, the curve passes through origo. Using this information, Figure 1 plots Φ as a function of F for four initial values of the benefit, F_0(i) ... F_0(iv).

All curves start from the 45 degree line as $\Phi(F_0,F_0) = F_0$, and they all obtain maximum at the unique $F^* > 0$ at which it is optimal to exercise the option. However, the maximum value is increasing in F_0, because it takes longer to reach F^* the smaller the initial value.

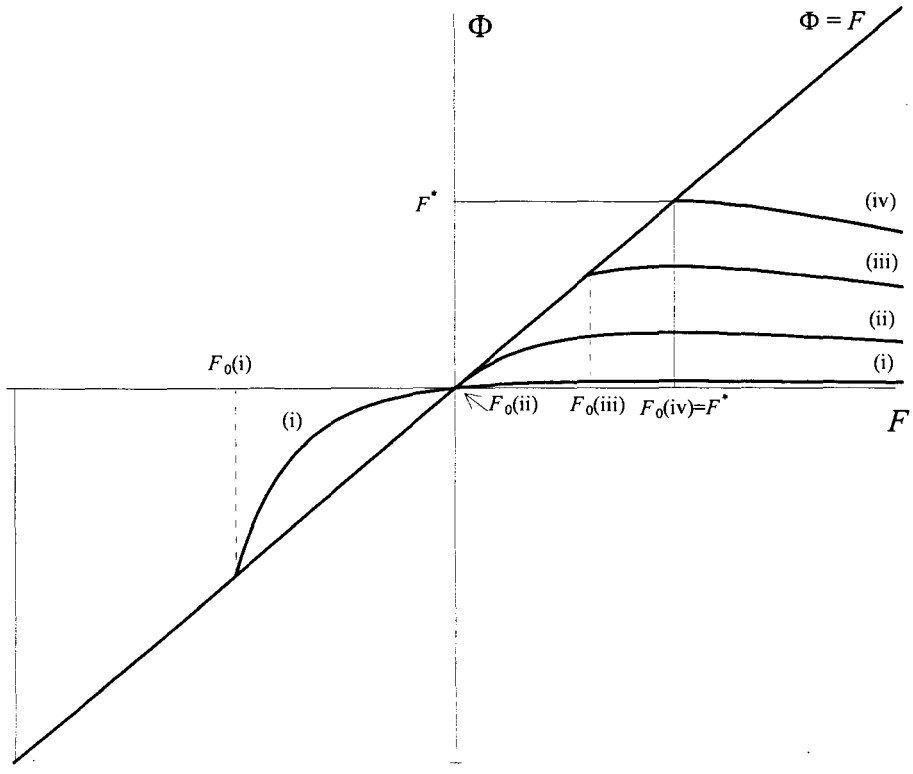

Figure 1. Expected Profit Functions.

Curve (i) starts from a negative benefit F_0(i). Initially, the slope is larger than unity as the marginal effect of discounting is positive. However, the slope decreases to zero as F approaches F^*. Curve (ii) takes off from origo, as the initial benefit is zero when F_0(ii) = 0. Since the marginal effect of discounting is also zero, the curve is tangential to $\Phi = F$ at the initial point. Curve (iii) assumes a positive starting value smaller than F^*. The initial slope is positive as there is a value from waiting, but less than unity due to discounting. Curve (iv) assumes an initial value F_0(iv) that is equal to F^*. That is right where the marginal value of waiting is zero, so the initial slope is zero. (The leftmost vertical line in Figure 1 is explained below.)

The standard forms of value matching and smooth pasting follow by assuming that the option is optimally exercised. Thus for a general initial $F < F^*$ with a corresponding $V < V^*$, the option value function is given by:

(14) $\overline{F}(C,V) \equiv \Phi(F,F^*)$.

By letting $F = F^*$, this gives the standard value matching condition (1) directly, as $Q(F^*,F^*) = 1$. Using the fact that $Q_1(F^*,F^*) = -Q_2(F^*,F^*)$, the standard smooth pasting condition (2) is also easily obtained from eqn. (11).

3. Relationship between the two approaches

The approaches to irreversible investment that have been discussed, are related by two elasticities that coincide in optimum. To see this, define the elasticity of the option value as

$$(15) \qquad \varepsilon_F \equiv \frac{V^* \cdot \bar{F}_V(C,V^*)}{\bar{F}(C,V^*)}.$$

By combining eqns. (1), (2), and (6), we have:

$$(16) \qquad \varepsilon_D = \varepsilon_F .$$

Figure 2 illustrates the relationship. On the left-hand side, $\bar{F}(C,V)$ is plotted as an increasing function of V. At the optimal V^*, the option value function hits the "profit line" V-C tangentially, according to the value matching and smooth pasting conditions. To the far right, $D(V_0,V)$ is plotted as a downward sloping demand function. The steep curve closer to origo represents a part of the marginal revenue function given by the left-hand side of eqn. (8). The optimal discount factor D^* is found where the marginal revenue equals the investment cost, with a corresponding threshold V^*. Observe that D^* is smaller than D^N, that would apply by a simple net present value rule.

The profit line V-C connects the two approaches. If the investment cost is increased, the line shifts vertically upwards. On the right-hand side, V^* increases via the marginal revenue function. On the left-hand side, the entire option value function shifts down (i.e. closer to the vertical axis), hitting the new profit line tangentially for a higher V^*.

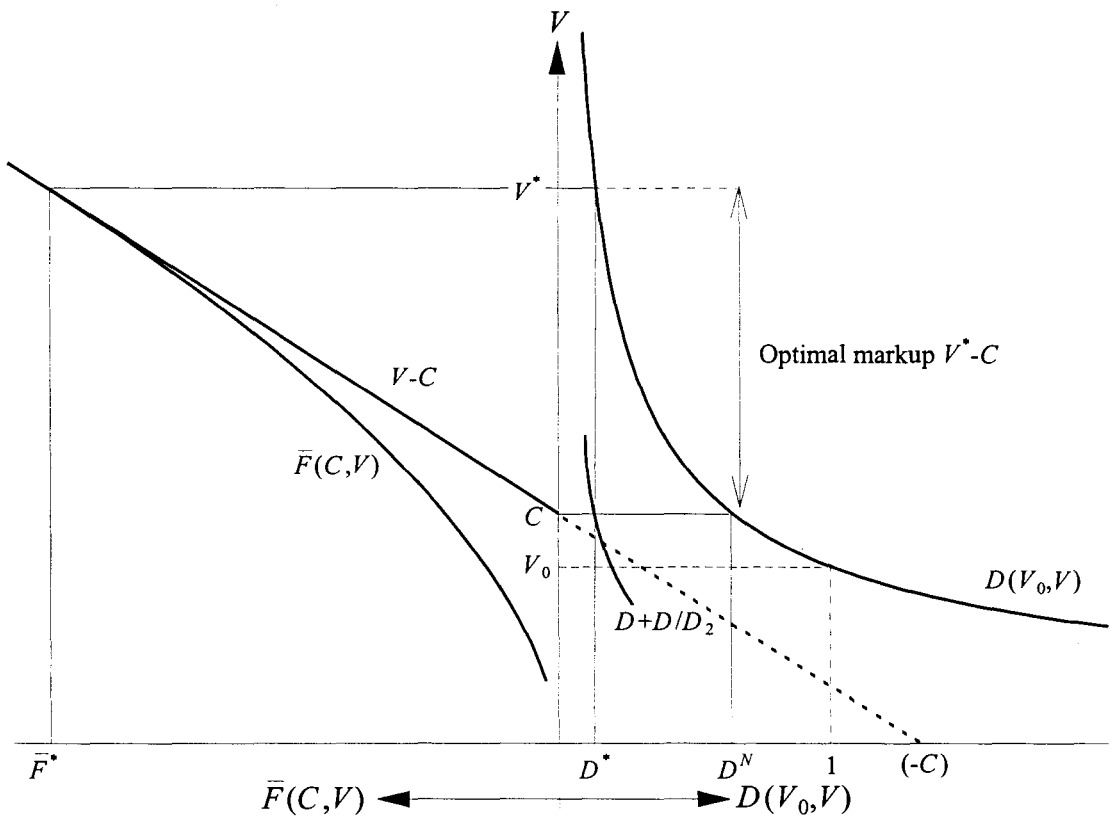

Figure 2. Optimal Investment Rules.

4. Example - a geometric Brownian price process

Assume a constant investment cost C and a geometric Brownian price

(17) $dV = \mu V dt + \sigma V dz$,

where μ is the trend and σ is the volatility. By Ito's lemma, the process for the benefit is given by

(18) $dF = \mu(F + C)dt + \sigma(F + C)dz$.

Following the approach described by Dixit et al. (1997) to finding expected discount factors, we have

(19) $Q(F_0, F) = \left(\frac{F_0 + C}{F + C}\right)^{\beta}$,

where β is the positive root of the following quadratic equation in x:

(20) $\frac{1}{2}\sigma^2 x(x-1)+\mu x-\rho=0$.

The elasticity of the discount factor becomes

(21) $\varepsilon_Q=\beta\frac{F}{F+C}$,

and setting this equal to unity according to eqn. (11), we have:

(22) $F^*=\frac{C}{\beta-1}$.

By inserting $F^* = V^* - C$, we obtain the familiar expression of the optimal investment rule for this problem:

(23) $V^*=\frac{\beta}{\beta-1}C$.

For convergence, we need $\beta > 1$, implying $\mu < \rho$. The option value follows by inserting optimal values into eqn. (14), yielding:

(24) $\bar{F}(C,V)=\frac{(\beta-1)^{\beta-1}}{\beta^\beta C^{\beta-1}}V^\beta$.

Thus the option value is an upward sloping convex function in V starting from origo. As the identity $Q(F_0,F) \equiv D(V_0,V)$ must hold, we also have:

(25) $D(V_0,V)=\left(\frac{V_0}{V}\right)^\beta$.

Hence the geometric Brownian price process corresponds to an isoelastic demand function with elasticity β. Eqn. (16) is also easily verified. In fact, Figures 1 and 2 correspond to a set of numbers for this process, although their general shape applies in a wider context as well. In Figure 1, the leftmost vertical line corresponds to $V_0 = 0$, implying $F_0 = -C$. If F_0(iv) were moved closer to $-C$, the initial slope would approach infinity, since $dV = 0$ at $V = 0$. For the same reason, Φ would go to zero at $F = F^*$ in that case.

5. Final remarks

The smooth pasting condition has been derived by optimization, considering an irreversible investment as a trade-off between the size of the net benefit by investing now, and the effect of discounting by waiting further. Smooth pasting turned out as the first-order condition that must hold to ensure maximum expected and discounted profit. Finally, it should be noted that the interpretation of the expected discount factor as a dynamic measure of quantity can be generalized beyond the level of a demand function. For example, it is straightforward to apply a similar approach to the related investment problem in which V is constant and C is fluctuating. If the cost process has the same properties as those assumed for V in the previous sections, then the discount factor will be analogous to a supply function.

Acknowledgements

I wish to thank Steen Koekebakker and Bernt Øksendal for helpful comments.

References

Brekke, K. A. and B. Øksendal, 1991, "The High Contact Principle as a Sufficiency Condition for Optimal Stopping", in: D. Lund and B. Øksendal, eds., *Stochastic Models and Option Values*, North-Holland, 187-208

Dixit, A. K. and R. S. Pindyck, 1994, *Investment Under Uncertainty*, Princeton University Press

Dixit, A., R. S. Pindyck and S. Sødal, 1997, "A Markup Interpretation of Optimal Rules for Irreversible Investment", *NBER Working Paper* 5971

McDonald, R. and D. Siegel, 1986, "The Value of Waiting to Invest", *Quarterly Journal of Economics* 101, 707-728.

Samuelson, P. A., 1965, "Rational Theory of Warrant Pricing", *Industrial Management Review* 6, 13-31

PART II

Equilibrium Models with Homogeneous Goods

CHAPTER 3

A Framework for Equilibrium Modelling

1. Introduction

The objective of the current chapter is to extend the approach to irreversible investment that has been discussed for the firm-level, to the most simple type of equilibrium models. In all models it is assumed that each firm manufactures a specific product, and that the size of the firm is fixed. The underlying assumption is that there exists a minimum firm size, but as long as this minimum is exceeded, the exact size does not matter. However, we require that each firm is small relative to the size of the market, and that the firms in the industry are symmetric by some measures to be defined.

The construction of models will benefit from the analogy from static to dynamic modelling that was demonstrated in *Chapter 1* and used in *Chapter 2*. The analogy is based on the following observation: A discount factor in a dynamic model with fixed-sized firms is sometimes analogous to a quantity measure in a static model with variable firm size.

Consider a firm with the option to obtain a net benefit (P - C) at some future date, which the firm is free to choose. Here P can be regarded as a unit price (or a compound project value V as in *Chapter 1*), and C as an investment cost. The net benefit fluctuates according to a continuous, autonomous Markov process. If ρ is a constant discount rate and T is the (stochastic) delay of the investment, then the firm's expected and discounted profit is $Q \cdot (P - C)$, where (P - C) is the net benefit at time of investment, and Q is the expected discount factor:

$$(1) \qquad Q \equiv E\left[e^{-\rho T}\right].$$

As long as there is any probability that the net benefit will exceed zero, there are expected profits from holding this option. The essence of the chapter can be summarized as follows:

We build a set of equilibrium models in which a large number of firms are acquiring, holding and exercising such options each period. Equilibrium will follow from a free-entry condition, where expected profits are zero if an optimal investment policy is followed by all firms.

Each option is acquired by an investment A that is similar to the ancillary investment in one of the examples in *Chapter 1*. By investing A, a firm-specific process for the benefit will be triggered from a fixed initial level, e.g. $(\overline{P} - C)$, where $\overline{P}$ is the initial value of a price process, and C is constant. The "dual" set of models from *Chapter 1, Note B*, where the price is fixed while the cost is fluctuating, will also be expanded to equilibrium.

As also discussed in *Chapter 1*, the first investment (A) may be related to R&D, advertising or a similar ancillary activity, whereas the second one (C) is normally related to establishment of a production line, e.g. for a pharmaceutical drug. Our objective is neither to give detailed characterizations of the two types of investments, nor to explain why this kind of investment problems are important in a broader context, but a couple of additional examples may be mentioned just to illustrate the wide range of possible applications.[1]

Cars, aircraft, computers, and other electronic equipment are all products that require an initial investment in some technology. Usually there is uncertainty on the income side at the time of the initial investment, as well as in the intermediate period before a (presumably larger) investment in production capacity. The demand may depend on what exactly the product is going to look like, and on how the taste for the product will develop in the future; in particular, it may be uncertain how the consumers respond to similar products developed by competitors. The expected demand may also increase over time, e.g. for a software product based on a hardware technology that improves, or simply as information about the product is spread. Moreover, the initial investment may be followed by uncertainty with respect to the cost of production, like if the price of intermediates fluctuate, or if learning-by-doing is important, which is often the case in hi-tech industries. With such dynamic effects, there may be gains from waiting.

[1] See Dixit and Pindyck (1994) for a general discussion, and for an equilibrium model in the field (pp. 267-277).

Extraction and development of a number of natural resources are also characterized by gains from waiting. The oil industry, where real options theory is extensively applied, is one example. In this industry, a large exploration cost is usually necessary before an oil field is developed, and optimal timing depends on oil prices as well as the development of technology that is of importance to the cost side. Farming and foresting are also industries where an initial investment and waiting apply before the market is provided with a "mature" product, and usually there is uncertainty involved. Product prices may fluctuate greatly, and there may also be firm-specific cost uncertainty, e.g. as a fish farm is exposed to diseases.

In all these cases, timing is an important part of the optimal decision rule. The examples span over homogeneous as well as differentiated goods, and we shall see that the methodological approach applies in both cases. Hopefully, the approach can also be applied to some cases with combined demand and cost uncertainty, but like in *Chapter 1*, we will restrict to models where all dynamics appear on just one of the two sides.

Part II, which is initiated by this chapter, is restricted to equilibrium models with homogeneous goods, but several of the results also carry over to product differentation, which is studied in *Part III*.

The formal descriptions will start with a structure of preferences, but due to the similarities noted in *Chapter 1*, it need not be described in depth for all cases. The rest of the chapter is structured as follows: Models with price variation (i.e., fluctuating P and constant C), are discussed in Section 2. The dual case with cost variation (i.e., constant P and fluctuating C) is discussed in Section 3, but is not treated as rigorously. This is even more true for the simple two-sector model that is set up in Section 4. A number of extensions are noted in Section 5, and conclusions are drawn in Section 6.

2. One-sector models with price variation

The basic setup in a deterministic case

As in the *Chapter 1* example with ancillary investments, each firm initially makes an irreversible investment A that establishes some kind of production right. We denote it the entry cost or the patent cost. Contrary to *Chapter 1*, this investment is fixed in the baseline

version of the model. It triggers a firm-specific demand process that starts from a fixed level $\overline{X}$ (yielding an initial price similar to V_0 in *Chapter 1*). However, the product cannot be consumed before the patent has been activated. This requires a second investment, C, which is called a cost of activation or production. Due to the demand process the price will fluctuate, and there will be value from postponing the second investment. As soon as activation takes place, the good must be consumed. The following example, which is chosen solely for its (hopefully!) pedagogical value, may serve as an illustration:

Imagine that the only commodity is wine manufactured in fixed-sized barrels. Let one barrel represent one firm. First, a fixed amount of wine is made and filled into a barrel. This is the "patent". Second, the barrel must be brought to the consumer. This is "activation". After activation, the wine must be consumed immediately. All agents discount future costs and benefits at a constant rate ρ, which can be thought of as a subjective rate of time preferences.

The consumers prefer old wine, as it matures at a fixed, positive rate μ ($< \rho$). However, the wine is homogenous in the sense that new and old wine substitute perfectly if adjusting for quality differences.

The symmetry of the problem suggests that all wine that is consumed will be of the same age. Figure 1 illustrates what this looks like in steady-state. The bullets represent entry (wine is made) and the squares represent activation (wine is brought to the consumer).

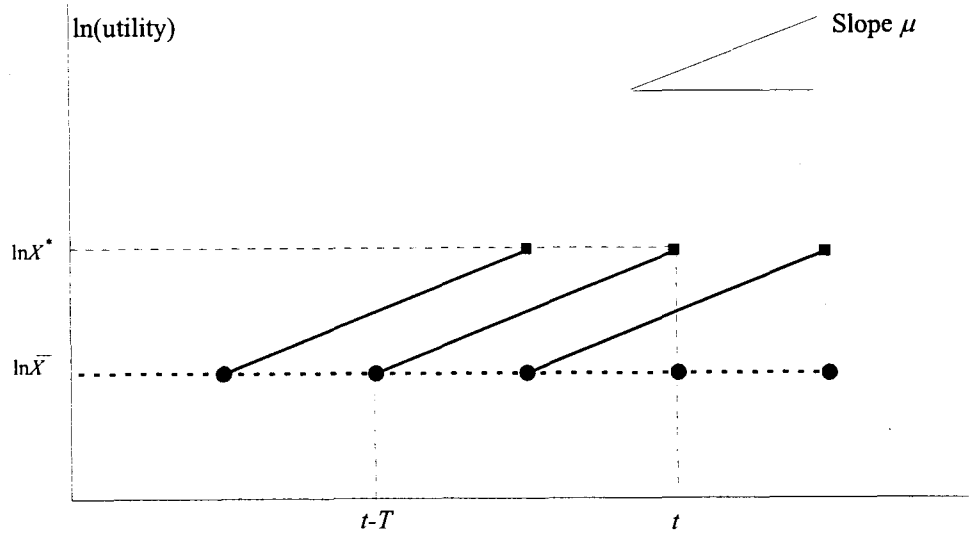

Figure 1. Steady-State with Firm-Specific, Growing Demand.

Each period a large number of barrels are filled, and the maturing processes take off. The utility (in logarithmic scale) that each barrel would yield, depending on how long the process is allowed to go on for, is depicted by the upward sloping lines. Utility from a new barrel equals a constant $\overline{X}$. Wine made at time (t - T) is consumed at time t. Thus, in equilibrium all wine that is consumed has age T, and the utility gain is X^*. The figure assumes $T = 2$. The instantaneous or periodic utility function can be stated in a simplified form as follows:

$$(2) \qquad u_t = N_t \overline{X} e^{\mu T}, \qquad t > T.$$

This says that the utility equals the number of barrels (N_t) times the state of the demand process. In this case, a process starting at $\overline{X}$ equals $X(T) = \overline{X}e^{\mu T}$ after T periods. One very important point must be emphasized here: The utility function measures the gain from those goods that are being consumed at time t, and only then. To find the actual T we must optimize, taking into account the time dimension.

To measure utility over an infinite time horizon, let us for a moment take the perspective of a social planner.[2] The social utility gain from immediate investment is $F_t = N_t(u_t - c)$, where c is the (constant) cost of activation measured in utility terms. Therefore the social planner's utility over an infinite time horizon starting now becomes:

$$(3) \qquad \int_0^\infty F_t e^{-\rho t} dt .$$

This is familiar. Each function F_t satisfies all the requirements that applied to the net benefit F that was discussed in *Chapter 2*, so the problem reduces to a question of marginal valuation. Thus, the decision of when to exercise the options will be a trade-off between larger versus later net benefits. The difference from *Chapter 2* is just that this is done for a (somewhat arbitrary) number of identical projects each period t over an infinite time horizon. For simplicity, we have also restricted to a deterministic case.

---

[2] This can be done with no loss of generality. As all agents share discount rate and have perfect expectations, the market outcome will be socially optimal as long as there are no distortions.

It follows from *Chapter 2*, eqn. (10), that the social planner will maximize the function $\Phi_t = Q_t F_t$; i.e.,

(4) $\Phi_t = N_t(X - c)e^{-\rho T}$.

Here T is a first-hitting time from $\bar{X}$ to some $X^* > \bar{X}$. Since $X = X(T) = \bar{X}e^{\mu T}$ in this deterministic example, all $X > \bar{X}$ will be hit just once until the optimal level X^* is reached. In a market solution, the costs (i.e., the second part of Φ_t) are carried by firms, so we can define a fictitious utility function that arises from the first part. Omitting time subscripts as the optimization is independent of absolute time, this utility function can be written as follows:

(5) $U = \int_0^N X_i e^{-\rho T_i} di$.

Here T_i is the delay of activation for each firm i, $X_i = \bar{X}e^{\mu T_i}$, and N is treated as a continuous variable for technical reasons.[3] This utility function can be maximized with respect to the T_i's, but it is better to use the transformation (1), since the discount factor can embody all the dynamics. (This approach will prove even more powerful in a stochastic case, as uncertainty can be embodied as well.) In this deterministic case, the utility function (5) transforms to

(6) $U = \int_0^N Q_i^{(\beta-1)/\beta}$,

where $Q_i = e^{-\rho T_i}$, $\beta = \rho/\mu$ (> 1), and utility is scaled by setting $\bar{X} = 1$. This looks like a static utility function that is strictly concave in each good. It can be illustrated by considering the effect of delaying the consumption of new wine marginally; i.e., decreasing Q_i slightly from unity. The closer μ is to ρ (the closer β is to one), the more the delay is compensated by improved quality, so the utility decreases more slowly than the discount factor. However, the longer the delay, the more quality improvements will be outdistanced by discounting. Therefore the utility must approach zero if $T_i \to \infty$ (i.e., if $Q_i \to 0$).

[3] Actually, N must be an integer, but this is not important as long as it is large.

With no maturing process, the standard discounting rules would apply, and the utility loss would correspond directly to the reduction in the discount factor. Indeed, this happens if $\mu \to 0$, as $\beta \to \infty$ in that case.

As all agents have rational expectations (and even perfect foresight in this case), any potential gains from waiting will be fully exploited. Firms discount the cost of activation at rate ρ, and since C is fixed, it follows that the cost function for a firm contemplating entry is:

$$(7) \qquad J_i = A + CQ_i .$$

From eqns. (6) and (7) it is observed that we have reached a model somewhat analogous to the standard monopolistic competition model with constant elasticities (β), fixed costs (A), and constant marginal costs (C). The resulting equilibrium conditions might readily be stated, but we leave this for a slightly more general formulation that includes uncertainty.

Growth of productivity

Up until now, ρ has been a rate of time preferences. However, in the rest of this chapter it is mostly regarded as a growth rate, so it is appropriate to explain how such an interpretation can come about. Thus, assume that productivity grows at a constant rate ρ, and that there are no longer any subjective time preferences in the traditional sense.

Productivity growth enables a steadily increasing number of fixed-sized firms (barrels), but growth could also be embodied in the size of each barrel instead of increasing the number. If labor is the only production factor, we might imagine that a fixed number of workers (A) are always used to fill up a new barrel, which is as large as possible. Since productivity grows at rate ρ, the size of the barrels belonging to successive firms will also grow at this rate. One convenient implication of this is that we get a stationary equilibrium, with fixed rates of entry and activation. (In cases of greater economic interest, the natural interpretation is to relate growth to improved quality as new generations of a product are introduced.)

Obviously, growth implies that more will be consumed, but how will the social planner respond? We assume that he is continuously updated, in the sense that he always perceives welfare relative to the current state of productivity (or technology). Hence, he is a "modern"

consumer, and his utility from consumption of a new good at any time equals a constant $\bar{X}$. Notice the difference from Figure 1. With the previous interpretation of ρ there was no growth, and the scale of utility measurement stayed fixed. Now the point of reference slides upward at rate ρ. If a patent is not activated, there will be a utility loss; not because the agents are impatient, but because of depreciation.

Due to this sliding scale, the marginal valuation of the utility gain (u_t) by waiting will effectively be discounted at rate ρ. This will also happen to the production cost. If a constant number of workers (C) can bring a new barrel to the customer immediately, only $Q_t C$ ($\leq C$) workers are needed if activation is postponed for T_i (≥ 0) periods. All this is illustrated in Figure 2.

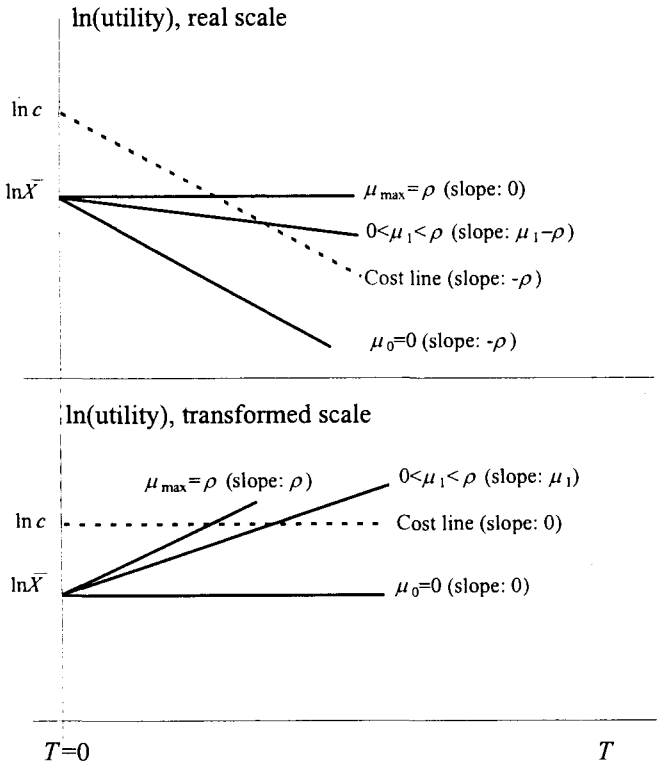

Figure 2. Depreciation in a Model with Updated Preferences.

The upper part shows, for three alternative values of μ, the real perception of utility depending on the age of a patent. The dotted cost line slopes downward, as productivity growth implies that the cost of activation decreases at rate ρ. If the maturing rate is $\mu_0 = 0$, the same happens to perceived utility, and there is no value from waiting. If the maturing rate is $\mu_{max} = \rho$, the

quality improvements compensate fully for growth, and there is no utility loss from waiting. Last year's barrel of wine is smaller than this year's, but the old wine is so much better that the consumer is indifferent between the two. In this case, there is no cost of waiting, and the model diverges. However, for a maturing rate in between these extremes, $0 < \mu_1 < \rho$, old barrels depreciate at a lower rate than the growth rate by which the cost of activation decreases. If that is the case, there is a value as well as a cost from waiting, and the model converges.

The lower part of Figure 2 shows the transformed scale, where all future values are related to the state-of-the-art at $T = 0$. Since the growth rate is ρ, the transformed scale appears by increasing slopes in the upper part by ρ. As observed, this yields a picture like in Figure 1. Hence, with the revised interpretation of the model, the social planner will still evaluate a fluctuating net utility $F_t = N_t(u_t - c)$ marginally when deciding on when to activate a patent. The utility function (6) appears again, but with a new interpretation of ρ.

These two interpretations of ρ, as well as some additional ones that are discussed in Section 5, are independent, and may therefore also be combined.

The general setup

Preferences

The instantaneous utility function can be stated as

$$(8) \qquad u_t = \int_{-\infty}^{t} \int_{0}^{N_s} X_{i,s}(t-s) \cdot c_{i,s}(t)\, di\, ds .$$

This is to be interpreted as follows: At some time s in the past, N_s firms acquired patents that triggered a set of independent demand processes $X_{i,s}$, $i = 1..N_s$. All previous entry times are included by the integration in s. A process starting at time s has age t - s at time t, so the level of process i from period s is $X_{i,s}(t\text{-}s)$ at time t. The terms $c_{i,s}(t)$ represent consumption of each good at time t. Since the utility is linear in each good, they are perfect substitutes in a static sense.

The utility function (8) does not consider two requirements in the model: First, the firm size is fixed, and as we shall let the entire demand be represented by one consumer, we set $c_{i,s} \equiv 1$ for all i,s. Second, each good is available just once, like a barrel of wine that is opened and consumed immediately. Therefore the utility function can be simplified to:

$$(9) \qquad u_t = \int_{-\infty}^{t} \int_{0}^{N_s} \delta_{i,s}(t - s - T_{i,s}) \cdot X_{i,s}(t - s)\, di\, ds .$$

Here $T_{i,s}$ is the age of a patent from period s that is activated in period t. The Dirac functions ($\delta_{i,s}$) ensure that the contribution to utility is zero except for those patents that are activated at time t, as consumption only takes place if $t = s + T_{i,s}$.[4] The rate of entry as well as the rate of activation (and consumption) will be constant in equilibrium from the symmetry and large-group assumptions. Further, the outer integral vanishes due to the Dirac functions, so the utility function simplifies to:

$$(10) \qquad u_t = \int_{0}^{N_t} X_i(T_i) di .$$

Here N_t is the rate of consumption, so the rate of utility is the sum of the demand shocks for products that are consumed in period t. Note that we have to distinguish between the rate of entry and the rate of activation because some patents may happen not to be used if there is uncertainty. Thus, the rate of entry, N_s, may be larger than the rate of activation, N_t.[5]

By similar arguments as in the deterministic model, aggregate utility over an infinite time horizon becomes:

$$(11) \qquad \int_{0}^{\infty} E\left[u_t \cdot e^{-\rho t}\right] dt .$$

---

[4] The δ-function is characterized by $\delta(x) = 0$ for all $x \neq 0$, and $\int_{-\varepsilon}^{\varepsilon} \delta(x) f(x) dx = f(0)$ for all $\varepsilon > 0$.

[5] The characterisation of equilibrium often simplifies if all patents are actually used (with probability one). We return to requirements for this in specific cases.

The main difference from the deterministic example is that we must take expected values because uncertainty will be allowed, but as in the deterministic case, utility must be optimized backwards by considering the marginal decision that leads up to maximum utility in each period. Further, we know from *Chapter 2* that the T_i's will be first hitting times up to some fixed level X^*. It follows that it suffices to optimize the function

$$(12) \qquad U = \int_0^N X_i \cdot Q_i di,$$

where N is the rate of entry (which is independent of time in equilibrium). Note that this is like (5), except that the discount factor is an expected value. Furthermore, as discussed in *Chapter 1*, Q_i will be decreasing in an arbitrary $X_i > \overline{X}$, since the farther X_i and $\overline{X}$ are apart, the longer it will take to reach X_i. Thus, the discount factor can be defined equivalently as a function $Q_i = Q_i(X_i)$, conditional on the particular process and considering the initial value as a constant. It also follows that the inverse function $X_i = X_i(Q_i)$ is well-defined. This is shown in Figure 3 for the case where the process starts at the fixed initial value $\overline{X}$. With no delay, $X_i(Q_i{=}1) = \overline{X}$, and $Q_i(X_i = \overline{X}) = 1$.

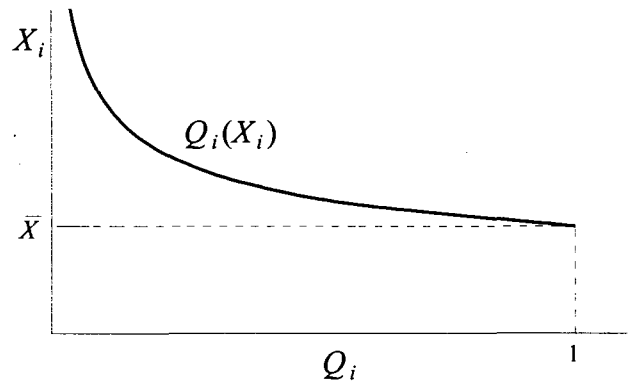

Figure 3. The Expected Discount Factor.

By inserting for X_i in eqn. (12), we get the following fictitious aggregate utility function:

$$(13) \qquad U = \int_0^N U_i(Q_i) di, \qquad \text{where} \qquad U_i(Q_i) \equiv X_i(Q_i) \cdot Q_i.$$

U_i is the dynamic analogue of a static utility function, defined in terms of the discount factor Q_i acting as a quantity measure. For this to make sense, U_i must be increasing in Q_i. The economic interpretation of the requirement is simply that, *ceteris paribus*, the consumer prefers to have the good soon rather than later. This is an obvious criterion for convergence. We assume that the parameters in the problem are well-defined so that this holds.[6]

Prices can be defined in terms of Q_i, so let $P_i(Q_i)$ be the price of a fixed-sized good depending on the discount factor that applies. Thus, in equilibrium, $P_i(1) = \overline{P}$ is the price of a new good. It follows that the consumer will maximize U given by (13) in terms of Q_i conditional on the income constraint $Y = \int_0^N P_i \cdot Q_i di$, where Y is total income in each period.

One way to think of this income constraint is to imagine that the consumer makes the following contract with all firms that make entry in each period: The firm is to provide the consumer with the good the first time the demand process hits the level X_i that corresponds to the discount factor $Q_i(X_i)$. For this contract the consumer pays P_i. In the following we describe what the price will be, assuming that the firm optimizes expected profit.[7]

Firm behaviour

As the patent cost and the production cost are fixed, the expected cost for a firm contemplating entry is like (7), but Q_i is now an expected value. No matter what the exact price process looks like, the identity $Q_i(\overline{X}, X^*) = Q_i(\overline{P}, P^*)$ will hold in equilibrium (including initial values explicitly in the discount factor function for completeness as in *Chapter 1*). It follows that the expected profit function is

---

[6] For example, the model diverges if $\mu > \rho$ in the wine example, as discounting is not sufficient to suppress the value of waiting. Then the wine would remain in the wine cellar forever, as the gain by waiting increases beyond all limits.

[7] As all agents are risk neutral, it makes no difference when contracts are written. The firm may equally well hold on to its patent until activation, or sell it for the correct price some day between entry and activation. In a specific model with geometric Brownian demand that follows, we will evaluate the correct price for the patent in such cases; i.e., the value of the firm at any time ahead of activation.

(14) $\Pi_i = (P_i - C)Q_i(P_i) - A,$

where $Q_i(P_i)$ follows from maximization of utility as discussed above. As usual, the firm will maximize expected profit by considering the value of a higher net benefit (P_i - C) in the future versus the cost of discounting. Then the optimal price P^* is given by the markup rule

(15) $$\frac{P^* - C}{P^*} = \frac{1}{\varepsilon_Q}, \quad \text{where} \quad \varepsilon_Q \equiv -\frac{dQ(P^*)}{dP} \cdot \frac{P^*}{Q(P^*)};$$

leaving out firm indices since all firms are alike. Intuitively, one should get somewhat similar price processes when starting with a particular set of underlying demand processes. However, most stochastic processes are so complex that it is impossible to obtain analytical expressions for the utility function (13) from a specific demand process; nor can the price process be derived from the utility function.

Fortunately, the relationship between the demand process and the price process can be derived analytically for the processes from *Chapter 1* (except the mean-reverting one). Below we demonstrate this by setting up the complete equilibrium model if the demand process is geometric Brownian. In that case we shall see that the price process is also geometric Brownian. The other processes from *Chapter 1* are discussed in *Appendix A*. We find that the demand process and the price process in those cases will be of a common type, but the parameters will not be exactly the same.

A specific model

Demand

Assume that the demand process for a specific good is geometric Brownian:

(16) $dX_i = \mu X_i dt + \sigma X_i dz.$

The interpretation of symbols is familiar: The drift μ is the expected growth of utility from consumption of a specific good by letting the patent mature. If $\mu > 0$, old products tend to be appreciated, but if there is uncertainty ($\sigma > 0$), this does not have to be the case. The expected discount factor for the geometric Brownian motion was shown in *Chapter 1* to be

(17) $Q_i(X_i) = \left(\frac{\overline{X}}{X_i}\right)^{\beta}$,

where β is the positive root of the following quadratic equation in x:

(18) $\frac{1}{2}\sigma^2 x(x-1) + \mu x - \rho = 0$.

Inserting the inverse function $X_i(Q_i)$ of (17) into eqn. (13), and choosing units by setting $\overline{X} = 1$, the utility function simplifies to:

(19) $U = \int_0^N Q_i^{(\beta-1)/\beta} di$.

The requirement for convergence is that β is finite and exceeds one. Notice that eqn. (19) is like eqn. (6), as the geometric Brownian motion is just an extension of the deterministic geometric motion from the introductory example.

From eqn. (18) it can be shown that $\beta > 1$ if $\mu < \rho$. The limiting behaviour towards infinity is discussed later. Maximization of (19) conditional on the income constraint $Y = \int_0^N P_i \cdot Q_i di$ yields:

(20) $Q_i = \left(\frac{\overline{P}}{P_i}\right)^{\beta}$, where $\overline{P} = Y / \int_0^N Q_i^{(\beta-1)/\beta} di$.

Since N is large, each firm will consider $\overline{P}$ to be constant. It follows that the elasticity of Q_i with respect to P_i will be regarded as constant and equal to β, just like the elasticity of the discount factor for the demand process. Thus, a geometric Brownian demand process leaves us (not surprisingly) with a price process of the same kind.

Analogous to a static model with isoelastic demand, the expected profit is maximized if the patent is activated when the price exceeds a fixed markup over C. Leaving out firm subscripts from now on, we have from (15):

(21) $P^* = \frac{\beta}{\beta - 1} C$.

Equilibrium

Free entry is obtained by the zero profit condition

(22) $(P^* - C)Q^* = A,$

where Q^* is the optimal discount factor. Using eqn. (21), this gives:

(23) $Q^* = R(\beta - 1),$ where $R \equiv A/C.$

In equilibrium, a constant number of firms, N^*, will pay the entry cost A each period. If ρ is interpreted as a growth rate, these firms will activate at expected cost Q^*C, which is smaller than C due to growth of productivity between entry and activation. Thus the steady state condition is $N^*\left(A + Q^*C\right) = W$, where W is the total wage bill (which is a measure of the size of the economy). Using eqn. (23), this simplifies to:

(24) $N^* = \frac{W}{A\beta}.$

As $\beta > 1$, eqn. (24) demonstrates an upper bound for the rate of entry. If $\beta \to 1$, then $Q^* \to 0$. In the limit there are no production costs, and the whole wage bill will be used for patents (i.e., $W = N^*A$).

The steady-state condition will be different and more complex if ρ is interpreted as a rate of time preferences instead of a growth rate. In that case, discounting to some extent only implies a delay of activation; not that aggregate costs of activation go down in steady-state. The exact relationship depends on the percentage of the patents that are actually activated. If all patents can be expected to be used, aggregate costs of activation do not go down at all by waiting; thus the steady-state condition becomes $N^*(A + C) = W$.

In general, however, the demand for some patents may never be high enough for them to be activated. Thus, if ρ is a rate of time preferences instead of a growth rate, the steady state condition becomes $N^*(A + prob(\overline{P}, P^*) \cdot C)$, where $prob(\overline{P}, P^*)$ is the probability that a

price process starting at $\overline{P}$ will ever reach P^*. The following properties can be shown to hold in this geometric Brownian case:

$$(25) \quad prob(\overline{P}, P^*) = \begin{cases} 1 & ,\text{if } \mu > \frac{1}{2}\sigma^2 \\ \left(Q^*\right)^{(1-2\mu/\sigma^2)/\beta} & ,\text{if } \mu \leq \frac{1}{2}\sigma^2 \end{cases}$$

Eqn. (25) can be derived from a result in Dixit (1993).[8] In particular, it can be shown that all values higher than the starting value will be hit with probability one as long as $\mu > \frac{1}{2}\sigma^2$. This explains the upper part. See also *Appendix B*, which determines the number of sleeping patents.

Further, it can be shown that $\beta \to 1$ if $\sigma \to \infty$, which also implies that $prob(\overline{P}, P^*) \to Q^*$. Hence, for high uncertainty the steady-state condition will be almost the same for the two interpretations of ρ. However, we have $prob(\overline{P}, P^*) > Q^*$ for any finite σ, so more resources will be used for activation if ρ is a rate of time preferences than if it is a growth rate. We do not pursue this point further here, but it will be important for the discussion of endogenous growth that follows in *Chapter 4*.

Figure 4 plots how the equilibrium with free entry is reached. Firms choose their price of activation so that the expected marginal revenue equates the expected marginal cost. If the entry rate is low, the profit function is OP', and the optimal policy corresponds to the point Z'. Then the expected profit is positive, and more firms are encouraged to enter. The initial price shifts down to $\overline{P}$ and the optimal discount factor to Q^*, corresponding to the point Z where the expected profit is zero.

---

[8] If P is geometric Brownian with drift μ and volatility σ, as in this case, it can be shown by Ito's lemma that $\ln P$ is arithmetic Brownian with drift $\mu - \frac{1}{2}\sigma^2$ and volatility σ. For an arithmetic Brownian motion with drift $\tilde{\mu} < 0$ and volatility $\sigma > 0$, the probability of reaching P_2 when starting from $P_1 < P_2$ equals $e^{2(P_2 - P_1)\tilde{\mu}/\sigma^2}$, according to Dixit (1993:54). The lower part of (25) follows by setting $\tilde{\mu} = \mu - \frac{1}{2}\sigma^2$, $P_1 = \ln \overline{P}$ and $P_2 = \ln P^*$, and using the optimal discount factor $Q^* = (\overline{P} / P^*)^{\beta}$.

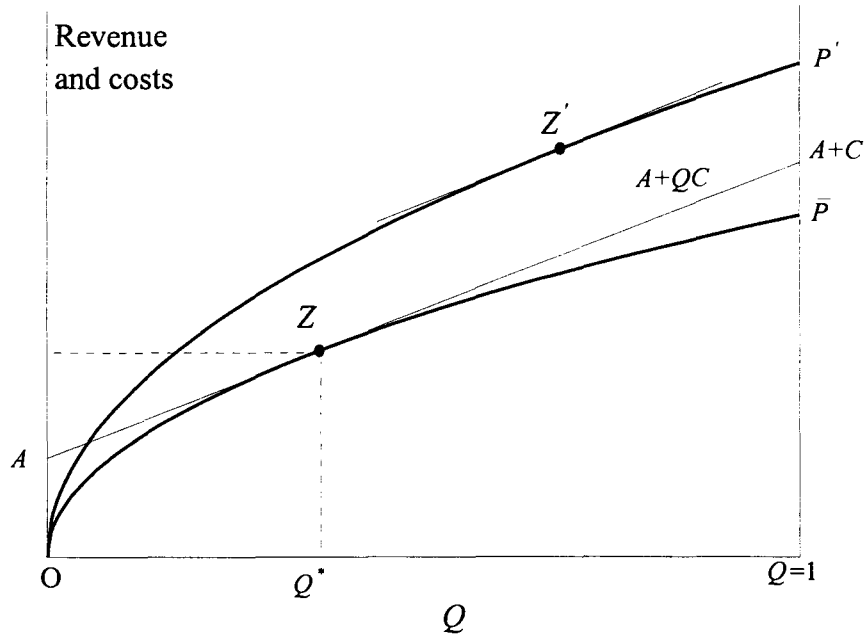

Figure 4. Equilibrium with Free Entry.

The results are analogous to those obtained by static models, except for one constraint. In a static model there is no limit to how much a firm can produce. In this model there is a time axis starting at $Q = 1$ and moving to the left in Figure 4. For convergence, Q^* cannot exceed unity, so the following inequality must hold:

(26) $R(\beta-1)\leq 1$.

This requirement does not have any static parallel, but it has a natural interpretation here: Waiting is not of interest if the patent cost is large relative to the production cost (i.e., if R is large), while at the same time β is not very small. As we will now discuss, β can be interpreted as a measure of "dynamic differentiation", somewhat similar to a static measure of product differentiation (despite the fact that the goods are homogeneous in a static sense). In equilibrium, β also shows up as a measure of dynamic economies of scale. Thus, the requirement of free entry will break down if economies of scale are too large.

Sensitivity

In equilibrium, the utility function (19) yields

(27) $U^* = N^*(Q^*)^{(\beta-1)/\beta}$,

where Q^* is given by eqn. (23). If ρ is interpreted as a growth rate, the entry rate N^* is given by eqn. (24). It can be shown that U^* is decreasing in A and C, which is intuitive. Note also that U^* is linear in W, which is simply the number of workers if the wage rate is normalized to unity. This result, which implies that there are no gains from scale, is discussed in more detail later. The opportunities for maturing are reflected in the fact that U^* is decreasing in β.

It can be shown from eqn. (18) that β is decreasing in σ, increasing in ρ, and decreasing in μ. If σ is large, goods considered to be equal today may be considered to be very different tomorrow, so these homogeneous products are more different in a dynamic respect the smaller β. However, β is increasing in ρ, as the future will be less important the larger the discount rate. Finally, a positive drift will counteract the effect of discounting, so β is decreasing in μ.

Similar to static models, β also shows up as an indicator of dynamic economies of scale in equilibrium, as observed from the markup price (21). In Figure 5, two paths for β are plotted as functions of σ, assuming $\rho = 0.04$. The limiting results are of particular interest.

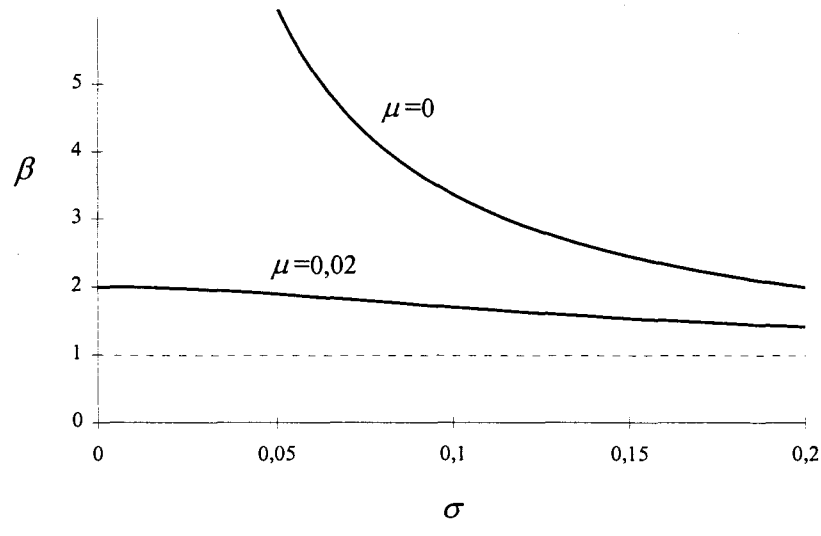

Figure 5. The Elasticity of the Discount Factor for a Geometric Brownian Motion.

First, $\beta \to 1$ if $\sigma \to \infty$, as this removes all product similarities in the dynamic sense. Second, $\beta \to \infty$ as $\sigma \to 0$ if μ is equal to zero (or negative), as this removes the forces that create differences by waiting, fluctuating prices and value from holding a patent. However, β is finite for $\mu \leq 0$ as long as $\sigma > 0$, since uncertainty always creates a probability of price growth.

This gives rise to value from waiting and markup pricing even if the price is expected to decline. Finally, $\beta = \rho/\mu$ if $\sigma = 0$ as long as $\mu > 0$. Thus, uncertainty is not needed if there is a positive drift.

Stability

The value of a firm that has just acquired a patent is found by inserting optimal values into the left-hand side of eqn. (22). This yields:

$$(28) \qquad F(\overline{P}) = \frac{(\beta-1)^{\beta-1}}{\beta^{\beta} C^{\beta-1}} \overline{P}^{\beta}.$$

Here $F(\overline{P})$ is the value of the option to activate as a function of the observed price, which happens to be $\overline{P}$ for a firm that has just made entry. However, the formula is perfectly general and shows the value of the firm at any time ahead of activation.

In Figure 6, the value of a patent is plotted as an increasing function (F) of the demand variable (X) in equilibrium. The value following from the initial shock $\overline{X}$ is denoted by $\overline{F}$, and the patent is activated the first time X reaches $X^* > \overline{X}$. Inserting eqn. (21) into eqn. (28), we have $F^* = C/(\beta - 1)$ at that point.

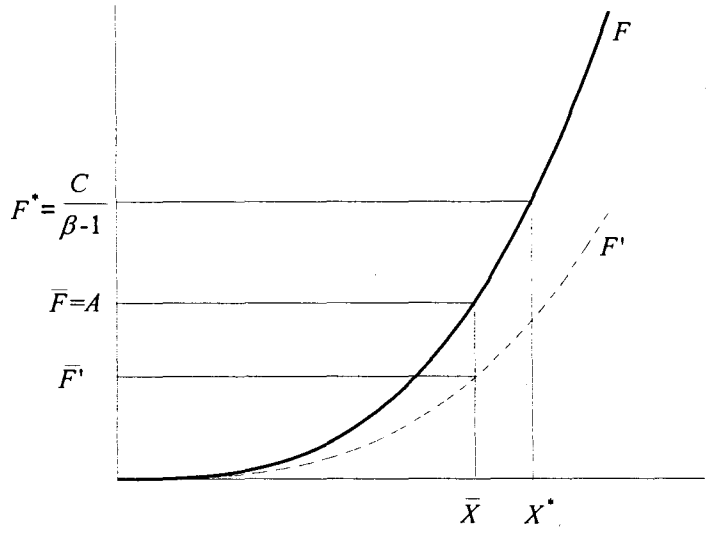

Figure 6. Option Value Functions.

As long as the requirement (26) holds, the entry cost equals the value of a firm that has just acquired a patent. Thus in equilibrium, we have $\overline{F} = A$ as well as $\overline{F} \leq F^*$.

If, by accident, the rate of entry were to increase slightly, starting from equilibrium, the entire value function would shift down as each firm would expect a smaller share of total sales. This is depicted by the dotted line in Figure 6. The initial value of a firm decreases to $\overline{F}'$, and since $\overline{F}' < A$, the expected profit from entry turns negative. For this reason, there will be a temporary stop in entry of new firms. Full employment in this period implies that the rate of activation increases above the equilibrium level. The transition back to equilibrium is speeded up by this effect, and it will go on until the "surplus" has been eliminated by activation of sleeping patents.

Similarly, the value function shifts upward if too few firms enter. New firms will bid up wages slightly to extract expected profits, and we get a transition phase with a higher rate of entry and a lower rate of activation than in equilibrium.

As this shows, any accidental perturbation away from equilibrium will be corrected by the rate at which new firms enter, and the equilibrium is stable.

3. One-sector models with cost variation

Introduction

In this section we discuss equilibrium models that are based on a fixed price and fluctuating costs. The models need not be spelled out in the same detail as above, since they have the "dual" character that was observed in *Chapter 1, Note B*. However, before the formal setup is described, it is convenient to illustrate what we are talking about.

For other reasons than ours, Baldwin and Krugman (1988) discuss the technology of RAM (Random Access Memory) chips for personal computers. This industry is illustrative for the methodological approach, as growth is mostly embodied in new product generations in a fixed-size manner. The technology has improved at an extreme rate since the mid-seventies, starting with the 4Kb and 16Kb generations. The current state-of-the-art is something like 32Mb.

By and large, such chips are homogeneous products, since 2x8Mb more or less equal 1x16Mb etc. Nevertheless, their production costs fluctuate separately. Learning seems to be very important in the first stage after a new generation has been introduced. As time goes by, its market share is reduced, and eventually brought to zero.

Roughly speaking, the 4K chip dominated the market in 1976, the 16K chip in 1980, and the 64K chip in 1983. In the respective years, average prices were \$4.35, \$4.77, and \$3.86, respectively. Taking these as representative cost figures as well, the numbers suggest growth rates of 30-40% per year. This can give a rough idea of the development of production costs relative to the state-of-the-art technology, which is our measurement of scale if ρ is interpreted as a growth rate. Thus, in Figure 7, Baldwin and Krugman's nominal prices for the 4K chip have been scaled up by 30% per year, starting in 1974.

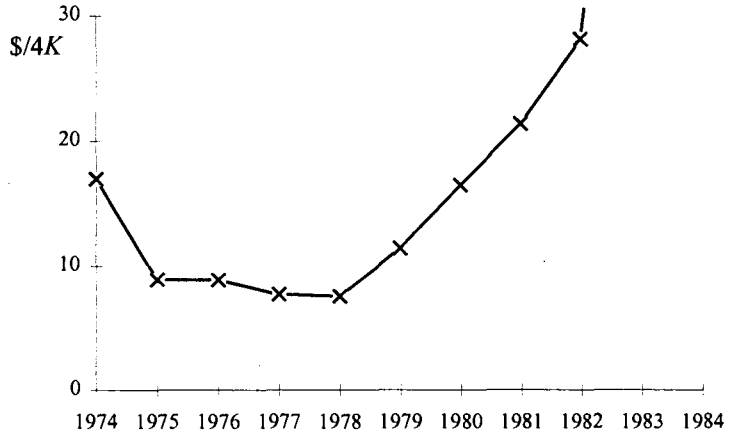

Figure 7. Growth-Corrected Price per 4K RAM.

This figure, which we shall return to later, shows that the average cost (per Kb) decreases faster than the growth rate at first, but more slowly after a while. In 1978, the cost of producing the 4K chip could no longer be reduced fast enough to cope with the next generation. A similar story applies to the 16K chip a few years later, the 64K chip thereafter, etc.

One interpretation of this observation is that learning is important, but that it is limited for each generation of technology. Hence, two opposing forces must be considered: Learning, which favors mature products, and growth, by which a specific generation depreciates.

If we neglect competition between firms that produce the same chip, it is reasonable to interpret the R&D effort by which a new chip is created as a patent. Activation can be interpreted as production. This is the basis for the first model that is presented below.

However, patents seem to be hard to protect in the computer industry (and several other industries). Information about innovations rapidly spills over to competitors, so a model based on exclusive patent rights will obviously lack some important real features. Thus, we shall also consider an extreme alternative where information is freely available. The latter will be analogous to a static model with perfect competition. Most real-world situations are probably somewhere in between these two extremes.

Baldwin and Krugman assume growth to be a result of externalities. Nevertheless, they model it as an exogenous process, arguing that each firm is so small that it will consider the growth rate to be fixed. We make a similar assumption.

Preferences

The instantaneous utility function can be stated as

$$(29) \qquad u_t = \int_{-\infty}^{t} \int_{0}^{N_s} c_{i,s}(t)\, di\, ds ,$$

where the interpretation of variables follows from (8). By similar arguments as in the previous model, this simplifies to

$$(30) \qquad u_t = \int_{0}^{N_t} 1 \cdot di ,$$

and the aggregate utility function to be maximized becomes:

$$(31) \qquad U = \int_{0}^{N} Q_i di .$$

If ρ is interpreted as a rate of time preferences, this says that the products are regarded as identical in all other respects than the time at which they are consumed. Thus they are truly perfect substitutes both in a dynamic and a static sense. With a growth interpretation of ρ,

which is mainly assumed below, the utility function (31) says that new goods are preferred, but two goods that are so old that the technology (productivity) has doubled since they were defined, satisfy the consumer equally as well as one new good. The products are homogeneous like 2x8Mb vs. 1x16Mb RAM. If the rate of income is Y, the preferences (31) give the following inverse demand functions:

$$(32) \qquad P_i = \frac{Y}{\int_0^N Q_i di}.$$

N is supposed to be large, so the influence of a single firm on the price can be neglected. Therefore we set $P_i \equiv P$ for all i. Then P is the price of a product based on a new patent (the current level of technology), while $Q_i P$ is the expected price conditional on a delay T_i.

If the cost of producing such goods were constant throughout time, we would be left with an infinite series of identical static models. Then the dynamic approach would yield no insight that could not be obtained from a static model. However, we now turn to dynamic models based on these preferences but where costs will change.

A model with patent rights

Firm behaviour

The life cycle of a firm is as usual: It makes entry by an irreversible patent cost A. The patent triggers a production cost process with the familiar characteristics. As a basic assumption, the initial cost $\overline{C}$ is so high that simultaneous entry and activation is not optimal. Then the firm has only one choice in order to make (or increase) profit: It must wait for a lower cost. Since prices are constant in terms of Q_i, the firm maximizes the expected profit function

$$(33) \qquad \Pi_i = (P - C_i)Q_i - A,$$

where C_i is the state of the cost process at time of activation. By arguments such as in *Chapter 1, Note B*, the optimal decision is to activate when a specific level $C^* < P$ is reached for the first time. The discount factor can be described as an increasing function of an arbitrary $C_i < \overline{C}$; i.e., $Q_i = Q_i(C_i)$. By inserting this into eqn. (33) and optimizing with

respect to C_i, the optimal decision for each firm follows the markup rule (leaving out firm subscripts from now on since all firms are alike):

$$(34) \quad \frac{P - C^*}{C^*} = \frac{1}{\varepsilon_S}, \text{ where } \varepsilon_S \equiv \frac{dQ(C^*)}{dC} \cdot \frac{C^*}{Q(C^*)}.$$

For comparison with the previous model we concentrate on geometric Brownian costs:

$$(35) \quad dC = -\mu C dt + \sigma C dz.$$

The economic interpretation of (35) is the following: The production cost will not be reduced by waiting exactly at the same rate as technology grows. There may be a deviating trend as well as uncertainty. If $\mu > 0$, some kind of learning is expected. The opposite ($\mu < 0$) is reasonable if knowledge or other specific resources must be restored as production is postponed, e.g. because workers leave. For a (downward moving) geometric Brownian motion, we have shown that the discount factor is

$$(36) \quad Q = (C / \overline{C})^{\alpha},$$

where α is the positive root of eqn. (18) if the minus sign in the parenthesis is changed to a plus sign. (See *Chapter 1, Note B.*) Thus $\varepsilon_S = \alpha$, and the optimal investment cost becomes a fixed fraction of the price:

$$(37) \quad C^* = \frac{\alpha}{\alpha + 1} P.$$

Equilibrium

Free entry is required, so the expected profit at time of entry must be zero in equilibrium. Setting $\Pi = 0$ in (33), and using (36) and (37), the equilibrium discount factor becomes:

$$(38) \quad Q^* = \left(\overline{R}\alpha\right)^{\alpha/(\alpha+1)}, \text{ where } \overline{R} \equiv A / \overline{C}.$$

As in the model with geometric Brownian price variation, eqn. (38) leads to a requirement for convergence. The expected discount factor cannot exceed unity, so

$$(39) \quad \overline{R}\alpha \leq 1$$

must hold for an equilibrium with free entry and zero profit in expectation. If ρ is interpreted as a growth rate, the steady-state condition is $N^*(A+Q^*C^*)=W$, with the same interpretation of variables as with price variation. This simplifies to

$$(40) \qquad N^* = \frac{W}{A(1+\alpha)}.$$

If $\alpha \to 0$, aggregate expected production costs approach zero, and the whole wage bill is used for patents ($N^*A \to W$) as in the previous model when $\beta \to 1$. Finally, the steady-state condition is altered if ρ is interpreted as a rate of time preferences, as opposed to a growth rate. The arguments are so similar to the previous model that we do not pursue this case. However, *Appendix B* argues that all patents are expected to be used if $\mu > -\frac{1}{2}\sigma^2$.

Sensitivity

In equilibrium, the utility function (31) simply yields

$$(41) \qquad U^* = Q^*N^*,$$

which is the expected rate of consumption. If inserting Q^* and N^* from eqns. (38) and (40), the same result is obtained as for similar values in the model with price variation; i.e., by inserting eqns. (23) and (24) into (27). We just have to replace R with $\overline{R}$, and β with $\alpha + 1$. This also holds for most other equilibrium formulas of interest, so a discussion on sensitivity can be kept short, focusing the relationship between α and its underlying variables.

It can be shown that $d\alpha/d\sigma < 0$, and that $\alpha \to 0$ as $\sigma \to \infty$. Hence, cost uncertainty increases utility in equilibrium, since cost savings become more likely. Furthermore, $\alpha = \rho/\mu$ if $\mu > 0$ and $\sigma = 0$, while $\alpha \to \infty$ as $\sigma \to 0$ if $\mu \leq 0$. The latter case shows how the value of waiting decreases to zero if the possibility for a cost reduction disappears. The limit with infinite α corresponds to no value from the patent, and a price equal to the cost of activation ($P = C^*$). We also have $d\alpha/d\mu < 0$, as the cost of activation will tend to be low if the expected learning rate is high. Finally, $d\alpha/d\rho > 0$, as any stochastic or deterministic property of the process is discouraged by discounting.

Hence, the relationship from α to ρ, μ and σ is very much like the relationship from β to its similar parameters. Therefore a reference to Figure 5 suffices as far as graphics are concerned. Also, α can be interpreted as a measure of dynamic economies of scale and of differences - this time with respect to costs (or technology). Notice, however, that $\beta(\mu,\sigma,\rho)$ does not equal $\alpha(\mu,\sigma,\rho)$ if $\sigma > 0$, so the partial derivatives are not identical.

According to eqn. (38), the optimal discount factor will be larger the larger A and the smaller $\overline{C}$. However, it can be shown that Q^* is U-shaped as a function of α. It follows that the allocation of resources between patenting and production, as well as the expected age of products that dominate the market, do not depend uniquely on α. *Chapter 5* demonstrates a quite surprising effect of this result in a location model with knowledge spillovers.

Finally, let us add some intuition to the convergence requirement (39) in the deterministic case ($\sigma = 0$). Then $\alpha = \rho/\mu$, and the criterion can be stated as:

$$(42) \qquad \frac{\overline{C}}{A} \cdot \frac{\mu}{\rho} \geq 1.$$

The first term on the left-hand side is the ratio of the initial production cost to the patent cost. It can be regarded as a static measure of the importance of learning. The second term, which is the ratio of the learning rate to the growth rate, makes up a similar dynamic measure. For waiting to be a matter of interest, the product of the two must exceed unity. In the RAM industry, it obviously does. Although the growth rate is extremely high ($\rho = 0.3 - 0.4$), the learning rate and the production cost share are large enough to pin it down. It usually takes several years from the technology of a microchip is available until production becomes so cheap that the chip dominates the market.

A model with perfect competition

Firm behaviour

A model without patent rights, but a production cost that fluctuates separately for each generation of technology, can be based on a cost function with two components as above. However, with no dynamic market power the investments will no longer be separated in time.

First there is a fixed cost which covers activities that are independent of the specific technology, and not exposed to learning. Mathematically, it will be somewhat analogous to the patent cost, so we use the term A for it. Here it may be more appropriate to call it a setup cost, which could be interpreted as costs related to infrastructure, administrative support etc. If ρ is interpreted as a growth rate, it follows that the fixed part of the effective unit cost equals A/Q ($\geq A$) for a product whose technology is T periods old, as the effective size of the good will be smaller than that of a new one.

The second cost component covers activities exposed to learning, and it will be described by a process with the familiar properties. Let us denote it by C and interpret it as a production cost. As usual, it is also measured relative to the state-of-the-art technology. If the production cost is some general C when the product is introduced in the market (by the undertaking of both investments), it follows that the total unit cost is:

(43) $J = A / Q + C$.

A large number of products are assumed to be launched each period. Because they substitute perfectly and do not change by ageing, as viewed by the consumer, the product price will be constant. Then the expected profit function is simply P - J, where P is fixed. It follows that expected profit is maximized if (43) is minimized.

When will a product be introduced? Since A is constant, it is clear from (43) that the optimal policy is to wait until the cost has reached a specific level C^* ($<\overline{C}$) for the first time. (By waiting until the second time, Q would decrease. However, since A is fixed, J then increases, and the expected profit decreases.)

As usual, we can set $Q = Q(C)$, where $dQ/dC > 0$. Inserting this into (43) and minimizing with respect to C give the first-order condition

(44) $$\frac{A}{Q^*C^*} = \frac{1}{\varepsilon_S},$$

where ε_S is defined as in eqn. (34). Hence, the ratio of the setup cost (A) to the expected production cost (Q^*C^*) equals the inverse of the elasticity of the discount factor. If waiting is

to apply, $dJ/dQ > 0$ is required at $Q = 1$. Assuming that this holds, the general shape of $J(Q)$ will be as plotted in Figure 8.

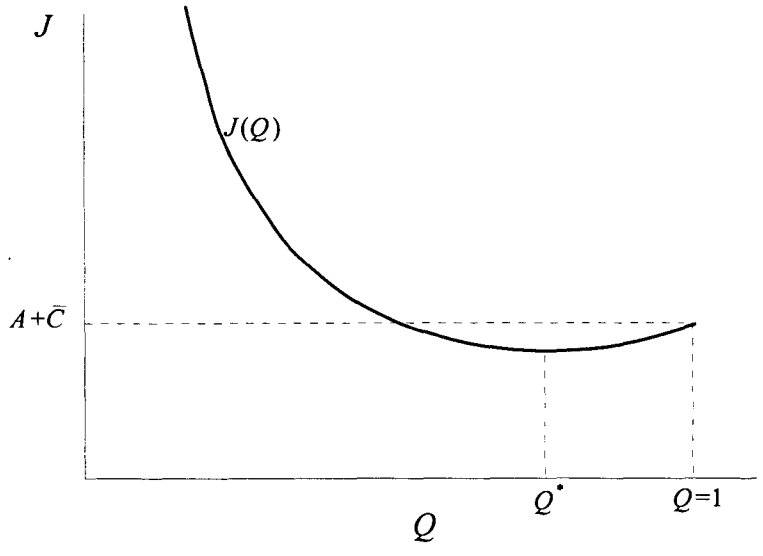

Figure 8. A Dynamic Cost Function with Learning.

Note that Figure 8 fits in well with the RAM industry studied by Baldwin and Krugman. (In Figure 7, the time axis goes from left to right while here it goes the other way, and uses a different scale. However, Figure 7 could easily be transformed to something like Figure 8.)

Equilibrium

If profits are driven to zero by free entry, we get the condition:

(45) $A / Q^* + C^* = P$.

Equilibrium is illustrated in Figure 9. Firms invest when located at the minimum of the cost-curve; i.e., when the potential for learning has been optimally exploited. With a large number of firms, this can be approximated by a horizontal supply curve (J_{min}). The downward-sloping demand curve $D(P)$ gives total demand as a function of prices at that particular time, and partial equilibrium is where demand intersects with supply. As long as no firm dominates the market, the size of the firm is irrelevant.

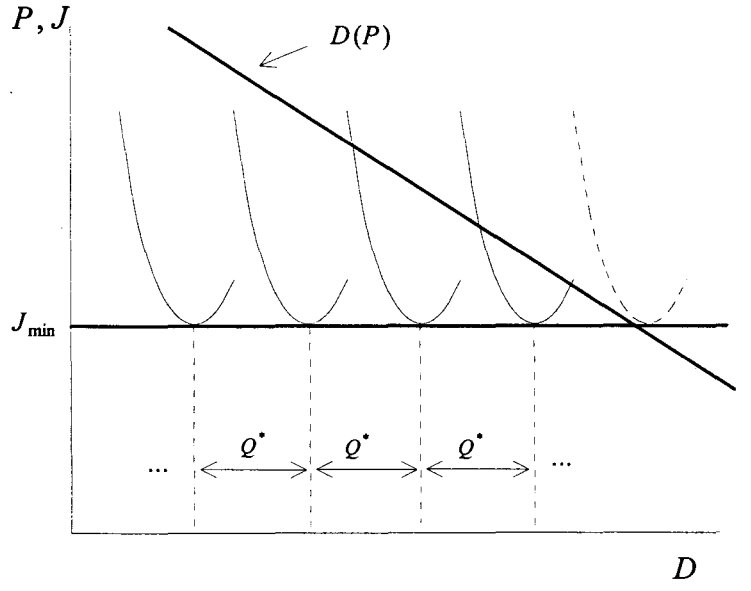

Figure 9. Perfect Competition in a Dynamic Model with Learning.

One interesting point to note is that (44) and (45) can be combined to yield the markup rule (34), and a one-sector equilibrium model follows from a full employment condition. If C is geometric Brownian, the optimal discount factor is like Q^* in eqn. (38), the entry rate like N^* in eqn. (40), etc. This also implies that the allocation of resources between the A-sector and the C-sector coincides in the two models.

The fact that the optimal investment rule coincides illustrates a more general result from Leahy (1993): A firm can act myopically when considering when to invest; i.e., it can believe that it has an exclusive option to the fluctuating benefit. It does not matter for the optimal decision whether its profit is actually brought to zero by others.

4. A two-sector model

In this section we combine some of the assumptions and results from previous sections. More precisely, the model with geometric Brownian demand in Section 2 will be combined with the setup of preferences from Section 3. Consider the following instantaneous utility function:

$$(46) \qquad u_t = \left(\int_{-\infty}^{t} \int_{0}^{N_s} X_{i,s}(t-s) \cdot c_{i,s}^{M}(t) di\, ds \right)^{\pi k} \cdot \left(\int_{-\infty}^{t} \int_{0}^{H_s} c_{i,s}^{H}(t) di\, ds \right)^{1-\pi}$$

This utility function has two parts as there are two goods: M ("manufactured goods") and H ("homogenous goods"). Further, π (<1) and k are positive constants which will be explained below. The interpretation follows from (8) and (29), but notation has been changed slightly for clarification.

If the manufactured good is wine, it may be reasonable to think of the other good as food (as our consumer probably prefers a proper meal). Food production is characterized by constant returns to scale, implying that this is simply a perfectly homogeneous good which can be produced (and consumed) in any quantity and at any time. Wine, however, is characterized as in Section 2, with some minimum firm size and two cost elements. Further, there is value from postponing the last investment. Simplifying as in Section 2, the utility function becomes:

$$(47) \qquad u_t = \left(\int_0^{N_t} X_i(T_i) di \right)^{\pi k} \cdot \left(\int_0^{H_t} 1 \cdot di \right)^{1-\pi} .$$

If the demand processes X_i are continuous and autonomous with the Markov property, it follows that utility is maximized if the patents to manufactured goods are activated when a specific threshold X^* is reached for the first time. Then the T_i's are first hitting times, and we can form a fictitious utility function

$$(48) \qquad U = \left(\int_0^{N} Q_i X_i di \right)^{\pi k} \cdot H^{1-\pi} ,$$

where N is the rate of entry. Assume that the demand processes for manufactured goods are geometric Brownian with drift μ and volatility σ. In that case, $X_i = \bar{X} Q_i^{-1/\beta}$, where $\bar{X}$ is initial demand and β is the well-known positive root of eqn. (18). If utility is scaled by setting $\bar{X} = 1$, and k is defined by $k = \beta/(\beta\text{-}1)$, the utility function can be rewritten as:

$$(49) \qquad U = M^{\pi} H^{1-\pi} ,$$

where M is an aggregate:

$$(50) \qquad M = \left(\int_0^N Q_i^{(\beta-1)/\beta} di \right)^{\beta/(\beta-1)} .$$

The preferences (49),(50) resemble a Dixit-Stiglitz model with constant elasticities, but our model has a different interpretation. Nevertheless, the procedures for maximizing utility conditional on the income constraint will be the same. By a standard argument it can be shown that the consumer each period will spend an income share π on wine and 1-π on food. If this had been a static model with small firms, each firm in manufacturing could act as a monopolist facing a constant price elasticity of demand equal to β. There would then be a set of demand functions for single goods

$$(51) \qquad Q_i = \left(\frac{P_i}{G} \right)^{-\beta} \frac{\pi Y}{G},$$

where G is a perfect price index:

$$(52) \qquad G = \left(\int_0^N P_i^{1-\beta} di \right)^{1/(1-\beta)} .$$

With our interpretation, N is the rate of entry, P_i is the activation price in terms of the discount factor, Y is the rate of income, and β is the elasticity of the discount factor with respect to the price of activation. Thus each small firm will act as a dynamic monopolist that faces an independent geometric Brownian price process. Since consumption shares are fixed, equilibrium for manufacturing follows from the model with geometric Brownian prices in Section 2.

Let us describe the behaviour of the model in some more detail. First, note that the utility function says that if the consumer cannot get wine with his dinner, then he will not enjoy his food either. In principle, if the size of the economy did not permit more than one glass of wine every other period, then nor would the consumer want any food in the other periods. If the size of the economy doubled, however, he could get half a glass of wine each period, and he would then also buy (half as much) food each period.

Since N is large, the consumer in this model will definitely get several meals each period, but does he gain from increasing the frequency? No, discounting is neglected within each period, so the frequency is not interesting. For example, if the period is a year, the consumer does not care whether he gets a piece of bread and a glass of wine once a day, or a bottle of wine and a loaf of bread once a week.

This argument complies with previous statements on scale effects: There are no gains from scale in any of the models in this chapter. This also means that there will not be intra-industry trade at positive trade costs as in static models with monopolistic competition. In *Chapter 6* and *Chapter 9*, however, we shall see how scale effects and gains from trade can arise by technically small changes in the utility function.

5. Extensions

A new interpretation of the firm

Up to now, a firm has been unit-sized in the sense that a patent enables it to produce one single unit at one moment in time. The models become more appealing by not letting the firm disappear as soon as its patent has been activated. Instead, let activation be interpreted as the establishment of an infinite-lived factory with a fixed capacity of one unit per period. This extension can be incorporated quite easily.

For simplicity, let us imagine that the consumers buy the consumption rights to all future production at time of activation. Hence, as soon as the patent has been activated, the consumer will get one unit of the good each period in all future. Technically, the δ-functions in the utility functions can be replaced by step functions to reflect this. In the model with geometric Brownian price variation, the resulting utility function will be like (19) except for a constant factor $1/(\rho-\mu)$. The interpretation of this factor is intuitive: It represents the multiplicative effect of being able to consume a good perpetually, when discounting at a rate which is corrected for expected maturing.

Each firm will observe an independent geometric Brownian flow price, and if the price takes on the value p at some time t, the net present price P (i.e., the discounted expectation of future values), is given by:

$$(53) \qquad P = E\left[\int_t^{\infty} p_t \cdot e^{-\rho t} dt\right] = \frac{p}{\rho - \mu}.$$

Thus P is also geometric Brownian, and essentially nothing has changed. We may also add a Poisson "death" process with a fixed intensity λ for operating factories to avoid an infinite number of firms in equilibrium. It is a standard result, see Dixit and Pindyck (1994:200), that this simply changes the proportional factor to $1/(\rho-\mu+\lambda)$.

Alternative interpretations of the discount rate

Two interpretations of ρ have been applied so far. It has either been a subjective rate of time preferences or an aggregate growth rate of productivity or technology. More options exist.

Up until now, a patent holder has been able to wait forever for its demand to rise to an optimal level. Such extreme assumptions can be avoided by a Poisson process somewhat similar to the one discussed above. Hence, we can introduce a fixed probability ρ_λ that a sleeping patent becomes worthless in the next period, e.g. because of a technological breakthrough. The effective discount rate increases accordingly. In the Section 2 model with geometric Brownian prices, β will increase. This affects the steady-state condition, but all major forces in the model remain. (The most reasonable assumption is probably to use the same process for both extensions, by setting $\rho_\lambda = \lambda$.)

Further, ρ might also embody growth of the factor endowment. Consider the wine example again, but exclude productivity growth. Normalizing the wage rate to unity and assuming labor is the only production factor, then A and C can be interpreted as fixed shares of the labor force. If the labor force grows at a fixed rate, a fixed share of the labor force will be able to make more wine and fill larger barrels as time goes by, even with no growth of productivity. (As the increased production must be shared by a similar increase in consumers, there are still no gains from scale.)

Hence, we can set $\rho = \rho_d + \rho_\lambda + \rho_g + \rho_l$, where ρ_d is a subjective rate of time preferences, ρ_λ is the intensity of a Poisson "death" process for sleeping patents, ρ_g denotes growth of productivity (or technology), and ρ_l is the growth of the labor force.

Endogenous patent costs

The patent cost can be endogenized as in *Chapter 1*, assuming that future revenues can be raised by increasing A. Now we have $Q = Q(P, A)$, and $Q_A > 0$. By including the second argument in the profit function (14) and optimizing with respect to P and A, eqn. (15) still holds. The following relationship (from *Chapter 1*) also applies:

$$(54) \qquad \frac{A^*}{Q^*P^*} = \frac{\varepsilon_Q^A}{\varepsilon_Q^P}, \quad \text{where} \quad \varepsilon_Q^A \equiv \frac{A^* Q_A(P^*, A^*)}{Q(P^*, A^*)}, \quad \varepsilon_Q^P \equiv -\frac{P^* Q_P(P^*, A^*)}{Q(P^*, A^*)}.$$

Hence, in optimum the ratio of the patent cost (A^*) to expected revenue (Q^*P^*) is given by the ratio of the elasticities. This can be extended to equilibrium. Combining eqn. (54) with free entry, the following condition is derived:

$$(55) \qquad \varepsilon_Q^A = 1.$$

If the price processes are geometric Brownian as in Section 2, eqn. (54) simplifies to:

$$(56) \qquad \frac{A^*}{Q^*P^*} = \frac{1}{\beta}.$$

We still have $Q^* = (\overline{P} / P^*)^\beta$, but now β is a function of A^*, so we ought to put an asterisk on it. Differentiation yields $Q_A^* = Q^* \beta_A^* \ln(\overline{P} / P^*)$, and eqn. (55) can be rewritten as

$$(57) \qquad Q^* = e^{-1/\varepsilon_\beta^A} \quad \text{where} \quad \varepsilon_\beta^A \equiv -\frac{A^* \cdot \beta_A^*}{\beta^*}.$$

The elasticity ε_β^A measures the effectiveness of increasing the ancillary investment. As eqn. (57) shows, the expected delay goes to zero ($Q^* \to 1$) as ε_β^A approaches infinity. It goes to infinity ($Q^* \to 0$) as ε_β^A approaches zero.

There are numerous ways by which β could be decreased: by increasing the expected price growth (μ) or the probability for such growth (σ), by decreasing the effective discount rate (ρ), or by a combination. *Chapter 1* discussed the first of these cases at the level of a firm. *Appendix C* extends all cases to equilibrium, also discussing interpretations in each case.

6. Final remarks

The framework that has been developed, relates a number of dynamic variables to static modelling. The characterization of costs and preferences, as well as several equilibrium formulas, coincide with analogous static models. Thus, some static models can possibly be reinterpreted and their results confirmed in a dynamic context. In addition, a number of additional questions arising from irreversibility and uncertainty can be addressed. This will become clearer as the approach is extended and applied in a number of directions in the chapters that follow. More variables are going to be endogenized, and by technically small changes in the utility functions, it will also be shown how the analogy to static modelling can be refined.

The specific models have also brought forward some criteria for waiting to matter in an industry-wide context. Although the criteria are stylistic as they are based on a particular set of processes, as well as on extremely symmetric models, they can provide some crude estimates that may supplement our introductory remarks on the importance of this kind of models.

For example, according to (23), waiting is not optimal in the model with geometric Brownian prices if $R(\beta - 1) \geq 1$. Empirical analyses indicate that firms typically require expected returns that could be three or four times the cost of capital; see Dixit and Pindyck (1994:7). Emphasizing that our calculations here are very rough, the investment rule (21) then suggests β-values that in many cases can be well below three. Then waiting will apply if R does not exceed one half. In broad terms, R defines a threshold for the ratio of typical ancillary investments (like R&D) to typical production investments, if waiting is to apply. Thus, ancillary investments should not represent more than one third of total investments. It seems reasonable to assume that the true number is often smaller; if this is the case, waiting matters.

Acknowledgements

I would like to thank Victor Norman, Terje Lensberg and Anthony Venables for helpful comments and clarifying discussions on this chapter.

APPENDIX A: Variable elasticities

Introduction

Consider a one-sector economy based on irreversible investments and price variation as in Section 2. Patent costs and production costs are constant, and preferences are given by an increasing, strictly concave function (13). As shown by Dixit and Stiglitz (1977) for the analogous static model, the price elasticity of demand facing a single, small firm is

$$\text{(A.1)} \quad \varepsilon_i(Q_i) = -\frac{U_i'(Q_i)}{U_i''(Q_i)\cdot Q_i},$$

where the primes denote first and second order derivatives. Krugman (1979) assumed that $d\varepsilon_i/dQ_i < 0$ in the static model, demonstrating two types of gains from trade even with no differences in preferences or factor endowments: First, there is a gain as economies of scale can be exploited. Second, the consumers get access to a larger number of products.

If the demand processes in the dynamic setup are geometric Brownian like the process in the first example in *Chapter 1*, ε_i is constant and equal to the elasticity of the discount factor (β). Thus, the demand process and the price process are practically the same. (In the static model, the only gain from scale in this case is due to variation.)

Below, we discuss the similar relationship if the demand processes are like the other processes from *Chapter 1*. As a matter of fact, $d\varepsilon_i/dQ_i < 0$ for all those that can be treated analytically. At first sight, this may not seem very interesting, as the dynamic model yields no scale effects. However, it will follow from *Chapter 6* that a similar argument also applies in a modified version with true product differentiation.

Arithmetic Brownian motion

If demand is described by arithmetic Brownian motions

$$\text{(A.2)} \quad dX_i = \mu dt + \sigma dz,$$

it was shown in *Chapter 1* that the discount factor is

(A.3) $Q_i(X_i) = e^{-\gamma(X_i - \bar{X})}$,

where γ is the positive root of the following quadratic equation:

(A.4) $\frac{1}{2}\sigma^2 x^2 + \mu x - \rho = 0$.

From (A.4), it follows that $d\gamma/d\mu < 0$, $d\gamma/d\sigma < 0$, and $d\gamma/d\rho > 0$, so γ is a measure of variation quite similar to β for the geometric Brownian motion. By inverting (A.3) and setting $U_i(Q_i) = Q_i \cdot X_i(Q_i)$ as in eqn. (13) in the text, we find:

(A.5) $U_i(Q_i) = \left(\bar{X} - \frac{\ln Q_i}{\gamma}\right) Q_i$.

Further, we have

(A.6) $U_i'(Q_i) = \gamma \varepsilon_i$,

where

(A.7) $\varepsilon_i = \gamma \bar{X} - 1 - \ln Q_i$.

(As observed, ε_i is decreasing in Q_i.) Since ε_i is also the price elasticity of demand, (A.7) yields

(A.8) $\frac{dQ_i}{dP_i}\frac{P_i}{Q_i} = 1 - \gamma \bar{X} + \ln Q_i$,

or:

(A.9) $\frac{dQ_i}{Q_i\left(1 - \gamma \bar{X} + \ln Q_i\right)} = \frac{dP_i}{P_i}$.

Noting that $\int 1/x(c + \ln x)dx = \ln(c + \ln x)$ plus a constant, both sides can be integrated, and we have:

(A.10) $\ln Q_i = \gamma \bar{X} - 1 + KP_i$.

Here K is a constant of integration that depends on the initial price. It is determined by the boundary condition $Q_i(\bar{P}) = 1$. The discount factor simplifies to

(A.11) $Q_i(P_i) = e^{-\tilde{\gamma}(P_i - \bar{P})}$, where $\tilde{\gamma} = \dfrac{\gamma\bar{X} - 1}{\bar{P}}$.

Referring to (A.3), this corresponds to an arithmetic Brownian price process slightly different from the demand process. The reason for the difference is that utility is discounted geometrically, while the demand process is arithmetic. Therefore the price process is affected by the initial demand $\bar{X}$.

As an example, we have $\tilde{\gamma} = 1/\bar{P}$ if $\bar{X} = 2/\gamma$. A free entry condition similar to eqn. (22) determines $\bar{P}$ in equilibrium as a function of A and C. By some algebra, it can be shown that $\bar{P}$ is given implicitly by the equation $A = \bar{P}e^{-C/\bar{P}}$ in this case.

The linear case

Let firm-specific demand be characterized by the process

(A.12) $dX_i = \mu(\phi - X_i)dt + \sigma(\phi - X_i)dz$,

where ϕ $(> \bar{X})$ is a constant. Then the discount factor is linear

(A.13) $Q_i(X_i) = a - bX_i$, where $a = \dfrac{\phi}{\phi - \bar{X}}$, $b = \dfrac{1}{\phi - \bar{X}}$,

as long as the combination of parameters (μ,σ,ρ) is such that the positive root of the following quadratic equation equals one:

(A.14) $\frac{1}{2}\sigma^2 x(x+1) + \mu x - \rho = 0$.

(See *Chapter 1, Note A.*) The utility function $U_i(Q_i) = X_i(Q_i)\cdot Q_i$ follows directly from (A.13):

(A.15) $U_i(Q_i) = \phi Q_i - (\phi - \bar{X})Q_i^2$,

and we obtain:

(A.16) $\varepsilon_i(Q_i) = \dfrac{\phi}{2(\phi - \bar{X})Q_i} - 1.$

Integrating (A.16) as in the previous example yields the following discount factor:

(A.17) $Q_i(P_i) = \tilde{a} - \tilde{b}P$, where $\tilde{a} = \dfrac{\phi}{2(\phi - \bar{X})}$, $\tilde{b} = \dfrac{\phi - 2\bar{X}}{2\bar{P}(\phi - \bar{X})}$.

Hence, the demand process leads to a linear discount factor, and a price process quite similar to the demand process.

Mean-reversion

Krugman (1979) does not give any reason why ε_i ought to decrease as a function of Q_i in the static model (but Krugman, 1980, argues for it in a footnote). In the dynamic setting there are good reasons to expect that this is the case. It is often expected that prices will be mean-reverting as in some sense they are related to long-run marginal costs. In our context, mean reversion can be related to stable preferences. If the price for product *i* is given by the process

(A.18) $dP_i = \eta(\hat{P} - P_i)P_i dt + \sigma P_i dz$,

where η and $\hat{P}$ denote the speed and the mean, it was argued in *Chapter 1* that the elasticity of the discount factor is increasing in P_i, because the price will tend to stay close to the mean; thus the discount factor will be larger in the vicinity of the mean when comparing with the geometric Brownian motion that appears if $\eta = 0$. It will be accordingly smaller than the discount factor for a geometric Brownian motion if the price is above a certain level. Since ε_i is constant for the geometric Brownian motion, it follows that it will increase in P_i and therefore decrease in Q_i for this mean-reverting process.

Final remarks

The main lesson from these examples is the observation that a particular demand process seems to lead to a somewhat related price process. This would probably also hold if starting with a mean-reverting demand process similar to (A.18). As for most other stochastic processes, however, an analytical solution is hard to obtain in this case.

APPENDIX B: The number of sleeping patents

In the models with price variation, the firm will either be waiting for the optimal price level to be reached from below, or it will already have reached this price and activated the patent. The question then arises: What is the number of sleeping patents in equilibrium?

In simple cases, the answer can be found by some standard results from stochastic calculus. We focus on the Section 2 model. If P is a geometric Brownian motion with drift μ and volatility σ as in this model, Ito's lemma says that $x = \ln(P)$ is arithmetic Brownian with drift $\tilde{\mu} = \mu - \frac{1}{2}\sigma^2$ and volatility σ. For such a process, Dixit (1993:56) finds that the expected time going from x_0 to $x_1 > x_0$ is

$$\text{(B.1)} \quad E_+[T] = \begin{cases} \dfrac{x_1 - x_0}{\tilde{\mu}}, & \text{if } \tilde{\mu} > 0 \\ \infty, & \text{if } \tilde{\mu} \le 0 \end{cases}$$

The fact that the expected time is infinite with a non-positive drift reflects a probability that the upper threshold will never be reached. With a positive drift; i.e., if $\mu > \frac{1}{2}\sigma^2$ in the geometric Brownian process, the expected time does exist, so any value higher than the starting value will be hit with probability one. In particular, this must hold for $\overline{P}$ and P^*, and we can be sure that all firms entering at $\overline{P}$ will become active some day. The rate of activation will then equal the rate of entry in equilibrium by the law of large numbers. (If it did not, the number of sleeping patents must either increase or decrease. It cannot decrease, however, since there is a fixed in-flow of firms and no other sources of exhaust. If, on the other hand, it were to increase, there would be a positive probability that the threshold would never be hit, which contradicts the result above.)

By inserting $x_0 = \ln(\overline{P})$ and $x_1 = \ln(P^*)$ into (B.1), and using the optimal discount factor $(\overline{P} / P^*)^{\beta} = R(\beta - 1)$, the expected time from $\overline{P}$ to P^* becomes:

$$\text{(B.2)} \quad E_+^*[T] = \begin{cases} \dfrac{\ln\left(1 / R(\beta - 1)\right)}{\beta\left(\mu - \frac{1}{2}\sigma^2\right)}, & \text{if } \mu > \frac{1}{2}\sigma^2 \\ \infty, & \text{if } \mu \le \frac{1}{2}\sigma^2 \end{cases}$$

To obtain the number of sleeping patents in steady-state, (B.2) must be multiplied by the entry rate, which is given by eqn. (24) if ρ is a growth rate. The number of sleeping patents is zero if $R(\beta - 1) = 1$, since this implies immediate activation. Further, the number goes to infinity if μ approaches $\frac{1}{2}\sigma^2$ from above, as the demand for some patents will never be high enough for activation in such cases. If $\mu \leq \frac{1}{2}\sigma^2$, the entry rate will be a factor higher than the rate of activation, corresponding to the probability that P^* is never hit. (The difference can be calculated using eqn. (25) in the text.)

If firms become operating factories that are exposed to a constant death rate λ after activation, the expected life time of an active firm is $1/\lambda$. The ratio of active firms to the total number of firms follows easily.

Similar arguments can be worked out for the models with cost variation. If the production cost is geometric Brownian with drift $-\mu$ and volatility σ, Ito's lemma and a relationship similar to (B.1) will show that the expected time from $\overline{C}$ down to C^* becomes:

$$\text{(B.3)} \quad E_{\overline{C}}^{C^*}[T] = \begin{cases} \dfrac{\ln\left(1/(\overline{R}\alpha)\right)}{\alpha\left(\mu + \frac{1}{2}\sigma^2\right)}, & \text{if } \mu + \frac{1}{2}\sigma^2 > 0 \\ \infty \quad , & \text{if } \mu + \frac{1}{2}\sigma^2 \leq 0 \end{cases}$$

Thus, some patents will never be used if there is not enough uncertainty to pin down an expected cost increase ($\mu < 0$). To obtain the number of sleeping patents in equilibrium, (B.3) must be multiplied by the entry rate, which in this case is given by eqn. (40) if ρ is a growth rate. The limiting behaviour follows from the discussion above.

APPENDIX C: Examples with endogenous patent costs

In this appendix we discuss three ways to endogenize the patent cost A for the model with geometric Brownian prices in Section 2.

Increasing the expected price growth (μ)

If some of the investment A is interpreted as an initial advertising campaign or another ancillary activity (see *Chapter 1*), it is reasonable to argue that the expected rate of increase in P can be raised by increasing A. Thus, set $\mu = \mu(A)$ and $\mu_A > 0$. Differentiating eqn. (18) in the text with respect to A and rearranging, we have (as in *Chapter 1*):

$$\text{(C.1)} \quad \varepsilon_Q^A = -\frac{A\mu_A \ln Q}{\sigma^2\beta + \mu - \frac{1}{2}\sigma^2}.$$

(Asterisks on optimal values are left out to save notation.) Setting ε_Q^A equal to unity according to eqn. (55), and using eqn. (23), the equilibrium patent cost A can be found. The effect of the patent cost on the discount factor becomes particularly simple if uncertainty is left out ($\sigma = 0$). Then we get

$$\text{(C.2)} \quad Q = e^{-1/\varepsilon_\mu^A}, \quad \text{where} \quad \varepsilon_\mu^A \equiv \frac{A \cdot \mu_A}{\mu}.$$

Hence, the age of patents that are activated will be lower the larger ε_μ^A; in other words, the more effective the advertising efforts.

One objection to this may be that an initial advertising campaign must be continued in order to sustain expected price growth. Such extensions require operating costs and exit options, or at least some kind of stepwise investment. This would need a much more complex model, but the qualitative conclusions may not be very different. (See Dixit, 1989, for a model with entry and exit, and *Chapter 10* for further discussion of the issue.)

Decreasing the effective discount rate (ρ)

As noted in the text, ρ can be interpreted as a rate of depreciation for a patent holder if growth is embodied in new patents. Sometimes a firm that makes an ancillary investment may also

have the opportunity to affect its actual rate of depreciation. For example, the firm could choose a more costly technology which does not pay off by higher prices in the short run, but which is profitable in the long run as upgrading becomes easier.

As in the previous example, a realistic representation requires operating costs and exit options, or upgrading could be represented by stepwise investments. For simplicity, assume that such investments can be neglected or included in A, while the effect is "pasted out" by decreasing the effective rate of depreciation. Formally, set $\rho = \rho(A)$ and $\rho_A < 0$, and differentiate as before. This yields

$$\text{(C.3)} \quad \varepsilon_Q^A = \frac{A\rho_A \ln Q}{\beta(\sigma^2\beta + \mu - \frac{1}{2}\sigma^2)},$$

which determines A when combined with eqns. (55) and (23). Once again, the result simplifies if there is no uncertainty ($\sigma = 0$). Then $\beta = \rho/\mu$, and we have:

$$\text{(C.4)} \quad Q = e^{-1/\varepsilon_\rho^A}, \quad \text{where} \quad \varepsilon_\rho^A \equiv -\frac{A \cdot \rho_A}{\rho}.$$

As observed, (C.4) is similar to (C.2).

Increasing the volatility (σ)

The first example assumed that the firm was able to increase the expected growth rate of the price process. In fact, less effective efforts might also work. Since the firm keeps the option not to activate if the price turns low, it could benefit from just raising the upside price potential by increasing σ. This flattens the distribution of future prices without lifting the expected level. If $\sigma = \sigma(A)$ and $\sigma_A > 0$, the following formula is obtained:

$$\text{(C.5)} \quad \varepsilon_Q^A = -\frac{(\beta - 1)\sigma A \sigma_A \ln Q}{\sigma^2\beta + \mu - \frac{1}{2}\sigma^2}.$$

As before, the optimal A can be found using eqns. (55) and (23). Simplifying by setting $\mu = 0$, a discount factor almost like the previous ones is found:

(C.6) $Q = e^{-1/(k\varepsilon_\sigma^A)}$, where $\varepsilon_\sigma^A \equiv \frac{A \cdot \sigma_A}{\sigma}$, $k \equiv \frac{\beta - 1}{\beta - \frac{1}{2}}$.

Since $\beta > 1$ and $d\beta/d\sigma < 0$, we have $0 < k < 1$ and $dk/d\sigma < 0$. The factor k acts like a drag on the elasticity ε_σ^A, and it becomes more important the larger the uncertainty. The drag factor shows that in terms of the elasticities ($\varepsilon_\mu^A, \varepsilon_\rho^A$ and ε_σ^A), increasing the uncertainty is not as effective as increasing the expected price growth or decreasing the effective rate of depreciation. The reason is that the probability of low prices also increases if σ increases. To compensate for this inefficiency, the investment A ought to raise both σ and μ at the same time.

As a final curiosity, it can be noted that the effect of increasing σ and μ simultaneously becomes very simple if they are linked by the functional relationship $\mu = \frac{1}{2}\sigma^2$. Then it can be shown that eqn. (C.6) holds with $k \equiv 1$.

References

Baldwin, Richard E. and Paul R. Krugman, 1988, "Market Access and International Competition: A Simulation Study of 16K Random Access Memories", in R. Feenstra (ed.): *Empirical Studies of International Trade*, Massachusetts Institute of Technology

Dixit, Avinash K. and Joseph E. Stiglitz, 1977, "Monopolistic Competition and Optimum Product Diversity", *American Economic Review* 67, 297-308

Dixit, Avinash K., 1989, "Entry and Exit Decisions under Uncertainty", *Journal of Political Economy* 97, 3:620-638

Dixit, Avinash K., 1993, *The Art of Smooth Pasting*, vol. 55 in *Fundamentals of Pure and Applied Economics*, eds. Jacques Lesourne and Hugo Sonnenschein, Harwood Academic Publishers

Dixit, Avinash K. and Robert S. Pindyck, 1994, *Investment Under Uncertainty*, Princeton University Press

Krugman, Paul R., 1979, "Increasing Returns, Monopolistic Competition, and International Trade", *Journal of International Economics* 9, 4:469-479

Krugman, Paul R., 1980, "Scale Economies, Product Differentiation, and the Pattern of Trade", *Americal Economic Review* 70, 950-959

Leahy, John, 1993, "Investment in Competitive Equilibrium: The Optimality of Myopic Behaviour", *Quarterly Journal of Economics* 108, 1105-1133

CHAPTER 4

Endogenous Growth

1. Introduction

In the same way as Paul Krugman and some of his contemporaries broke the monopoly held by the standard trade theory around 1980, his namesake Paul Romer in 1983 initiated a process that was going to turn a number of results from neoclassical growth theory upside down.[1] In both cases, product differentiation and economies of scale played an important role. In particular, the Dixit-Stiglitz (1977) framework has been a cornerstone in many new growth models. However, the Dixit-Stiglitz model is a static model. This often makes it quite complicated to include a number of typical dynamic features of importance, such as fluctuating demand, irreversibility and improvements in product qualities.[2]

In this chapter, we attempt to circumvent many difficulties of this kind by endogenizing the growth rate in the "Dixit-Stiglitz-like" one-sector model with geometric Brownian prices in *Chapter 3*. It should be noted that many of the results that are obtained will hold also if the goods are true differentiated products, as in the Dixit-Stiglitz model. However, to avoid too many sets of assumptions floating around at the same time, and since product differentiation has not been discussed yet (see *Chapter 6*), we stick to homogeneous goods as in *Chapter 3*.

Before proceeding with the analytical treatment, it is convenient to review some empirical facts about growth as well as typical results obtained by new growth models.[3] Here we will

---

[1] See Dixit and Norman (1980), Krugman (1979), Krugman (1980) and Romer (1983).

[2] See Aghion and Howitt (1992) for a model with quality improvements, and Helpman (1992) for a survey of models of this kind.

[3] See Grossman and Helpman (1991), and Barro and Sala-i-Martin (1995) for surveys of the field.

just briefly discuss some major findings. In the introduction to their book, Grossman and Helpman (1991) refer to a number of empirical studies of interest. Typically, long-run growth tends to be higher:

- the larger the export share (Michaely, 1977, Feder, 1982)
- the lower the population growth (Baumol et al., 1989)
- the more scientists an industrialized country has (Romer, 1989)
- the larger the economy (Syrquin and Chenery, 1989)
- the larger the manufacturing share (Syrquin and Chenery, 1989)
- the smaller the government share of total consumption (Landau, 1983, Barro, 1989)
- the larger the government investment rates (Landau, 1983, Barro, 1989)
- the smaller the marginal tax rates for a fixed average tax rate (Koester and Kormendi, 1989)
- the more outward orientation or openness to trade (Syrquin and Chenery, 1989).

Some of these findings have been questioned in more recent studies. For example, Jones (1995a,b) argues that evidence from industrialized countries does not confirm the important role of scale effects, as the factor productivity growth rates in major OECD countries have not increased along with the number of scientists during the last decades.[4] On the other hand, Romer (1996) and others continue to emphasize that scale effects are among the main causes of economic growth.

New growth theory considers accumulation of physical capital to be of minor importance for economic growth. The main explanation is related to spillovers from aggregate accumulation of knowledge, a concept that at least goes back to Arrow (1962). Technical progress arises as an unintended byproduct of private investment decisions because knowledge cannot be protected effectively. When firms invest in R&D that bring about a new product, they contribute to a common pool of knowledge that other innovators can also extract from. The larger this pool, the better the conditions for long-run growth.

In all new growth models, the thing that really matters is the total amount of knowledge-accumulating activities. Normally, this can be characterized by some aggregate measure of

[4] See also Young (1995).

investment. With respect to normative issues, new growth concepts and models do not, as a general rule, predict as unambiguous results as may be indicated by the empirical analyses that were referred to. For example, the theory does not predict a strictly positive relationship between growth and openness to trade. If domestic research in sectors where a country has an international advantage is discouraged by import, its growth rate may go down if the trade barriers are lowered. Although the general rule in this particular case seems to be that openness encourages growth,[5] the example shows a typical feature of new growth models as well as other models based on imperfect competition: History can be "replicated" by formulating the model in a specific manner, but the predictive power may be weak.

By using a large number of dynamic models, Grossman and Helpman (1991) find that the equilibrium growth rate typically is higher the larger the economy, the more productive the research laboratories, the more patient the households, and the larger the perceived differentiation of products. It is also a typical result that the related parameters must exceed specific thresholds in order for innovation and endogenous growth to take place.

With respect to welfare, no clear-cut conclusion exists on whether a market left to itself will yield too little or too much growth, although the most common outcome seems to be that the growth rate will be too small. The reason for the general ambiguity is that there can be at least three effects from knowledge spillovers: Contemporary consumers and future innovators may gain, and contemporary producers lose.

Finally, if knowledge spillovers are hampered by national borders, it is clear that accidental events can be very important for growth. A country that happens to get a head start in the accumulation of knowledge may increase its lead over time.[6] However, as shown by e.g. Scherer (1982), and Bernstein and Nadiri (1988,1989), national borders do not have to be

---

[5] See e.g. Coe and Helpman (1993), and Harrison (1995) for empirical analyses, and Baldwin and Forslid (1996) for a theoretical discussion. By a new "q-theory" approach, somewhat related to the approach in this dissertation, Baldwin and Forslid identify a number of links between openness and growth.

[6] The importance of history is also typical for new economic geography models, as discussed by Krugman (1991).

decisive for the flow of information. Thus, country size should not be used uncritically when trying to measure the size of the knowledge pool. Indeed, when considering the growth performance in small countries like the Scandinavian or Benelux, it does not seem as if country size is of any major importance as long as the barriers to trade are small.

2. Non-technical description

The baseline setup that will be used is the one-sector model with geometric Brownian demand from *Chapter 3*, and it does not have to be spelled out again. However, the model is extended in a simple way: We shall let the growth rate of technology (productivity) depend on the investments in activities that are assumed to create spillovers. It was argued that the growth rate can be embodied additively in the discount rate ρ. By letting a part of ρ be an increasing function of growth-creating investment, growth becomes endogenous. The approach can be illustrated by returning to the wine example.

Suppose that there is only one production factor: labor services provided by the workers (who are also the consumers in the economy). As before, two types of irreversible investments are necessary before the wine can be consumed. It is the division of labor between these activities that is important with respect to growth, since the character of the two investments normally is quite different.

Assume that one of the investments, say harvesting grapes and making the wine ("patenting"), consists of an activity that increases the physical condition of the worker. In the other activity (i.e., getting the wine to the consumer, or "activation") he will be more effective the better his physical shape; however, this does not contribute to further improvements. Then it is obvious that the equilibrium growth rate of productivity, or the rate at which the physical condition of a representative worker improves, will be higher the more he is involved in the first activity.

Of course, there are decreasing returns in physical activities, so the growth rate will eventually go to zero in this case. This may not be the case, however, if we talk about knowledge that tends to spill over to contemporaries and future generations. If knowledge is allowed to accumulate in this aggregate manner, the growth rate will be higher the more investments in activities that (mainly) contribute to accumulation.

To get a more complete picture of this, two questions need to be answered: First, which investments are related to knowledge accumulation? Second, is it absolute or relative investments that matter?

As far as the first question is concerned, we have to consider the conceptual difference between the two investments. Normally, the patent is an investment in technology, like R&D, which brings about a drug, a car or a computer. We may denote this by the term *product innovation*. It establishes the technology for the specific product, which will also lay the basis for future generations. In many cases, the production process by which a specific good becomes available, will be a simpler effort. Therefore it is also reasonable to assume that the latter activity does not contribute to knowledge accumulation in the same way as the development of new products. There may be exceptions to this, however, so we also briefly discuss the alternative case, in which *process innovation* is assumed to be the major growth-creating activity.

The second question is related to the previous discussion on empirical and theoretical results with respect to country size. The size of the economy will be important, since there is more knowledge to spill over, the greater the aggregate innovation. However, since information does not flow perfectly within an economy, and normally less perfectly the larger its size, relative size will also matter. The conditions for spillovers between a fixed number of innovators are probably better in a small community, where they are more likely to get in touch by coincidence. Therefore a measure of spillover ought to be strictly concave as a function of the size of the economy.

As already noted, what is meant by "size", is also important. With international trade, information may accompany products across borders, and the effective size with respect to spillovers can be far larger than the size of the country. In any case, our objective is not to answer questions on how to define an economy or the scope of spillovers in general, so we will put such issues aside. The model is simple and stylistic, and will be used just to highlight some points of particular interest.

3. Endogenous growth

Assume that growth is induced by knowledge accumulation stemming from product innovation. Such activities are located in the patent sector, where new products are created by some kind of R&D. What matters for growth, are the total expenditures on patents. In other words, there exists an increasing function $g = g(A_{tot})$, where A_{tot} is the total rate of patent expenditures in terms of labor, and

(1) $$\rho = \rho_0 + g(A_{tot}).$$

Let us assume for now that the consumer has no time preferences, so ρ_0 is simply a constant exogenous rate of productivity growth or labor growth (or a combination). Thus, ρ is the total growth rate, also acting as the effective discount rate. Since each firm is small, it will neglect its own contribution to the growth rate. If follows that the firms will consider g and ρ to be fixed, although they are really endogenous variables, so equilibrium will be given by the same formulas as in *Chapter 3*.

With full employment, the total wage bill is $W = wL$, where w is the wage rate and L is the size of the labor force. The wage rate is normalized to unity in all periods. Then $W = L$, and costs are measured directly in terms of workers. The patent cost is A, and the production cost is C, as before. In equilibrium, $A^*_{tot} = N^* A$, and from eqn. (24) in *Chapter 3*, we have:

(2) $$A^*_{tot} = \frac{L}{\beta^*}.$$

Thus the growth rate will be higher the larger the market (L) and the greater the price variation (the smaller β^*). The asterisk on β refers to the fact that endogenous growth is embodied in it. As also discussed in *Chapter 3*, we have $d\beta^*/d\mu < 0$ and $d\beta^*/d\sigma < 0$. Hence, increasing the drift or uncertainty of demand at the firm level decreases β^*, and increases the share of patents to total investments. The reason is that the value of a patent will increase if β^* decreases, so the firms will wait longer before activating. This saves production costs due to growth, there is room for a larger number of firms, and growth is encouraged.

It can also be shown that $d\beta^*/d\rho > 0$, which implies that exogenous growth discourages endogenous growth.

In general, the relationships that demonstrate these conclusions are complex, as endogenous growth is embodied in ρ. More insight can be gained by looking at a simple example. Assume that the spillovers are characterized as follows:

(3) $g(A_{tot}) = \gamma A_{tot}$.

Here γ (> 0) is a measure of knowledge accumulation, or more indirectly: a productivity parameter in researching. Referring to the wine example once again, suppose that a fixed number of workers learn so much and exchange so much information each year, that the total amount of wine that can be produced increases by one percent per year. After two years, they have learned enough to increase production by approximately two percent. Eqn. (3) says that the same growth of productivity is obtained in one year by doubling the number of workers.

In the previous section we argued that spillover effects are probably strictly concave in the size of the economy. Eqn. (3) breaks this assumption by assuming that they are linear. This is done partly to make things simple, and partly due to the lack of data that could bring about a more realistic form of the g-function. In any case, this representation can be considered a first-order approximation in equilibrium for an appropriate choice of ρ_0 and γ.

Uncertainty is not important for the qualitative conclusions with respect to other parameters, so we set $\sigma = 0$ to obtain an analytical solution. (See *Appendix A* for a discussion of uncertainty.) We then have $\beta^* = \rho^* / \mu$, where ρ^* is the equilibrium effective discount rate. From eqns. (1), (2) and (3), the following relationship is obtained:

(4) $g^* = \frac{1}{2}\left(\sqrt{\rho_0^2 + 4\gamma\mu L} - \rho_0\right)$.

This growth rate is always positive, but it is smaller than what might be expected. According to eqn. (3), g increases linearly in A_{tot} and γ. However, the equilibrium rate increases only with the square root of L and γ. This is because some of the spillover gains are used to

increase present consumption. The level of consumption is lifted at the cost of growth of consumption.

From eqn. (4) it can be shown that $dg^*/d\rho_0 < 0$. This confirms the statement that the endogenous growth rate decreases by increasing the exogenous growth rate.

The results can also be related to standard measures of saving and investment: A patent (investment A) does not contribute to present consumption. It is an investment for the future. On the other hand, investment C represents production that is directly related to current consumption, so the ratio of patents to total investments can be regarded as a savings rate. Let us denote this ratio by s. From *Chapter 3*, eqn. (23), it is found to be:

$$(5) \qquad s \equiv \frac{A}{A + Q^* C} = \frac{1}{\beta^*}.$$

Eqn. (5) holds regardless of whether growth is endogenous or exogenous. However, in any case, s should be considered an endogenous variable, and we have three interpretations for β^*. At first sight, it is a measure of variation analogous to a static measure of product differentiation. As in static models, it shows up as an inverse measure of scale economies in equilibrium. In addition, it is also an inverse measure of saving in this model. With spillovers, it follows that more growth is obtained the larger the savings rate (the smaller β^*).

Finally, if we set $L \equiv 1$, the calculations above also apply to a setting where only relative investments in growth-creating activities matter.

4. Extensions

General comments

The model in the previous section can be extended along the same lines as in *Chapter 3*. A second, homogeneous good with no dynamic properties can be included, and the definition of a firm can be extended. Instead of assuming a "one-shot game" where the firm disappears immediately after its patent has been activated, we can interpret C as an investment in production capacity that makes the firm into an operating factory.

Chapter 3 also discussed extensions where β was endogenized via the patent cost A, assuming that future revenues could be raised by increasing A. The effectiveness was measured by the elasticity of β with respect to A. This extension also applies to endogenous growth. Then the equilibrium β^* becomes a result of two opposing forces: By increasing A, β^* decreases as each firm observes that its effective demand increases. (This can happen either as the drift μ or the volatility σ of the firm's demand increases or as its effective discount rate ρ decreases.) At the same time, β^* increases at an aggregate level, as all firms contribute to increase the aggregate growth rate (g) by a small amount each.

Process innovation

We have consistently assumed that growth arises from patents, as this is the most easy to justify. However, it does not always have to be the case. It could be that a patent takes the form of a simple ancillary investment like advertising, while growth may stem from process innovation. This can be illustrated by the "just-in-time" concept that was introduced in the car industry 10-15 years ago, where, apparently, there was much learning in production activities which contributed to aggregate growth of productivity.[7] New techniques were quickly adopted by other firms in the industry as well as by other industries. The concept is still being improved, and innovations are hard to protect. This has not only implied lower production costs; product quality and failure rates have also improved more generally. (The computer industry discussed in *Chapter 3* may be another example.)

If knowledge accumulation is a result of process innovation of this kind, most results are reversed. Growth is still encouraged by size, but e.g. demand uncertainty will make firms more hesitant to investing in production. The share of such investments decreases if the uncertainty increases, and there will be less accumulation of knowledge to induce growth.

Alternatively, if growth is a result of accumulation in both types of investments, the relative importance determines the final outcome with respect to growth. If the spillover effects are equally strong, dynamic parameters will not matter at all, and there is little more to be said.

---

[7] See Abegglen and Stalk (1985, ch. 5).

When considering the character of the two types of investments, it seems likely that patents are more important for knowledge accumulation and growth, but in the last resort, this is, of course, an empirical question.

Cost variation

So far, the analysis has been based on a model with geometric Brownian prices and constant costs. It is straight-forward to apply a similar approach to models with fluctuating costs and a fixed price. For example, let wine be replaced by oil (which needs millions of year to mature, so we make no major mistake by neglecting it). The patent cost A could be interpreted as a fixed exploration cost by which an oil reservoir is revealed, while C is a development cost that depends on where it is located; onshore or offshore, shallow water or deep water, etc. The exact nature is not important. The thing that matters, is that C fluctuates separately for each firm (or oil field), and that there is a probability of reducing it below the price by waiting.

As in *Chapter 3*, let us assume geometric Brownian cost processes with drift $-\mu$ and volatility σ. By endogenizing the growth rate in the same manner as above, we end up with equilibrium conditions that are very similar to the previous ones. We just have to replace the constant β^* (> 1) by $\alpha^* + 1$, where α^* (> 0) is the positive root of eqn. (12) in *Chapter 1, Note B*. For example, the expression analogous to eqn. (2) is:

$$(6) \qquad A_{tot} = \frac{L}{\alpha^* + 1}.$$

If, as before, the effect of knowledge accumulation is represented by eqn. (3), and uncertainty is left out by setting $\sigma = 0$, the growth rate in equilibrium becomes:

$$(7) \qquad g^* = \tfrac{1}{2}\left(\sqrt{(\mu + \rho_0)^2 + 4\gamma\mu L} - \rho_0 - \mu\right).$$

This is like eqn. (4) except that ρ_0 is replaced by $\rho_0 + \mu$. The qualitative results can be summarized as follows: The endogeneous growth rate is higher the larger the economy (L), the learning rate (μ) and the spillover parameter (γ), and the smaller the exogeneous growth rate (ρ_0).

Time preferences

Up until now, ρ_0 in eqn. (1) has been interpreted as a growth parameter, and time preferences have been excluded. The reason is that all parameters that build up to the effective discount rate then contribute to decrease the required amount of labor by waiting, and equilibrium formulas are easily derived due to the simplicity of eqn. (2).

Referring to the discussion in *Chapter 3*, the algebra gets more messy if ρ_0 is a rate of time preferences, since this affects the division of labor differently. Consider first the case with no spillovers, and let ρ_0 be a rate of subjective time preferences instead of an exogenous growth rate. Then the required amount of labor for activation does not go down by waiting, so the steady state condition becomes $N(A+C) = L$ as long as all patents are expected to be used. The savings rate becomes $s = A/(A+C) = R/(R+1)$, where $R = A/C$. This does not at all depend on the demand variables (μ and σ) or time preferences (ρ_0).

However, time preferences still play a role for endogenous growth. The more impatience, the less the required amount of labor for activation will go down for each patent, as firms will tend to activate earlier. For this reason the savings rate, which is positively related to the growth rate, ought to be larger the smaller the subjective rate of time preferences. *Appendix B* shows this for the deterministic case. If the spillovers are represented by eqns. (1) and (3), and ρ_0 denotes time preferences, equilibrium growth, g^*, is given implicitly by the following equation:

$$(8) \qquad \left[R\left(\frac{L\gamma}{g^*}-1\right)\right]^{1+\rho_0/g^*} = R\left(\frac{\rho_0+g^*}{\mu}-1\right).$$

As long as the model converges, it can be shown that g^* in this equation increases by R.[8] Hence, the initial cost share does not cancel out if the consumers are impatient: the more

[8] Note that the right-hand side of eqn. (8) is the optimal discount factor, so there are two requirements for convergence. First, the expressions on each side of the equation must be strictly positive. This requires that $\rho_0 + g^*$ ($= \rho^*$) is larger than μ. Second, the same expressions cannot exceed unity, since this implies profit from immediate activation.

research that is needed (the larger R), the greater the spillovers and the larger the growth rate. It can also be shown that g^* is decreasing in ρ_0, so growth is indeed discouraged by impatience. The growth rate in this case is also increasing in μ, L and γ, just like when ρ_0 is interpreted as exogenous growth.

The shape of g^* as a function of R is plotted in Figure 1 for the two sets of assumptions. The straight horizontal line corresponds to eqn. (4), where g^* is independent of R. The lower curve corresponds to eqn. (8), which assumes that all patents are used. As discussed in *Chapter 3*, the latter holds if $\mu > \frac{1}{2}\sigma^2$. However, *Chapter 3* also argued that the aggregate division of labor between patenting and production will be closer by the two interpretations of ρ_0 if σ increases (so that $\mu > \frac{1}{2}\sigma^2$ no longer holds). Then the growth rate increases for the time preference interpretation, as indicated by the arrows in Figure 1. Similarly, we should expect something in between the two curves if ρ_0 were a combination of exogenous growth and time preferences.

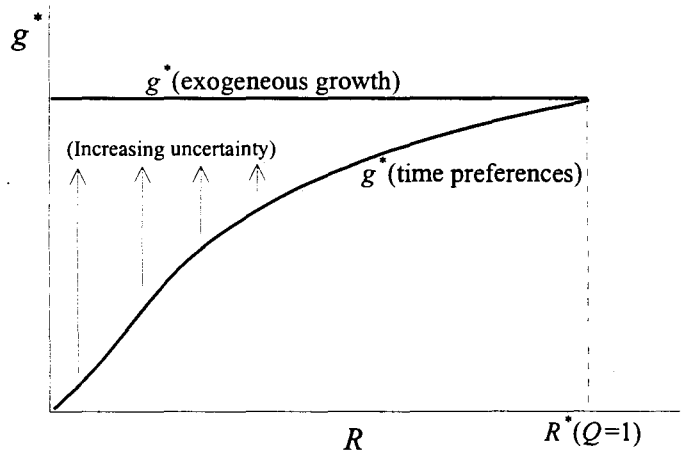

Figure 1. Endogenous Growth Rates.

It can be shown that g^* from eqn. (8) approaches g^* from eqn. (4) if R approaches the limit with immediate activation, where $Q = 1$. In Figure 1, this occurs for $R = R^*$. Finally, note that eqn. (8) collapses as expected to $g^* = \sqrt{\mu L \gamma}$ for $\rho_0 = 0$, just like eqn. (4).

5. Final remarks

We started this chapter with a list of "empirical facts" about growth, and some typical results obtained by new growth models. Considering the simplicity of the model, it has confirmed a surprisingly large number of previous findings, in addition to shedding light on the relationship between dynamic uncertainty and growth. Let us sum up the finds:

First, the growth rate will be higher the larger the economy (L). However, we also observed endogenous effects that tend to decrease the importance of size, as some of the benefit is used to increase present consumption instead of growth.

Second, the growth rate will be higher the larger the spillover parameter (γ). This is somewhat analogous to the Grossman and Helpman (1991, ch. 4) model with rising product quality. They measure the step size on a "quality ladder" by a specific parameter, and find (not surprisingly) that there is more growth the larger its value. In effect, γ does the same thing: it is an exogenous parameter that increases the gap between successive product generations.

Third, endogenous growth is smaller the larger the exogenous growth rate. This is also a typical result, and it is particularly interesting as far as developing countries are concerned, since one interpretation of exogenous growth can be quantitative growth of the labor force.

Fourth, the growth rate is larger the more patient the agents, as in almost every new growth model. However, in a sense, it can be said that time preferences are not as harmful for endogenous growth as is exogeneous growth, since time preferences have less influence on the division of labor between the two types of investment.

Fifth, the model can indirectly be used to argue that the growth rate will be higher the larger the manufacturing share; i.e., the share of the economy characterized by the kind of economies of scale that have been studied. To simplify, we did not consider more than one sector. If the economy is extended to two sectors along the lines of *Chapter 3*, it is clear that the growth rate will be higher for a given total size the larger the manufacturing share. From a statistical point of view, we also argued that the conditions for knowledge accumulation and

accidental spillover effects are better the larger the share of innovators in the economy. If such forces are active, the conclusion is strengthened.

Sixth, the growth rate is higher the larger the expected growth or uncertainty of demand at the firm-level (μ or σ). Such parameters give rise to dynamic economies of scale. In a sense, we obtain similar results as in new growth models with static product differentiation, saying that growth is caused by differentiation and economies of scale.

From the latter result we can also deduce a result of importance to policy. With demand uncertainty ($\sigma > 0$) it is clear that *ex post* profits will differ highly although expected profits are zero at time of entry. One matter of interest is what the government will do with firms that make a lot of profits. If profits are heavily taxed, typically by progressive tax rates, the firm's value of waiting decreases, as an unproportionately large part of the benefit is lost. The result in equilibrium will be the same as if β^* increases: A larger share of total resources will be used for production, and the growth rate decreases along with the savings rate. This also fits in well with the empirical results from the introduction.

Acknowledgements

I am grateful to Victor Norman for valuable suggestions and inspiring discussions on the research idea in this chapter, and to Terje Lensberg for helpful comments on a preliminary version.

APPENDIX A: Growth with uncertainty

To study how the growth rate in equilibrium is affected by uncertainty, let us start with the following equations (leaving out asterisks on optimal values):

(A.1) $\frac{1}{2}\sigma^2\beta(\beta-1)+\mu\beta-\rho=0$

(A.2) $A_{tot}=\frac{L}{\beta}$

(A.3) $\rho(A_{tot})=\rho_0+\gamma A_{tot}$.

Eqn. (A.1) is the familiar quadratics; (A.2) and (A.3) follow from the text. By inserting (A.2) and (A.3) into (A.1), we have:

(A.4) $\Psi(\beta,\sigma)\equiv\frac{1}{2}\sigma^2\beta^2(\beta-1)+\mu\beta^2-\rho_0\beta-\gamma L=0$.

This third degree equation can be solved analytically, but the algebra gets terrible. Instead, implicit differentiation gives

(A.5) $\frac{\partial\Psi}{\partial\beta}=\frac{3}{2}\sigma^2\beta^2+2(\mu-\frac{1}{2}\sigma^2)\beta-\rho_0$,

and

(A.6) $\frac{\partial\Psi}{\partial\sigma}=\beta^2\sigma(\beta-1)$.

Thus we have

(A.7) $\frac{d\beta}{d\sigma}=-\frac{\partial\Psi/\partial\sigma}{\partial\Psi/\partial\beta}\overset{(A.4)}{=}\frac{-\beta^3\sigma(\beta-1)}{\frac{3}{2}\sigma^2\beta^3+2(\rho_0\beta+\gamma L-\frac{1}{2}\sigma^2\beta^3)-\rho_0\beta}=\frac{-\beta^3\sigma(\beta-1)}{\frac{1}{2}\sigma^2\beta^3+\rho_0\beta+2\gamma L}<0$.

The inequality follows as $\beta > 1$. As expected, β is decreasing in the uncertainty, and the growth rate increases if the externalities are located in the patent sector. Due to the complexity of (A.7), it seems difficult to get any further with the study of uncertainty in this case without turning to numerical experiments.

APPENDIX B: Growth with time preferences

If ρ_0 is interpreted as a rate of time preferences instead of an exogenous growth rate, the steady state condition in the deterministic version of the model can be stated as follows:

(B.1) $NA + Ne^{-g^* T^*} C = L$.

The first term on the left-hand side is the rate of entry times labor requirement for each patent. The second term contains the similar requirement for production. Since productivity grows at rate g^*, the labor requirement in case of immediate production (C) is reduced by the exponential term, as T^* is the delay. With full employment, and costs measured in terms of labor, the two terms sum up to the total labor force (L). Then total investments in patents can be written as

(B.2) $$A_{tot} = \frac{RL}{R + e^{-g^* T^*}} = \frac{g^*}{\gamma},$$

where $R = A/C$ as before, and the last transition follows from eqn. (3) in the text. The effective discount rate is

(B.3) $\rho^* = \rho_0 + g^*$,

and the equilibrium discount factor follows from *Chapter 3*, eqn. (23):

(B.4) $$e^{-\rho^* T^*} = R\left(\frac{\rho^*}{\mu} - 1\right).$$

By combining (B.2), (B.3) and (B.4), eqn. (8) in the text is obtained.

References

Abegglen, James C. and George Stalk Jr., 1985, *Kaisha, The Japanese Corporation*, New York: Basic Books

Arrow, Kenneth J., 1962, "The Economic Implications of Learning by Doing", *Review of Economic Studies* 29, 155-173

Aghion, Philippe and Peter Howitt, 1992, "A Model of Growth Through Creative Destruction", *Econometrica* 60, 323-351

Baldwin, Richard and Rikard Forslid, 1996, "Trade Liberalization and Endogenous Growth: A q-Theory Approach", *NBER Working Paper* 5549

Barro, Robert J., 1989, "Economic Growth in a Cross Section of Countries", *Quarterly Journal of Economics* 106, 2 (May), 407-443

Barro, Robert J. and Xavier Sala-i-Martin, 1995, *Economic Growth*, McGraw Hill

Baumol, William J., Batey Blackman, Anne Sue, and Edward J. Wolff, 1989, *Productivity and American Leadership: The Long View*, MIT Press

Bernstein, Jeffrey I. and M. Ishaq Nadiri, 1988, "Interindustry R&D Spillovers, Rates of Return, and Production in High-Technology Industries", *American Economic Review* 78 (Paper and Proceedings), 429-434

Bernstein, Jeffrey I. and M. Ishaq Nadiri, 1989, "Research and Development and Intra-Industry Spillovers: An Empirical Application of Dynamic Duality", *Review of Economic Studies* 56, 249-268

Coe, D. and E. Helpman, 1993, "International R&D Spillovers", *CEPR Discussion Paper* 840

Dixit, Avinash K. and Victor Norman, 1980, *Theory of International Trade*, Cambridge University Press

Dixit, Avinash K. and Joseph E. Stiglitz, 1977, "Monopolistic Competition and Optimum Product Diversity", *American Economic Review* 67, 297-308

Feder, Gershon, 1982, "On Exports and Economic Growth", *Journal of Development Economics* 12, 59-73

Grossman, Gene M. and Elhanan Helpman, 1991, *Innovation and Growth in the Global Economy*, MIT Press, Cambridge, Massachusetts

Helpman, Elhanan, 1992, "Endogenous Macroeconomic Growth Theory", *European Economic Review* 36, 237-267

Harrison, Ann, 1995, "Openness and Growth: A Time-Series, Cross-Country Analysis for Developing Countries", *NBER Working Paper* 5221

Jones, Charles I., 1995a, "Time Series Tests of Endogenous Growth Models", *Quarterly Journal of Economics* (May), 495-525

Jones, Charles I., 1995b, "R&D based Models of Economic Growth", *Journal of Political Economy* (August), 759-784

Koester, Reinhard B. and Roger C. Kormendi, 1989, "Taxation, Aggregate Activity and Economic Growth: Cross-Country Evidence on Some Supply Side Hypotheses", *Economic Inquiry* 27, 367-386

Krugman, Paul R., 1979, "Increasing Returns, Monopolistic Competition, and International Trade", *Journal of International Economics* 9, 4:469-479

Krugman, Paul R., 1980, "Scale Economies, Product Differentiation, and the Pattern of Trade", *American Economic Review* 70, 950-959

Krugman, Paul R., 1991, *Geography and Trade*, MIT Press

Landau, David, 1983, "Government Expenditure and Economic Growth: A Cross-Country Study", *Southern Economic Journal* 49, 783-792

Michaely, Michael, 1977, "Exports and Growth: An Empirical Investigation", *Journal of Development Economics* 40, 49-53

Romer, Paul, 1983, *Dynamic Competitive Equilibria with Externalities, Increasing Returns and Long Run Growth*, Ph.D. dissertation, University of Chicago

Romer, Paul, 1989, "What Determines the Rate of Growth and Technical Change?", *The World Bank Policy, Planning and Research Working Paper* WPS 279, Washington D.C.

Romer, Paul, 1996, "Why, indeed, in America? Theory, History, and the Origins of Modern Economic Growth", *NBER Working Paper* 5443

Scherer, F. M., 1982, "Interindustry Technology Flows and Productivity Growth", *Review of Economics and Statistics* 64, 627-634

Syrquin, Moshe and Hollis Chenery, 1989, "Three Decades of Industrialization", *World Bank Economic Review* 3, 145-181

Young, Alwyn, 1995, "Growth withouth Scale Effects", *NBER Working Paper* 5211 (August)

CHAPTER 5

Agglomeration with True Externalities

1. Introduction

One of the most remarkable characteristics of this century is the development toward concentration of human activity. For example, roughly five percent of the world's population lived in cities with a population exceeding 100,000 at the beginning of the century, whereas today almost fifty percent of us do. Of course, the total world population has also increased, but that just explains a small part of this dramatic change. Among economists, the development has brought forward a new, formal approach to the study of agglomeration. The new theory is often called "new economic geography", although "geographical economics" seems to be a better name.[1]

Agglomeration, which in broad terms can be defined as geographical concentration that cannot be traced directly to immobile resources, occurs at many levels. It may be urbanization, which includes a rich variety of activities, or concentration of specific industries. At the industry level, agglomeration seems to be particularly common for those industries that are based on new technology, although Krugman (1991a) argues that this dimension tends to be exaggerated. Industrialized production of textiles used to take place within just a few regions in Britain and other parts of Europe in its early days; for a long time, most of the car industry was localized in the Northeastern part of the United States and in Central Europe; and the computer industry in Silicon Valley has been a popular example among currently advanced industries.

---

[1] "Location and trade" is also a common term for this theory. See Fujita and Thisse (1996), Ottaviano and Puga (1997), and Fujita, Krugman and Venables (1997) for surveys. A simple, yet thorough, discussion is also given by Venables (1996).

However, the main part of the textile production has now moved to Asian regions, the car industry has spread to newly industrialized countries in Asia as well as to other parts of the world, and the Silicon Valley is not as important for the computer industry as it used to be. Nevertheless, industries continue to agglomerate even if production is established in new locations and larger numbers. Thus, the underlying forces do not vanish.

Irreversibility is obviously important for the rise and decline of agglomerated economies, but few models in the new economic geography literature address this issue. In particular, dynamic uncertainty is more or less non-existent. Certainly, the existence of multiple equilibria and the importance of history are emphasized, but not as a result of dynamic uncertainty.

The fact that little attention has been paid to uncertainty and irreversibility becomes even more surprising when looking at simple statistics about the geographical sustainability of agglomeration. Dicken and Lloydd (1990:165) claim that close to 80 percent of all manufacturing investments in the advanced nations consist of expanding existing plants. Thus, there is not only a tendency towards agglomeration; as we all know, there are also strong forces that tie an agglomerated economy to the particular location where it happened to be established.

To obtain agglomeration there must be cost or demand linkages. That is, costs must decrease or demand must increase by extending the number of firms. Such linkages can take many forms, based on pecuniary (market) externalities or true externalities. This chapter focuses on true externalities, while pecuniary externalities are discussed in *Chapter 7*.

The approach will be similar to that of *Chapter 4*, which had as a starting point the model with fluctuating demand and fixed costs from *Chapter 3*. The effective discount rate (ρ) was endogenized, assuming that growth of technology was an increasing function of aggregate patent investments.

In the study of agglomeration that follows, models with fluctuating costs and fixed prices are at least as interesting. Thus we will initially consider a homogeneous good, like computer chips, that pays a fixed price per byte, but for which production costs fluctuate separately for each generation.

Furthermore, we will introduce externalities via the production cost process, assuming that expected learning increases with the total level of production in the economy. The explanation is some kind of knowledge spillovers, but we are not concerned about the precise description of these. One possibility is that information is exchanged through workers who bring along their experience when changing employers.[2]

In addition to the industry that has been mentioned, we envisage a perfectly competitive industry with no externalities. As in Krugman (1991b) and many similar models, the consumption shares are fixed, and we adopt the term manufacturing for the industry with (dynamic) economies of scale, and agriculture for the other industry. Furthermore, the two goods are produced by specific factors: workers and farmers.

The economy consists of two regions separated by a trade cost barrier that only applies to manufactured goods. Farming is tied to the land so the farmers do not move, whereas the workers are attracted to the region that offers the higher real wage. The main role of agriculture is that the farmers make up a demand for manufactured goods. The objective is to discuss the stability of an equilibrium with industrial agglomeration; i.e., a "core-periphery" pattern with all manufacturing in one region.

The outcome will depend on the following opposing forces: If a core-periphery pattern has been established, consumer prices will be higher in the periphery. This will make it attractive for a new firm to "defect" by locating there instead of in the core. However, costs will also be higher for a defecting firm, partly because the workers must be compensated for a higher cost level in the periphery, and partly because the defecting firm will not gain from knowledge spillovers to the same extent as firms in the core.

As in many other new economic geography models, we find that agglomeration is most stable for large manufacturing shares and for intermediate trade costs. The relationship between agglomeration and parameters like firm-specific uncertainty and aggregate growth is complex, and it could be ambiguous. However, it is concluded that agglomeration is encouraged by uncertainty and discouraged by growth for the most realistic assumptions.

---

[2] See Midelfart Knarvik (1995) for a location model with such spillovers.

The rest of the chapter is structured as follows: First the main equations that characterize manufacturing are listed (Section 2). The setup will be the same as in the *Chapter 3* model because of the large-group assumption. Since each firm is small, it will consider learning to be fixed in equilibrium although it is, in fact, endogenous. As in the growth model, it follows that externalities can be characterized separately. Then we introduce trade costs (Section 3), and a specific representation of externalities (Section 4). This completes the framework that is required for a discussion of agglomeration (Section 5). After this discussion, some extensions are noted (Section 6) before conclusions are drawn (Section 7).

2. The model

Agriculture needs no discussion, as it is produced by a specific factor with constant returns to scale, and can be traded at no cost. Its price will then be the same in both regions, and we just have to specify consumption shares for later use. As in *Chapter 3*, let the manufacturing consumption share be π, and the agricultural share be 1 - π.

In the following, assume that all firms in manufacturing are concentrated in a core region. The basic setup for this industry is the *Chapter 3* model with patent rights and firm-specific, geometric Brownian production costs:

(1) $dC = -\mu C dt + \sigma C dz$.

The elasticity of the discount factor, which is a measure of cost variation, is then the positive root (α) of the following quadratic equation in x:

(2) $\frac{1}{2}\sigma^2 x(x+1) + \mu x - \rho = 0$.

The expected profit for a firm contemplating entry is

(3) $\Pi = (P - C)Q - A$,

where $Q = (C/\overline{C})^{\alpha}$ is the expected discount factor, and P is the price in terms of Q. The optimal cost at which to invest becomes

(4) $C = \frac{\alpha}{\alpha + 1} P$,

and the optimal discount factor

(5) $Q=(\overline{R}\alpha)^{\alpha/(\alpha+1)}$, where $\overline{R}\equiv A/\overline{C}$,

follows from a free entry requirement. For simplicity, ρ is considered to be a fixed, exogenous growth rate of technology. Then the equilibrium rate of entry becomes

(6) $N=\dfrac{L}{A(\alpha+1)}$,

where L is the total number of workers, and the wage rate has been normalized to unity. For free entry and no expected profits to apply, we must have:

(7) $\overline{R}\alpha\le 1$.

Later we shall also need the equilibrium price, which from previous equations can be shown to be:

(8) $P=\dfrac{(\alpha+1)A}{(\overline{R}\alpha)^{\alpha/(\alpha+1)}}$.

3. Trade costs

The trade barrier is represented by an *ad valorem* ("iceberg") trade cost for manufactured goods. More precisely, let τ be a constant between zero and one. If $\tau=1$, there are no trade costs, and $\tau=0$ implies they are infinite. Since prices are measured in terms of Q, we can imagine that if a quantity (represented by the discount factor) Q_1 is shipped from the core, only a fraction $Q_2=\tau Q_1$ arrives in the periphery. It follows that the consumer price will be a factor $1/\tau$ higher in the periphery than in the core.

Alternatively, τ can represent a delay from production to consumption. This is most easily observed in the deterministic case, where $Q_1=e^{-\rho T_1}$ and $Q_2=e^{-\rho T_2}$. Then T_1 is the delay from patenting to production, with both activities taking place in the core. Similarly, T_2 is the delay from patenting in the core to consumption in the periphery. The difference is $T_2-T_1=\ln(1/\tau)/\rho$, which is a nonnegative constant that equals zero if $\tau=1$, and that goes to

infinity as τ approaches zero. Hence, a constant τ can be interpreted as if it takes a fixed time to transport the good from the core to the periphery. The effect of the trade barrier is that consumers in the periphery do not get access to products of the same quality or effective size as those who live in the core, due to growth that is embodied in new products.

In principle, we can also think of τ as a delay if there is uncertainty. This can be seen by use of a relationship that was derived in *Chapter 1* for various discount factors. Let $X_0 < X_1 < X_2$ (or $X_0 > X_1 > X_2$) be three states of a stochastic (price or cost) process with the familiar properties, while $Q(X_i, X_j)$ is the expected discount factor going from X_i to X_j for the first time. Then we have $Q(X_0, X_2) = Q(X_0, X_1) \cdot Q(X_1, X_2)$. Hence, by setting $Q_1 = Q(X_0, X_1)$, $Q_2 = Q(X_0, X_2)$, and $\tau = Q(X_1, X_2)$, it is observed that a constant τ acts like a fixed discount factor, which in due term corresponds to a fixed delay.

In this model there is no firm-specific process by which the quality of a traded product changes. Therefore the loss rate by delayed consumption equals the growth rate ρ. If the product had been wine, however, the loss might not be that large, as wine matures. Instead of thinking of the trade cost as if an iceberg shrinks, we might speak of a "drift bottle cost", imagining that a bottle needs time to cross an ocean but that the wine matures on the way.

4. Externalities

A simple specification of externalities follows by setting

(9) $\mu(C_{tot}) = \gamma C_{tot}$,

where γ (> 0) is a constant, and C_{tot} is the aggregate rate of production in terms of labor. As N firms each period produce their specific good at expected cost QC, we have $C_{tot} = NQC$ in equilibrium. The zero profit condition yields $QC = \alpha A$ by setting expected profits to zero in (3) and using (4). Then insert the rate of entry from (6), and the expected rate of learning in equilibrium becomes:

(10) $\mu = \dfrac{\gamma L \alpha}{\alpha + 1}$.

This can be inserted into eqn. (2) and solved for α. For simplicity, set $\sigma = 0$. Then $\alpha = \rho/\mu$, and eqn. (10) gives the following solution for μ:

$$(11) \qquad \mu = \frac{\rho}{2}\left(\sqrt{1+4\gamma L/\rho}-1\right).$$

This resembles the growth rate in *Chapter 4*, eqn. (4). Here the equilibrium rate of learning grows with the square root of γL although it is linear in the total rate of production. As in the growth model, the spillovers change the allocation of investments. Because the learning rate increases, it is optimal to wait for a lower cost before activating. This causes a shift away from production, and the net effect on learning becomes less than linear. We also find

$$(12) \qquad \alpha = \frac{2}{\sqrt{1+4\gamma L/\rho}-1},$$

noting that this is decreasing in γL and increasing in ρ. In *Appendix A* it is also shown that α is decreasing in cost uncertainty (σ). These characteristics are used below when discussing how size, growth and uncertainty affect the stability of an equilibrium with agglomeration.

5. Agglomeration

Since the farmers are immobile, there will always be a demand for manufactured goods in both regions. Thus, since zero trade costs ensure equal prices for agriculture while there are positive trade costs in manufacturing, the latter industry will also be spread as long as the linkages are weak. To study how strong linkages can establish a sufficient force for agglomeration, assume, as before, that a core-periphery pattern is the initial situation. Further, assume that a manufacturing firm must place both of its investments in the same region.

Since manufactured goods are homogeneous (like agriculture), a small firm that locates in the periphery will only sell in its home market. Trade costs imply that those who live in the periphery must pay a factor $(1/\tau)$ more for manufactured goods than those who live in the core. Therefore the product price that the defecting firm can obtain is $P_d = P/\tau$.

Since manufactured goods represent a share π of total consumption, the cost level will be a factor τ^{π} higher in the periphery than in the core. To attract workers, the defecting firm must

increase the wage rate accordingly. To acquire a patent, A workers will be needed for one period, and they must be compensated for having to buy all goods in the periphery during this period. The wage rate has been normalized to unity in the core, so it follows that the patent cost in the periphery will be $A_d = \tau^{\pi} A$. Similarly, the initial production cost is $\overline{C}_d = \tau^{-\pi}\overline{C}$.

Spillovers are also hampered by the trade barrier, and they ought to be more hampered the larger the barrier. A simple specification follows by letting the elasticity of the discount factor that applies to the defecting firm be:

(13) $\alpha_d = \alpha / \tau$.

With this specification, the defecting firm gets access to spillovers just like firms in the core if there is no trade cost ($\tau = 1$). If the trade cost is infinite ($\tau = 0$), then $\alpha_d = \infty$. This implies no value from waiting in the periphery, as there are no activities to learn from. Note also that the specification yields $\alpha_d = \rho/(\tau\mu)$ in the deterministic case, implying that the learning rate in the periphery is proportional to τ, which is actually the inverse of a trade cost.

If all other firms are located in the core and gain no profits in expectation, a new firm would locate in the periphery if its expected profit were positive. Therefore we must have

(14) $\Pi_d = (P_d - C_d)Q_d - A_d < 0$,

for agglomeration to be stable. Here C_d is the cost at which the defecting firm decides to activate, and Q_d is the discount factor that applies. The optimal cost follows by setting $P = P_d$ and $\alpha = \alpha_d$ in eqn. (4); i.e.,

(15) $C_d = \frac{\alpha}{\alpha + \tau} \cdot \frac{P}{\tau}$.

The optimal discount factor is $Q_d = (C_d / \overline{C}_d)^{\alpha_d}$, which by use of (8) and (15) gives

(16) $Q_d = \left(\tau^{-(1-\pi)} \cdot \frac{\alpha + 1}{\alpha + \tau} \cdot \left(\overline{R}\alpha \right)^{1/(\alpha+1)} \right)^{\alpha/\tau}$

in equilibrium. If waiting applies ($Q_d < 1$), we find, by inserting (15) and (16) into (14), that agglomeration is stable if K_1 defined by the following expression is less than one:

$$(17) \qquad K_1 = \tau^{\pi-\alpha(1-\pi)/\tau}\left(\frac{\alpha+1}{\alpha+\tau}\right)^{(\tau+\alpha)/\tau}\left(\overline{R}\alpha\right)^{\alpha(1-\tau)/[\tau(1+\alpha)]}.$$

However, there is less value from waiting in the periphery than in the core, so waiting does not have to be optimal for the defecting firm even if it applies in the core. There are two reasons for this. First, the cost variation is smaller in the periphery than in the core ($\alpha_d > \alpha$). Second, the product price is a factor $1/\tau$ higher, while the cost level is only a factor $\tau^{-\pi}$ higher. The implied difference in value of waiting is reflected in the discount factors (5) and (16), as $Q < Q_d$.

If waiting is not optimal in the periphery (i.e., if $Q_d = 1$), then K_1 must be replaced by $K_2 = (P_d - \overline{C}_d)/A_d$, or:

$$(18) \qquad K_2 = \tau^{-(1-\pi)}(\alpha+1)\left(\overline{R}\alpha\right)^{-\alpha/(\alpha+1)} - 1/\overline{R}.$$

The overall stability can be characterized as a function $K(\pi,\tau,\alpha,\overline{R})$ that equals K_1 if $Q_d < 1$, and K_2 otherwise. It can be shown that $K_1 = \tau/(\overline{R}\alpha) = K_2$ if $Q_d = 1$, so K is continuous, as expected. The following properties of K are derived:

First, $dK/d\pi < 0$, implying that agglomeration becomes more likely or stable the larger the consumption share. This is due to cost differences. The more that is spent on manufactures, the larger the difference in cost levels, and the more a defecting firm must pay to attract workers.

Furthermore, $dK/d\overline{R} > 0$, which implies that agglomeration becomes more likely, the larger the initial share of learning activities in the cost function (i.e., the smaller A relative to $\overline{C}$). This is an effect of dynamic economies of scale: the larger the patent cost, the less competition.

The K-function is so complex that it is hard to track down its relationship with α. Numerical analysis shows that it is ambiguous if waiting applies in the periphery (i.e., if K_1 is to be used). However, it can be shown that $dK_2/d\alpha > 0$, and it will be argued below that $dK/d\alpha > 0$ can also be expected in most cases. To see the economic implications, we must consider how α

depends on its underlying variables. We find that $d\alpha/d\rho > 0$ from eqn. (12), so agglomeration becomes less stable, the larger the growth rate for parameter combinations where $dK/d\alpha > 0$. On the other hand, $d\alpha/d(\gamma L) < 0$, so agglomeration becomes more stable the larger the economy (L) and the higher the measure of externalities (γ). Finally, it follows from *Appendix A* that agglomeration also becomes more stable the greater the cost uncertainty, as $d\alpha/d\sigma < 0$.

The relationship from τ to K is also too complex for analytical treatment. Numerical results show that it is ambiguous. Notice first that $K(\tau = 1) = 1$. Hence, location is irrelevant with no trade cost. The ambiguity in α and τ is seen from Figure 1, which plots four paths of $K(\tau)$ using different α-values. In all cases, $\overline{R} = 0.03$, and $\pi = 0.4$. The solid curves are most interesting, as the α-values for the other ones are extremely small (with accordingly extreme learning).

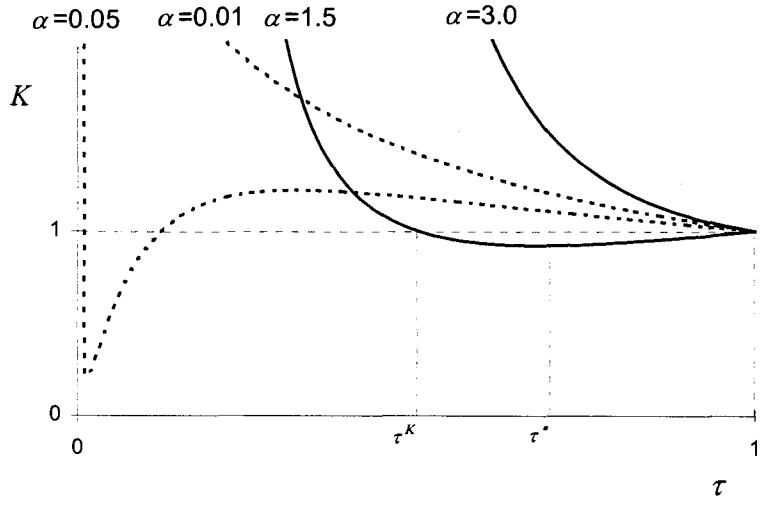

Figure 1. Stability of Agglomeration.

Agglomeration is never stable if there is not enough learning ($\alpha = 3.0$). However, agglomeration is stable if there is more learning ($\alpha = 1.5$), and trade costs are below some specific level ($\tau > \tau^K$). In that case agglomeration is most stable for some intermediate trade cost further below (τ^*). Similar U-shaped relationships are found in other models, like Krugman (1991b), and Krugman and Venables (1996), which are discussed in more detail in *Chapter 7*.

In the cited references, agglomeration is always stable if scale economies are very large. In this model agglomeration is never stable if the trade cost is sufficiently high. This can be observed by inspecting (16) and (18). The right-hand side of (16) approaches infinity as τ approaches zero, implying that waiting never applies in the periphery if the trade cost is very high. However, K_2 - which is the proper function to use in that case - also approaches infinity as τ approaches zero, so K is definitely larger than one for τ sufficiently small.

Subtle effects arise if α approaches zero, as shown by the curves for $\alpha = 0.05$ and $\alpha = 0.01$. This follows from the fact that the optimal discount factor (5) is U-shaped as a function of α. The underlying reason is as follows: If α is fairly large, e.g. as the uncertainty or the expected learning rate is small, the discount factor decreases by decreasing α as it becomes optimal to wait for a lower cost. However, there is a lower cost barrier at zero. If the uncertainty (or the expected learning rate) gets extremely high, the expected increase in the speed (which is increasing in σ and μ) down to the optimal cost dominates over increase in distance. Then the discount factor rises again. In this model, such effects shift the allocation of investments towards production, with corresponding implications for externalities and stability of agglomeration.

If $\alpha = 0.05$, agglomeration becomes stable for intermediate trade costs, while being unstable for small trade costs (because the spillovers to the periphery are large enough to cope with the core). Once again, however, agglomeration is unstable for very large trade costs.

Agglomeration is never stable if $\alpha = 0.01$, and we are back to the situation as for small cost variation ($\alpha = 3.0$). However, the reason is completely different. Now there is so much learning that firms in both regions are able to produce at very low costs. Thus the initial cost disadvantage by locating in the periphery is compensated, and the benefit from higher prices in the periphery dominates for any level of trade costs.

The two latter cases illustrate what may happen if assumptions are brought to the extremes, and they do not seem to be very realistic. For example, $\alpha = 0.01$ corresponds to a learning rate that is a hundred times as large as the growth rate in the deterministic case. If similar curve shapes did appear for much higher α-values by changing other parameters, this would

not be a valid argument, but numerical experiments indicate that the result remains, even with drastic changes in $\overline{R}$ and π.

In any case, Figure 1 supports the main conclusion from many models in the new economic geography literature: Agglomeration is more likely for intermediate trade costs than for very large or very small trade costs.

Finally, note that the requirement (7) implicitly defines a set of thresholds for dynamics to matter.[3] For example, learning is not important enough for waiting to be optimal, even in the core, if the economy is not large enough (i.e., if L is so small that $\overline{R}\alpha > 1$). Such cases must be studied by other tools, but agglomeration does not seem likely. By numerical experiments it can be shown that there are similar thresholds for π and $\overline{R}$. If the manufacturing share is very small or the ratio of the patent cost to the initial production cost is very large, agglomeration is unstable for all trade cost levels. In broad terms, externalities are not present in a large enough part of the economy.

6. Extensions

All extensions that were discussed in *Chapter 3* and *Chapter 4* also seem to apply to this model: The firm can be interpreted as an infinite-lived factory, a Poisson "death" process for operating factories might be included, the patent cost and the growth rate can be endogenized etc. Most of these extensions are more or less straight-forward to develop, so we restrict to a few that are of particular interest.

Time preferences

As discussed in previous chapters, ρ can be defined as a subjective rate of time preferences instead of a growth rate. This changes the steady-stated condition from $N(A+QC) = L$ to $N(A+C) = L$ as long as all patents are expected to be used. If externalities are characterized by

[3] Such minimum levels are also common in new growth models; see e.g. Grossman and Helpman (1991).

eqn. (9), total learning-creating investment in steady-state becomes $C_{tot} = \gamma NC$. The relations that led up to eqn. (11), now yield:

$$(19) \qquad \mu = \gamma L\left(\frac{\alpha}{\left(\bar{R}\alpha\right)^{\alpha/(\alpha+1)} + \alpha}\right).$$

The equilibrium elasticity (α) is found by inserting this into eqn. (2), and it can be shown that the qualitative relationship between α and its underlying variables is the same as before.

The first term in the denominator of (19) is the discount factor (5), which is less than one if waiting applies. In the limit with Q = 1, eqn. (19) coincides with eqn. (10), and all equilibrium formulas will be the same as if ρ were a growth rate.

Chapter 3 concluded that all patents will be used with probability one as long as $\mu + \frac{1}{2}\sigma^2 > 0$. Thus, if μ is negative and the uncertainty is not large enough for this to hold, eqn. (19) is no longer valid. As this case does not seem very interesting, we do not pursue it any further.

Alternative characterizations of externalities

The behaviour of the model does not depend much on the specific characterization of externalities. To illustrate this, let (9) be replaced by

$$(20) \qquad \mu(C_{tot}) = \mu_{\max} \cdot \frac{\gamma C_{tot}}{\gamma C_{tot} + 1},$$

where $\mu_{\max}$ is the maximum expected learning rate. A similar exercise as in Section 4 ends up with the following learning rate in the deterministic case:

$$(21) \qquad \mu = \frac{\rho}{2}\left(\sqrt{(1+\gamma L)^2 + 4\gamma L\mu_{\max}/\rho} - (1+\gamma L)\right).$$

It can be shown that α also in this case decreases in γ and L, and increases in ρ. It also decreases in $\mu_{\max}$. Thus, the qualitative results are not changed by the new specification.

More surprisingly, neither does the behaviour of the model depend highly on where the spillovers are located. If μ is a function of aggregate patent investments, like $\mu = \gamma A_{tot}$, eqn. (6) yields

(22) $\mu = \gamma L / (\alpha + 1)$

in equilibrium. As $\alpha = \rho/\mu$, this also gives $\mu = \gamma L - \rho$, and:

(23) $$\alpha = \frac{1}{\gamma L / \rho - 1}.$$

It follows that α is decreasing in (γL) and increasing in ρ. As shown in *Appendix B*, it is also decreasing in uncertainty (σ). Hence, as long as α is not extremely small, agglomeration will be more stable the larger the economy or the manufacturing consumption share, and the higher the uncertainty. The fact that uncertainty encourages agglomeration in both cases can be explained as follows:

If learning stems from patenting (A), then the expected production cost (QC) at the firm level decreases by uncertainty for two reasons (as long as α is not very small): First, a smaller production cost would be more likely even with no spillovers. This gives room for a larger number of firms, and with spillovers we get an additional boost downwards as more patents increase the learning rate. Total spillovers as well as economies of scale increase, and agglomeration becomes more stable.

If learning originates from production (C), the forces arising from increased uncertainty are opposing. A smaller production cost becomes more likely at the firm level, so aggregate production and therefore also the total amount of spillovers, decrease. However, more uncertainty also gives more dynamic economies of scale, which encourage agglomeration. The latter effect dominates for the most realistic α-values.

Price variation

The same approach as above can be applied to the alternative models with geometric Brownian prices and fixed costs. In this case, spillovers from R&D could be represented by the expected rate of quality improvements for existing products or patents (also denoted by μ).

It seems likely that this kind of spillovers will be hampered by trade barriers somewhat similar to cost spillovers.

Hence, starting with a core-periphery pattern, one can use exactly the same procedures as above to describe costs and benefits for a defecting firm. The only difference is that the representation of spillovers losses (13) ought to be replaced by $\beta_d = \beta / \tau$. Here β, which is the familiar positive root of eqn. (18) in *Chapter 3*, is the demand elasticity applying to firms in the core, while β_d applies to the defecting firm. If waiting is optimal in the periphery, it can be shown by similar arguments as in Section 5, that agglomeration is stable if

$$(24) \qquad \tilde{K}_1 = \tau^{1-\beta(1-\pi)/\tau} R^{(1-\tau)/\tau} \frac{(\beta-\tau)^{(\beta-\tau)/\tau}}{(\beta-1)^{(\beta-1)/\tau}}$$

is less than one. If waiting does not apply in the periphery, agglomeration is stable if

$$(25) \qquad \tilde{K}_2 = \tau^{-(1-\pi)} \beta \left(R(\beta-1)\right)^{-(\beta-1)/\beta} - 1/R$$

is less than one. The final criterion $\tilde{K}(\pi,\tau,\beta,R)$ that arises by joining $\tilde{K}_1$ and $\tilde{K}_2$ correctly, have most characteristics in common with $K(\pi,\tau,\alpha,\bar{R})$, just replacing $\alpha + 1$ by β, and $\bar{R}$ by R. For example, $d\tilde{K} / d\pi < 0$ and $d\tilde{K} / dR < 0$. Note also that $\tilde{K}_2$ equals K_2 if $\beta = \alpha + 1$, so there is not much of a difference if waiting does not apply in the periphery.

However, K_1 and $\tilde{K}_1$ are quite different when evaluated as functions of α and β, respectively. This is due to the discount factor. As noted above, the optimal discount factor in the model with cost variation is U-shaped in α. The optimal discount factor in case of price variation is $R(\beta - 1)$, which is strictly increasing in β. The reason is that there is no upper price barrier similar to the cost barrier at zero.

From an analytical point of view, the model with price variation is more appealing than the model with cost variation because of this difference. It can be shown that $\tilde{K}_1$ (like $\tilde{K}_2$) is strictly increasing in β. This also affects the limiting behaviour of $\tilde{K}(\pi,\tau,\beta,R)$ when $\beta \to 1$. For example, the overall shape of $\tilde{K}$ as a function of τ will be like the rightmost curve in Figure 1 if $\beta = 4$, $R = 0.03$ and $\pi = 0.4$. (This is comparable with $\alpha = 3$ in the previous

model). Similarly, the shape will be like the curve for $\alpha = 1.5$ if β is reduced to 2.5. Most interestingly, however, nothing like the two dotted curves appears by decreasing β further down towards unity. The U-shape just get wider and deeper, implying that agglomeration stabilizes for an increasingly larger trade cost (smaller τ).

Thus, if externalities arise from the demand side, agglomeration always becomes more likely by increasing the uncertainty or expected growth of demand (σ or μ). Similarly, it always becomes less likely the less patient the agents and the larger the growth rate (ρ).

Intuitively, these results strengthen the conclusion that the dotted curves in Figure 1 depict exceptional cases. One might say that they reflect the lack of realism in using a geometric Brownian cost process if the uncertainty (or the expected learning rate) is very large. A process with mean-reversion seems more realistic to reflect large uncertainty in such cases.

7. Final remarks

This chapter has addressed the question of agglomeration and trade in a setting with externalities that contribute to learning. The results resemble those of static models, in particular with respect to trade costs, as we found that agglomeration is more likely with intermediate trade costs. By and large, agglomeration is also more likely the greater the economies of scale.

The effect of the dynamic parameters of interest was shown to be quite complex. However, if we disregard some unrealistic parameter combinations in the model with cost variation, the results are clear: Firm-specific uncertainty makes agglomeration more likely. Agglomeration is also encouraged by increasing the size of the economy or the manufacturing share, as long as learning is positively related to size. On the other hand, agglomeration is discouraged by growth and time preferences.

As in previous chapters, specific combinations of parameters were required for convergence. If the dynamic effects are not strong, there may not be enough value from waiting for an equilibrium with free entry and zero profits in expectation. This is an interesting topic for further research, that seems to need some kind of oligopoly models.

APPENDIX A: Cost uncertainty and spillovers from production

To determine the elasticity α if there is cost uncertainty, and spillovers stem from production, we must solve the third-degree equation that appears by inserting μ in eqn. (10) into eqn. (2):

$$\text{(A.1)} \quad \Psi(\alpha,\sigma) \equiv \tfrac{1}{2}\sigma^2\alpha^3 + (\sigma^2 + \gamma L)\alpha^2 + (\tfrac{1}{2}\sigma^2 - \rho)\alpha - \rho = 0.$$

The analytical solution is messy, but differentiation gives

$$\text{(A.2)} \quad \frac{\partial \Psi}{\partial \alpha} = \tfrac{3}{2}\sigma^2\alpha^2 + 2(\sigma^2 + \gamma L)\alpha + (\tfrac{1}{2}\sigma^2 - \rho),$$

and

$$\text{(A.3)} \quad \frac{\partial \Psi}{\partial \sigma} = \sigma\alpha^3 + 2\sigma\alpha^2 + \alpha\sigma.$$

Thus we have

$$\text{(A.4)} \quad \frac{d\alpha}{d\sigma} = -\frac{\partial \Psi / \partial \sigma}{\partial \Psi / \partial \alpha} = -\frac{\sigma\alpha^3 + 2\sigma\alpha^2 + \sigma\alpha}{\tfrac{3}{2}\sigma^2\alpha^2 + 2(\sigma^2 + \gamma L)\alpha + (\tfrac{1}{2}\sigma^2 - \rho)} \overset{\text{(A.1)}}{=} -\frac{\sigma\alpha^4 + 2\sigma\alpha^3 + \sigma\alpha^2}{\sigma^2\alpha^3 + (\sigma^2 + \gamma L)\alpha^2 + \rho} < 0,$$

since all terms are positive for positive roots of α.

APPENDIX B: Cost uncertainty and spillovers from patenting

A slightly different third-degree equation than in *Appendix A* appears if learning stems from patent investments as opposed to production. By inserting eqn. (23) into eqn. (2), we find:

$$\text{(B.1)} \quad \Psi(\alpha,\sigma) \equiv \tfrac{1}{2}\sigma^2\alpha^3 + \sigma^2\alpha^2 + (\tfrac{1}{2}\sigma^2 + \gamma L - \rho)\alpha - \rho = 0.$$

Differentiation yields

$$\text{(B.2)} \quad \frac{\partial \Psi}{\partial \alpha} = \tfrac{3}{2}\sigma^2\alpha^2 + 2\sigma^2\alpha + (\tfrac{1}{2}\sigma^2 + \gamma L - \rho),$$

and

$$\text{(B.3)} \quad \frac{\partial \Psi}{\partial \sigma} = \sigma\alpha^3 + 2\sigma\alpha^2 + \alpha\sigma,$$

and thus:

$$\text{(B.4)} \quad \frac{d\alpha}{d\sigma} = -\frac{\partial \Psi / \partial \sigma}{\partial \Psi / \partial \alpha} = \frac{\sigma\alpha^3 + 2\sigma\alpha^2 + \alpha\sigma}{\tfrac{3}{2}\sigma^2\alpha^2 + 2\sigma^2\alpha + (\tfrac{1}{2}\sigma^2 + \gamma L - \rho)} \overset{\text{(B.1)}}{=} -\frac{\sigma\alpha^4 + 2\sigma\alpha^3 + \sigma\alpha^2}{\sigma^2\alpha^3 + \sigma^2\alpha^2 + \rho} < 0.$$

References

Dicken, Peter and Peter E. Lloyd, 1990, *Location in Space. Theoretical Perspectives in Economic Geography*, Harper-Collins Publishers, New York

Fujita, Masahisa and Jacques-Francois Thisse, 1996, "Economics of Agglomeration", *Journal of the Japanese and International Economies* 10, 339-378

Fujita, Masahisa, Paul Krugman and Anthony J. Venables, 1998, *The Spatial Economy,* book manuscript

Grossman, Gene M. and Elhanan Helpman, 1991, *Innovation and Growth in the Global Economy*, MIT Press, Cambridge, Massachusetts

Krugman, Paul R., 1991a, *Geography and Trade*, MIT Press, Cambridge, Massachusetts

Krugman, Paul R., 1991b, "Increasing Returns and Economic Geography", *Journal of Political Economy* 99, 3:483-499

Krugman, Paul R. and Anthony J. Venables, 1996, "Integration, specialization, and adjustment", *European Economic Review*, Papers and Proceedings 40, 959-967

Midelfart Knarvik, Karen Helene, 1995, "Technological Spillovers, Industrial Clusters and Economic Integration", *Discussion Paper* 6, Institute of Economics, Norwegian School of Economics and Business Administration (NHH)

Ottaviano, Gianmarco I. P. and Diego Puga, 1997, "Agglomeration in the Global Economy: A Survey of the New Economic Geography", *CEP Discussion Paper* 356

PART III

Equilibrium Models with Product Differentiation

CHAPTER 6

Product Differentiation

1. Introduction

In static modelling there is a close link between economies of scale and product differentiation. The Dixit-Stiglitz (1977) model[1], which describes this relationship in a general equilibrium framework, has become a standard model in new trade and growth theory as well as in new economic geography.[2] While the first models of this kind introduced differentiation from the demand side, assuming that the consumers have a preference for variation, Ethier (1982) introduced it from the cost side, assuming that production of a homogeneous good is based on a set of intermediates produced under economies of scale.

The equilibrium models that have been discussed in previous chapters share many characteristics with static models based on product differentiation. Various measures, quite similar to static measures of product differentiation, were also used to describe the fact that costs or demand for specific goods may change over time. Nevertheless, the goods have been perfect substitutes in a static sense, so the dissertation so far has really been a story about homogeneous goods.

The objective of this chapter is to extend the methodology in a way that brings about a true dynamic interpretation of the Dixit-Stiglitz model with constant elasticities. The most important equilibrium formulas will coincide with those of the static model.

---

[1] See also Spence (1976).

[2] See Krugman (1990), Grossman (1991), and Grossman and Helpman (1991) for a number of path-breaking models in trade and growth theory. Fujita and Thisse (1996), and Ottaviano and Puga (1997) survey the somewhat related geography literature, which mainly stems from the 1990's.

The Dixit-Stiglitz model is characterized by its structure of preferences. The preference for variation in differentiated goods can be expressed in compact form by a utility function $U(q) = Nq^{(b-1)/b}$. Here N is the number of categories, q is the quantity in each one, and b (> 1) is a constant. It follows that the consumer will be better off the larger the number of categories over which he can spread a fixed total consumption q_{tot}. Then the amount of each category is $q = q_{tot}/N$, and the utility becomes $U(N) = N^{1/b} q_{tot}^{(b-1)/b}$. This is increasing in N as long as b is finite.

The constant b is a measure of product differentiation. However, it does not say whether the differences stem from the demand or the supply side; i.e., whether it is the consumer who has a fine taste or whether the goods are different as measured by a fixed scale. Of course, this is not important in the static model. The matter of interest is product differences as they are perceived by the consumer. The distinction is more interesting in the dynamic setting that is developed below, where the dynamic elasticity β will be a true measure of differentiation.

In *Chapter 3*, it was argued that β arises from a mixture of underlying parameters μ, σ and ρ. Some of these are technological, whereas some are more naturally related to taste. For example, we could let $\mu > 0$ represent the growth process of a pine tree, but as an extreme alternative, it could also represent the satisfaction by repeated listening to a Beethoven symphony. Similarly, ρ is a growth rate by one interpretation, and a rate of time preferences by another.

In the dynamic version of the Dixit-Stiglitz model that follows, not all interpretations of μ, σ and ρ that were discussed in *Chapter 3* are applicable. Partly for this reason, and partly to bring forward the main idea, most of the discussion is based on a very narrow interpretation of these variables. For simplicity, we also leave the discussion of uncertainty to the end.

2. A dynamic representation of differentiated goods

For a dynamic approach to be of any interest, something must change with time. The main thing that changes in this economy is productivity. As before, labor is the only production factor, there is a fixed endowment of it, and we denote the growth rate of productivity by a

constant ρ. There are no subjective time preferences, so ρ is simply the rate at which a representative worker improves in his job throughout history.

Scale economies are reflected in a minimum labor requirement for construction of specific goods. As usual, there are two types of investment, but no maturing process applies. Thus, to avoid confusion, let us replace the wine industry from *Chapter 3* with a cake industry, emphasizing that this example is also chosen for pedagogical reasons. Let one cake correspond to one firm. First, all the ingredients must be mixed together. Second, the cake must be baked and served.

The first activity requires a fixed investment similar to the patent cost in previous chapters. A minimum number of workers must be employed to undertake such an effort. For example, the details of the process may prevent one person from making more than one cake each period. Therefore he must stay with this rate, but as his productivity improves, the size of the cake that he can make will grow at rate ρ. Thus, the example assumes physical growth, although increased quality is often a more interesting interpretation, as also noted in *Chapter 3*.

In the other activity, denoted by activation or production, there is no minimum level as to the patent. If a fixed number of workers are needed to bake a new cake, productivity growth implies that the requirement goes down if the cake is not put into the oven immediately.

So far, nothing has changed from the wine example, except that a maturing process applied to the wine. For the moment, assume that cakes which are not yet baked, do not change by getting older, and that the preferences for specific cakes are fixed.

However, in this model the consumer cares about both the quantity of each good and the number of categories, since the cakes will not be exactly alike: some may happen to be a little sweeter than others etc. The consumer enjoys this kind of variation, so he prefers two different pieces of cake as opposed to one piece twice as large.

How can variation be obtained in a world like this, where the firm size is fixed and all consumers have the same taste? By time! If the cakes that are baked (and consumed immediately afterwards) are old and therefore small, baking will not need as many workers. Then more workers will be involved in making new cakes. Since all investments take place at

different times, the implications are clear: the stronger the preference for variation, the more often the consumer will be eating cake, but the piece will be smaller each time.

If the marginal utility when increasing the quantity of a specific cake is decreasing, his preferences are strictly concave, as required. However, the consumer is "modern" in the sense that his preferences are updated to the prevailing level of productivity (or technology) at any time, as discussed in *Chapter 3*. Thus, assume that the gain from consuming one new cake always equals one. Figure 1 illustrates such a single-good utility function.

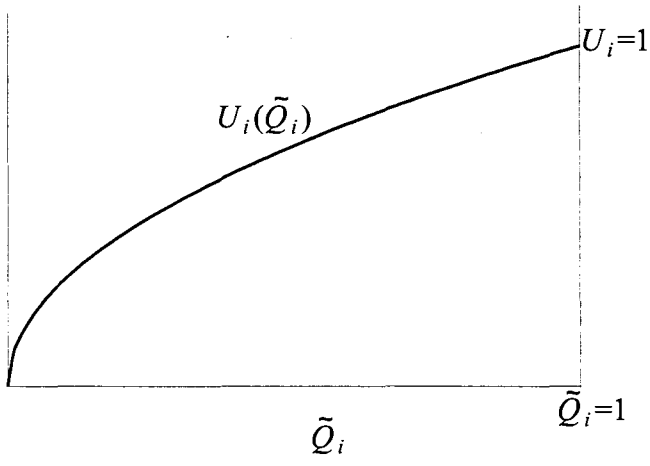

Figure 1. Single-Good Utility.

In the figure, $\tilde{Q}_i$ is the quantity relative to one new cake (i.e., the state-of-the-art) at any time. It is a product of two factors: The size of the whole cake is measured by the discount factor $Q_i = e^{-\rho T_i}$ (≤ 1), which takes into account the age of the cake, T_i (≥ 0). In addition, the representative consumer will not be the only one, so the share q_i (≤ 1) of the cake that he receives also matters. Hence, the correct quantity to insert into the utility function becomes:[3]

(1) $\tilde{Q}_i = Q_i q_i$.

To describe such preferences formally in a two-sector model, we will derive the dynamic response to the demand functions that arise from a standard Dixit-Stiglitz model.

---

[3] For example, a cake of age $T_i = \ln 2/\rho$ yields $Q_i = \frac{1}{2}$. If the consumer gets half of it, we have $q_i = \frac{1}{2}$, and $\tilde{Q}_i = \frac{1}{4}$.

3. The dynamic Dixit-Stiglitz model

Periodic utility is expressed by a Cobb-Douglas function

$$U = H^{1-\pi} M^{\pi}, \tag{2}$$

where π is constant ($0 < \pi < 1$). Further, H and M are consumption of homogeneous goods and manufactured (differentiated) goods, respectively. M is a CES-aggregate

$$M = \left(\int_0^N \tilde{Q}_i^{(b-1)/b} di \right)^{b/(b-1)}, \tag{3}$$

where N is the number of goods that are consumed in the particular period, $b > 1$ is a constant elasticity of substitution, and $\tilde{Q}_i$ is consumption of each good. The utility function is like a static one, but it expands the static definition as it applies in all periods throughout the history for a consumer who adjusts his perception of utility to the current level of productivity. This has been anticipated by the definition (1) that is embodied in (3). As the firm size is fixed, the relative size, which matters for utility, decreases along with Q_i. Hence, the discount factor (which here might be called a productivity factor) establishes the updating mechanism. There are no subjective time preferences in the traditional sense, so equilibrium will be determined by maximizing utility each period conditional on an income constraint.

By a standard argument, the consumers will spend an income share π on manufactures, and $1-\pi$ on the homogeneous good. We get a set of demand functions for manufactured goods

$$\tilde{Q}_i = \left(\frac{P_i}{G} \right)^{-b} \frac{\pi Y}{G}, \tag{4}$$

where P_i is the price in terms of $\tilde{Q}_i$, Y is the total income rate, and G is a perfect price index:

$$G = \left(\int_0^N P_i^{1-b} di \right)^{1/(1-b)}. \tag{5}$$

Since N is assumed to be large, each firm can act as a monopolist facing a constant price elasticity of demand equal to b. This is plotted in Figure 2.

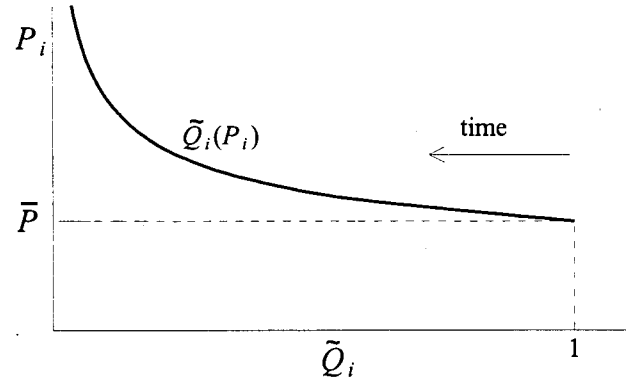

Figure 2. Demand Functions in a Dynamic Dixit-Stiglitz Model.

Contrary to a static model, the firm size is fixed, so the firm has no opportunity to scale its production at a specific time. However, the patent fixes the size of the product, and growth (at rate ρ) is embodied in new patents. For this reason, a firm that does not activate immediately will find itself moving to the left in Figure 2, as its effective (relative) size shrinks. The effective size of a new firm is one, so the process will start at $\tilde{Q}_i = 1$, as indicated in the figure.

As long as the properties of the goods do not change with time, we know from previous chapters how this development will be experienced by a small, waiting firm in equilibrium: It can act as if it faces a firm-specific price that increases at a constant rate, because this corresponds to a constant elasticity of the discount factor.

Let us show this formally. At entry, we have $\tilde{Q}_i = 1$, but the effective size shrinks by the process $d\tilde{Q}_i = -\rho\tilde{Q}_i dt$ due to growth. The demand function in eqn. (4) can be written as $\tilde{Q}_i = (\bar{P}/P_i)^b$, where $\bar{P}$ is regarded as constant as N is large. This yields: $dP_i/dt = (dP_i/d\tilde{Q}_i)\cdot(d\tilde{Q}_i/dt) = (-P_i/\tilde{Q}_i b)\cdot(-\tilde{Q}_i\rho) = (\rho/b)P_i$. Hence, the firm will observe a price that increases at rate $\rho/b \equiv \mu_{\text{det}}$, which is smaller than the growth rate as $b > 1$.

Notice that μ_{det} does not describe a "magic" preference for old products. It is simply a dynamic parameter that arises from fixed, but continuously updated, static preferences for variation. The underlying force is not like a process that makes homogeneous trees grow.

To a large extent, the difference between the static and the dynamic model is a matter of when the goods are consumed. In the static model this must take place instantaneously. Here, variation is obtained by time, as a number of goods are consumed each period.

This is a quite nice story about growing demand for old products. To put it simply, a veteran car may be popular not because it is old but because it is different. In such cases, price growth ($\mu_{\text{det}} > 0$) does not mean that preferences change; it is a consequence of a preference for variation. The larger μ_{det} (i.e., the smaller b), the more the consumer is willing to reduce the quantity of each good to obtain a larger number of categories. As there are no subjective time preferences, equilibrium follows by repeating this story each period, and by requiring optimal investment decisions, free entry, and clearing of the factor market.

The cost function of a representative firm becomes

$$(6) \qquad J_i = (L_A + \tilde{Q}_i L_C)w,$$

where w is the wage rate, L_A is the fixed number of workers required for a patent, and L_C is the similar fixed number of workers that would be needed to activate a patent immediately. Due to growth, the labor requirement for activation decreases along with the discount factor for each good. Notice that the firm size is one, so we might have left out the tilda on Q_i in eqn. (6). Nevertheless, we keep it just to emphasize the difference from the models with homogeneous goods in previous chapters.

Profits are maximized if the firm charges a fixed markup price as discussed several times in previous chapters. Leaving out firm subscripts and asterisks for optimal values, the optimal price at which to activate is

$$(7) \qquad P = \frac{\beta}{\beta - 1} wL_C,$$

where $\beta = \rho/\mu_{\text{det}}$.[4] Free entry is obtained by the zero profit requirement:

[4] The replacement of b by β is a matter of convention, as we assume that the firm acts according to the observed dynamic variables, μ_{det} and ρ.

(8) $\left(P - wL_C\right)\tilde{Q} - wL_A = 0.$

This free entry condition is *ex ante*, but with no uncertainty (as assumed for the moment), *ex post* profit will also be zero for all firms. By combining eqns. (7) and (8), we also find

(9) $\tilde{Q} = R(\beta - 1)$, where $R \equiv L_A/L_C$.

Similar to the *Chapter 3* model, $\tilde{Q} \leq 1$ is required, so there cannot be too much scale economies. Eqn. (9) shows that the age of each patent that is activated will be higher (the size of the good relative to a new one will be smaller) the stronger the preference for variation; i.e., $\tilde{Q}$ will be smaller the smaller β (= b). This conforms with the intuitive arguments in the introduction.

Further, $\tilde{Q}$ in eqn. (9) does not depend on more than β and the ratio of the two exogeneous cost elements. In a multi-region setting with "iceberg" trade costs (see below), the allocation of resources between the two types of investment (patenting and production) will then be independent of the distribution of firms, so the relative number of firms will be equal to the relative number of workers.

Finally, if the total labor endowment in manufacturing is L_M, the steady-state condition is $N(L_A + \tilde{Q}L_C)w = L_M w$. Using eqn. (9), this simplifies to:

(10) $N = \frac{L_M}{L_A \beta}.$

If all patents are used (which is true in this deterministic case), N is the constant rate of entry, activation and consumption in equilibrium. As eqn. (10) shows, the number of categories will be larger the stronger the preference for variation (i.e., the smaller β). This also conforms with the introductory remarks.

4. Relationship between the static and the dynamic approach

There is just one technical difference between this model and the two-sector model in *Chapter 3*: The quantity measure c_i (= q_i) in the utility from manufactures in the *Chapter 3*

model has effectively been replaced by $q_i^{(b-1)/b}$ in eqns. (1),(3). However, in both models the firm-size is fixed at unity, so unless we have a trade context with different consumer prices, the models will behave similarly in most respects.

Although the technical change is small, the interpretation of the new model is highly different, as there is now a true preference for variation. The *Chapter 3* model contained no gains from scale. By increasing the size of the economy, each consumer would get a larger number of glasses each period, but less wine in each one, so to speak. That did not make him wealthier, because the wine was homogeneous and the total quantity each period stayed the same. This model, however, is analogous to the static Dixit-Stiglitz model with respect to individual welfare. Welfare increases with size because the cakes are different, and the consumers prefer variation.

The close relationship to the static Dixit-Stiglitz model can be illustrated very simply: If Q_i is fixed at unity, all dynamic effects are removed, and the static model appears by allowing q_i to fluctuate. On the other hand, if q_i is fixed at unity for a firm activating its patent, while Q_i is allowed to fluctuate downwards by postponing this investment, we get the dynamic model with analogous preferences. Demand will be represented by isoelastic demand functions $\tilde{Q}_i(P_i)$, and the preference for variation is symmetric in quantity and time.

This symmetry is convenient as it implies that we do not have to be concerned about whether market power is exploited instantaneously (by splitting a fixed quantity) or dynamically (by waiting). Therefore the model can be treated as a static model although it is actually dynamic. To see why, consider the options facing a forward-looking firm with a new patent; i.e., a firm that finds itself at $\tilde{Q}_i = 1$ (the start of the arrow) in Figure 1. The firm size is fixed at one, but the firm can choose any relative quantity $\tilde{Q}_i$ by deciding on a particular discount factor Q_i. Furthermore, if there is more than one region, the firm can split such a quantity optimally by selling to each of them, and according to the demand function (4).

But these are exactly the same options as for a firm which has paid the fixed cost in the analogous static model. Thus, we have to get the same equilibrium conditions as well, as long as the optimal quantity in the dynamic model is less than one; i.e., if waiting applies.

5. Trade costs

One of the most interesting properties of the static Dixit-Stiglitz model is its ability to explain intra-industry trade. We shall see how this can be incorporated into the dynamic setting with trade barriers. However, start with the most simple case, assuming no trade costs. Sales are to be split between two regions (1 and 2) populated by consumers with identical preferences.

Since activation (baking a cake) cannot be split up in time, the firm in region 1 will activate according to the usual markup rule (7), with a corresponding discount factor Q_1. (Note that the subscripts now refer to a region, not to be confused with the firm index i above.) With the firm size fixed at unity, the total supplied relative quantity is $\tilde{Q}_1 = Q_1$, and as the firm decides on a particular Q_1, total sales are split up between region 1 and 2 as follows:

$$(11) \qquad \tilde{Q}_1 = Q_1 q_1 + Q_1 q_2 \equiv \tilde{Q}_{11} + \tilde{Q}_{12} .$$

Here q_1 is the region 1 share, and $q_2 = 1 - q_1$ is the region 2 share. Thus, $\tilde{Q}_{11}$ and $\tilde{Q}_{12}$ are relative quantity measures like $\tilde{Q}_1$, and the corresponding price in each region is reduced according to eqn. (4). This gives the following total demand:

$$(12) \qquad \tilde{Q}_{11} + \tilde{Q}_{12} = \left(\frac{P_{11}}{G_1}\right)^{-\beta} \frac{\pi Y_1}{G_1} + \left(\frac{P_{12}}{G_2}\right)^{-\beta} \frac{\pi Y_2}{G_2} .$$

Here P_{1j} is the consumer price of the region 1 good that is sold in region j. For zero trade costs, the price indices are the same, and $G_1 = G_2$. By scaling the total rate of income to one ($Y_1 + Y_2 = 1$), profits are maximized if $P_{11} = P_{12}$, $q_1 = Y_1$, and $q_2 = Y_2$. This exploitation of market power by splitting up sales is exactly as in the static model.

Two alternative *ad valorem* trade barriers can be introduced: Either an "iceberg" cost as in many static Dixit-Stiglitz models, or a time cost as in *Chapter 5* (where it was called a "drift bottle cost", but that name does not apply here as there is no maturing). If τ is a constant between zero and one, $\tau = 1$ corresponds to no trade barrier, and $\tau = 0$ to an infinite barrier.

If $\tilde{Q}_{12}$ is the quantity shipped from region 1, a time barrier can be illustrated by expressing the quantity arriving in region 2 as $\tau\tilde{Q}_{12} = (\tau Q_1)\cdot q_2$. Thus the discount factor is smaller, so the

importing region gets the good later. Then the relative size, and thereby the *effective* arrived quantity, is smaller. Iceberg costs follow by an alternative interpretation of the same expression: $\tau\tilde{Q}_{12} = Q_1 \cdot (\tau q_2)$. Then the actual arrived quantity is also smaller, but with no delay.

In any case, the demand for a region 1 good in region 2 is given by $\tau\tilde{Q}_{12} = (P_1 / \tau G_2)^{-\beta} \pi Y_2 / G_2$, where P_1 is the producer price. That is, only a share τ of the shipped quantity arrives, which also implies a consumer price that is a factor $1/\tau$ higher than the producer price. If $\tilde{Q}_{11}$ denotes the demand for a good from region 1 in the domestic market, total demand can be summarized (assuming a common producer price P_1) as

$$(13) \quad \tilde{Q}_{11} + \tilde{Q}_{12} = \left(\frac{P_1}{G_1}\right)^{-\beta} \frac{\pi Y_1}{G_1} + \left(\frac{P_1}{G_2}\right)^{-\beta} \tau^{\beta-1} \frac{\pi Y_2}{G_2},$$

and similarly for a region 2 good. Thus, the firm faces an isoelastic demand curve in both regions, so it will want to activate the patent (i.e., select the appropriate markup price) according to the same investment rule as if only selling in the domestic market.

Notice that the effect of the trade cost is the same for both interpretations of τ, although the good does not "melt" physically by the time barrier interpretation. But how can there be a real loss in this case? The loss arises because the good must be produced earlier. Hence, production requires more labor than without the barrier, and there is what we may call a dynamic productivity loss decreasing the level of welfare. This can be illustrated by thinking of a core-periphery equilibrium with all manufacturing in one region. The firms will activate when obtaining a fixed markup price, and at a constant rate in equilibrium. However, the consumers in the periphery will not get access to any of these goods before later. They will experience the same growth of real income as the consumers in the core, but lagging behind. As noted in *Chapter 5*, the time lag is $\ln(1/\tau)/\rho$, which is zero for $\tau = 1$, and positive for $\tau < 1$.

The time barrier interpretation may be convenient for studies of the product cycle. Iceberg costs seem more natural in most of our applications, although one reason for this may simply be that they are more familiar from static modelling. In any case, the result is the same by both interpretations for a large-group equilibrium.

6. Uncertainty

As long as the specific goods do not change in any respect by getting older, as assumed up until now, there is a unique relationship between the presented model and the static Dixit-Stiglitz model with constant elasticities. Below, we show that uncertainty can be included in the dynamic setting, but that a constraint on the stochastics is necessary. The appropriate set of firm-specific stochastic processes will be found by the same technique as when studying linear demand and supply curves in the notes to *Chapter 1*.

In the deterministic case in previous sections, it was noted that the effective size of each fixed-sized good shrinks according to the process $d\tilde{Q}_i = -\rho \tilde{Q}_i dt$. Uncertainty can be introduced by assuming that the consumer does not perceive it this way; instead, he perceives each good as if it starts to contract or expand stochastically. More precisely, let the perceived process be an independent geometric Brownian motion:

(14) $\quad d\tilde{Q}_i = -\tilde{\mu}\tilde{Q}_i dt + \tilde{\sigma}\tilde{Q}_i dz$.

By Ito's lemma, it follows from the demand function, eqn. (4), that the price process will also be geometric Brownian, with drift $\mu = \tilde{\mu}/b + (b+1)\tilde{\sigma}^2/2b^2$, and volatility $\sigma = \tilde{\sigma}/b$.

As all agents are risk neutral (and we may imagine that contracts are written when each process starts; i.e., whenever a patent is acquired), we are looking for a price process that yields an elasticity of the *expected* discount factor that equals b. To obtain this symmetry, b must be the positive solution to the characteristic quadratic, eqn. (18) in *Chapter 3*. By inserting for μ and σ into that equation (and setting $x = b$), we find that eqn. (14) complies in expectation with the fixed static preference for variation if the following formula holds:

(15) $\quad \tilde{\sigma}^2 + \tilde{\mu} - \rho = 0$.

Hence, there are an infinite number of stochastic processes matching a constant static elasticity. Among them is the deterministic case, where $\tilde{\sigma} = 0$, and $\tilde{\mu} = \rho$.

In order to obtain this result, we had to assume that all consumers were alike. The problem would be more complex if consumers in different regions could develop different firm-

specific preferences; i.e., if more than one realization of the price process applied to each good.

So long as the cost side is not affected, it seems to be a matter of interpretation whether we consider the stochastics above to be a result of things that happen with the good, or whether it is the consumer who changes his perceptions. As in the static model, the fact that counts is how a specific good is perceived by the representative consumer, not how somebody else might have looked at it.

Finally, by similar arguments as in *Chapter 3*, we can conclude that all equilibrium formulas remain, except that $\tilde{Q}$ becomes an expected value if it is uncertain when the markup price (7) is reached for the first time. There is still a constant preference for variation in a large-group sense, as the rate of entry (10) remains. However, with high uncertainty (more precisely, if $\frac{1}{2}\sigma^2 > \mu$), some patents will never be used, and the rate of activation will be lower than the rate of entry.

7. Final remarks

When restricting to a deterministic model and a fixed growth rate of productivity, it has been shown that the relationship between the static and the dynamic approach is unique. The taste parameter (μ_{det}) in the dynamic model arises as a consequence of fixed preferences for variation, and fixed preferences for specific goods. If departing from the latter assumption, a requirement must be put onto the stochastics. In one sense this makes the model less appealing, but the number of alternative interpretations also increases.

It is not fair to regard this dynamic model of monopolistic competition as a generalization of the static model. Technically, the static Dixit-Stiglitz model is obtained by reinterpreting the cost variables, and by removing the dynamics as well as the assumption of fixed-sized firms. Thus, the interpretation is somewhat different. The dynamic model also has more requirements for convergence, as waiting does not always apply. Such cases, as well as relaxing the assumption of a fixed firm size, are interesting topics for further research.

It should be emphasized that we have applied a more narrow interpretation of the variables μ, σ and $\dot{\rho}$ than in previous chapters. For example, it is not straight-forward to interpret ρ as a rate of time preferences instead of a growth rate. A technical trick (or a more creative interpretation than I have come up with yet) might do the job, but the problem is that the effective size of a firm does not shrink by ageing in this case. Therefore waiting does not bring about a larger number of products, and it becomes harder to obtain the required symmetry between the time and quantity dimensions. On the other hand, it seems possible that ρ could be interpreted as a growth rate of the labor force instead of a growth rate of productivity, as in the *Chapter 3* model (section 5). In this case, the underlying assumption would be that a firm can not be too small relative to the size of the economy, e.g. as the optimal scale of ancillary activities, like advertising, will be determined by the size of the economy. If each patent requires a share L_A/L_M of the labor force in manufacturing, and L_M grows at rate ρ, then L_A must increase at the same rate. Thus, we still get a stationary equilibrium that is described by the same equations as above.

In cases where the symmetry between the time and quantity dimensions cannot be retained, one might want to use a utility function with inner terms $U_i = Q_i^{(\beta-1)/\beta} q_i^{(b-1)/b}$, where $b \neq \beta$. Then b could take care of static preferences for variation, and β take care of the dynamics. This yields a more general model, but also one that is more difficult to use.

Although this discussion on product differentiation has focused on the isoelastic case, the interpretation seems to apply more broadly. For example, we could have started with a single-sector static utility function $U = \int_0^N U_i di$ with variable elasticities as in *Chapter 3, Appendix A*. The related demand functions correspond to the same price processes as those derived in that appendix, but in this case, we should also expect scale effects as discussed by Krugman (1979) and others for the analogous static model.

References

Dixit, Avinash K. and Joseph E. Stiglitz, 1977, "Monopolistic Competition and Optimum Product Diversity", *American Economic Review* 67, 297-308

Ethier, Wilfred J., 1982, "National and International Returns to Scale in the Modern Theory of International Trade", *American Economic Review* 72, 389-405

Fujita, Masahisa and Jacques-Francois Thisse, 1996, "Economics of Agglomeration", *Journal of the Japanese and International Economies* 10, 339-378

Grossman, Gene M. (ed.), 1991, *Imperfect Competition and International Trade*, MIT Press

Grossman, Gene M. and Elhanan Helpman, 1991, *Innovation and Growth in the Global Economy*, MIT Press

Krugman, Paul R., 1979, "Scale Economies, Product Differentiation, and the Pattern of Trade", *American Economic Review* 70, 950-959

Krugman, Paul R., 1990, *Rethinking International Trade*, MIT Press

Ottaviano, Gianmarco I. P. and Diego Puga, 1997, "Agglomeration in the Global Economy: A Survey of the New Economic Geography", *CEP Discussion Paper* 356

Spence, Michael, 1976, "Product Selection, Fixed Costs, and Monopolistic Competition", *Review of Economic Studies* 43, 217-235

CHAPTER 7

Agglomeration with Pecuniary Externalities

1. Introduction

Chapter 5 initiated a discussion of economic geography by a model with true externalities. However, this is not the only way that agglomeration forces can appear. The new economic geography literature may have brought more important contributions in pointing at how agglomeration can be encouraged by forward and backward linkages arising from pecuniary externalities. In this chapter we shall see how a pathbreaking static geography model with pecuniary externalities can be coupled with the dynamic approach to product differentiation that was discussed in *Chapter 6*. First, consider the following simple story:

Suppose that a number of workers are employed in an industry, named manufacturing, that makes up a significant share of the total economy. Manufactured goods have many characteristics in common, but in some respects they are also different.

Further, the economy consists of several regions separated by trade barriers. Each region may have a share of the labor force employed in manufacturing, while the rest is occupied by some spread activity, say farming.

If all manufacturing happened to be concentrated in one region, is it likely that new firms will also be established there, even if the workers were not tied to the land like the farmers? That is: will agglomeration be a stable equilibrium? Due to trade costs, the cost level will be lower in the large market, where manufacturing is concentrated. However, consumer prices will be higher in other markets. If manufactured goods are fairly alike and trade costs are high, new firms will be pulled out to the periphery by these consumer prices. The workers must be compensated for a higher cost related to living in the periphery, but since manufactured goods only represent a share of total consumption, the difference in prices will count more than the cost difference that must be compensated.

This may be different if the consumers have strong preferences for differentiated consumption; i.e., if they are willing to pay much for getting access to a larger number of categories of manufactures. Then there will be a lot of intra-industry trade, implying that a firm locating in the periphery must base a large share of its sale on exports. There are two drawbacks by locating in the periphery in such cases: High production costs and a small domestic market. Possibly, those drawbacks may be more important than the gain from less competition. If firms enter and leave independently, there seems to be, and there indeed is, a possibility that manufacturing will remain concentrated for some combinations of parameters.

This is an old story. It was told by Krugman (1991) in the introduction to his "core-periphery" model, and economic geographers told similar stories even long before. However, a static model like Krugman's cannot say much about how agglomerated economies actually rise or decline conditional on dynamic variables like growth, preference changes etc. The presented dynamic interpretation of the Dixit-Stiglitz (1977) model can enrich the understanding of agglomeration in this respect.

As a matter of fact, we shall be able to construct a dynamic version of Krugman's model that shares its mathematical form. Most of the discussion will be focused on interpreting results that can be extracted from the dynamic framework, but for completeness we spell out the whole model first. Before doing so, it is convenient to summarize how the main assumptions and results relate to the static model:

1. The dynamic model is based on the same static preferences as Krugman's model at any point in time (possibly in an expected sense). However, the model spans over an infinite time horizon, which includes exogenous growth of productivity. The representative consumer responds to this growth process by adjusting his perception of utility to the current level of productivity at any time.

2. Scale economies and market power are obtained by minimum requirements for patent investments, analogous to the fixed cost that applies in the static model.

3. All the rest of the main requirements in the static core-periphery model apply over the entire time horizon. This includes assumptions on trade costs, factor endowments, factor mobility etc.

4. In broad terms, the dynamic framework brings forward a new interpretation of the core-periphery model. This can be used to study the two types of equilibria in the model: either a symmetric equilibrium, or one with agglomeration.

Finally, note that the static core-periphery model has been extended in a number of directions since the original publication; see e.g. Krugman and Venables (1995), Puga (1996), Baldwin and Forslid (1997), and Fujita, Krugman and Venables (1998). Hopefully, the analogy that is developed below also applies to several of these extensions.

2. The model

Non-technical description

The model has two regions (1 and 2), two goods (manufactures and agriculture), and two specific production factors (workers and farmers). The workers produce manufactured goods, the farmers produce agricultural goods, and the total endowments of workers and farmers are fixed. All agents are forward-looking, optimizing decision makers.

Agriculture is a homogenous, perfectly competitive good that can be traded at no cost. Half of the farmers live in each region, and they are not allowed to move. This implies equal wage rates in farming, and equal shares of total farmer income to each region. As in *Chapter 5*, the main role of the farmers is to make up a demand for manufactured goods.

Manufactures are differentiated goods, produced by firms that undertake irreversible investments. An infinite number of possible products exist, and a large number are produced each period. Hence, each firm is small compared to the size of the market. The workers are mobile, and will move - at no cost and with no delay - to the region that offers the higher real wage.

Two investments are necessary before a manufactured good can be consumed. First, there is an entry cost that gives the firm the exclusive right to produce a fixed quantity of the good. This patent investment is irreversible in the sense that it locks on to a particular technology, and fixes the size of the firm. However, the firm is free to choose when to activate the patent. Whenever activation takes place, the good must be consumed immediately.

The productivity of the workers grows at a fixed, exogenous rate ρ in both activities. There are scale economies, as a fixed number of workers always must be hired for one period to acquire a patent. Thus, growth of productivity is embodied linearly in new patents, since the effective size of an old patent will be smaller than a new one, as in several previous models.

The formal setup

Preferences are fixed in the sense that the following Cobb-Douglas utility function applies in all periods:

(1) $U = H^{1-\pi} M^{\pi}$.

Here H is consumption of the agricultural good, and M is a CES-aggregate in manufacturing:

(2) $M = \left(\int_0^N \tilde{Q}_i^{(b-1)/b} di \right)^{b/(b-1)}$, where $\tilde{Q}_i \equiv Q_i q_i$, and $Q_i \equiv e^{-\rho T_i}$.

In eqn. (2), $b > 1$ is a constant elasticity of substitution, and N is the number of manufactured goods that are consumed each period. The quantity measure for each one, $\tilde{Q}_i$, takes into account the age T_i of each good that is consumed, as the age is embodied in the discount factor Q_i. Further, q_i is the share of the good that each consumer gets, so we set $q_i = 1$ when discussing the entire market. As in Krugman's model, the number of consumers can be scaled to one.

By letting this utility function apply in all periods, we make an important assumption about the consumer: He is "modern", as his perception of utility is related to the current state of productivity at any time. Utility is scaled so that the inner term in eqn. (2) equals one for a specific product if it is based on a new patent ($Q_i = 1$), and the consumer gets all of it ($q_i = 1$). No subjective time preferences in the traditional sense apply, so equilibrium follows by maximizing utility conditional on an income constraint each period, and by requiring optimal decisions, free entry, and full employment.

By a standard argument, the consumers will spend an income share π on manufactured goods, and $1-\pi$ on agriculture each period. The demand function facing each patent holder is

$$(3) \qquad \tilde{Q}_i = \left(\frac{P_i}{G}\right)^{-b} \frac{\pi Y}{G},$$

where G is a perfect price index

$$(4) \qquad G = \left(\int_0^N P_i^{1-b} di\right)^{1/(1-b)},$$

and Y is the rate of income, which is scaled to one each period. As in previous chapters, P_i is the price in terms of a good based on a new patent. Since each firm is small, and since trade costs will be of the iceberg type, it follows that a small firm can act as a monopolist with isoelastic demand in both markets.

The firm size is fixed, but as noted above, growth is embodied linearly in new patents as a fixed number of workers, L_A, are needed for one period to acquire a patent at any time. Similarly, L_C workers would be needed for one period to activate such a patent immediately.

Hence, the fixed quantity embodied in the patent i that L_A workers were able to make T_i periods ago, is only Q_i (≤ 1) times as large as the quantity embodied in a new patent. On the other hand, due to growing productivity, only $Q_i L_C$ ($\leq L_C$) workers will be needed to activate the old patent. Thus, if w is the wage rate in manufacturing, and we define $A = wL_A$ and $C = wL_C$, the total cost function for a firm contemplating entry in equilibrium becomes:

$$(5) \qquad J_i = A + C\tilde{Q}_i, \quad \text{where} \quad \tilde{Q}_i = 1 \cdot Q_i \ .$$

As in *Chapter 6*, we could have left out the tilda on Q_i in this equation, but we keep it to emphasize that this is a model with true product differentiation.

As also discussed in the previous chapter, a firm that does not activate a patent immediately will observe the demand function (3) as a firm-specific price that increases at a fixed rate $\rho/b \equiv \mu_{\text{det}}$ ($< \rho$). Hence, instead of being able to choose the relative quantity that it can supply at a specific point in time, the firm can use the option established by the patent, to slide along the demand curve. Then it is optimal to activate when time has brought the firm a fixed

markup price over the cost of activation. The unit cost of activation for a small firm in equilibrium is C from eqn. (5), so the markup price becomes

$$(6) \qquad P = \frac{\beta}{\beta - 1} C;$$

leaving out asterisks for optimal values, and following the convention of replacing the static elasticity b with the dynamic elasticity $\beta = \rho/\mu_{det}$, as we imagine that the firm makes its decision conditional on observed dynamic variables. As in all previous chapters, there is a unique, decreasing relationship between the price and the discount factor, so the optimal markup price equivalently can be stated in terms of the discount factor. Free entry follows from the zero profit requirement

$$(7) \qquad (P - C)\tilde{Q} - A = 0,$$

and by inserting eqn. (6), we also get:

$$(8) \qquad \tilde{Q} = \frac{L_A}{L_C}(\beta - 1).$$

Finally, the rate of entry follows from the condition $N(L_A + \tilde{Q}L_C)w = wL_M$, where L_M is the total number of workers. This simplifies to:

$$(9) \qquad N = \frac{L_M}{L_A \beta}.$$

In this deterministic case, N is the rate of entry, activation and consumption in equilibrium.

3. Two regions, trade costs and factor mobility

Regional demand

While agricultural goods can be freely traded, manufactured goods are exposed to *ad valorem* trade costs, characterized by a constant τ between zero and one, where $\tau = 0$ implies an infinite trade cost, and $\tau = 1$ implies zero trade cost. As discussed in *Chapter 6*, there are two technically equivalent interpretations of τ. One is the standard iceberg cost, by which a fixed

share of a shipped quantity arrives in the importing region immediately afterwards. Alternatively, τ can reflect that it takes a fixed time to transport the good.[1]

If the demand in region k for a typical product from region j is denoted by $\tilde{Q}_{jk}$, and P_j denotes the producer price, eqn. (3) yields domestic demand $\tilde{Q}_{jj} = P_j^{-\beta} G_j^{\beta-1} \pi Y_j$. Similarly, for the export market where only a fraction τ of the shipped quantity arrives, we have $\tau \tilde{Q}_{jk} = (P_j / \tau)^{-\beta} G_k^{\beta-1} \pi Y_k$, $j \neq k$. Total demand for a representative product from region 1, $\tilde{Q}_1 = \tilde{Q}_{11} + \tilde{Q}_{12}$, becomes:

$$(10) \quad \tilde{Q}_1 = P_1^{-\beta} G_1^{\beta-1} \pi Y_1 + P_1^{-\beta} G_2^{\beta-1} \tau^{\beta-1} \pi Y_2 .$$

Similarly, total demand for a representative region 2 good, $\tilde{Q}_2 = \tilde{Q}_{21} + \tilde{Q}_{22}$, becomes:

$$(11) \quad \tilde{Q}_2 = P_2^{-\beta} G_1^{\beta-1} \tau^{\beta-1} \pi Y_1 + P_2^{-\beta} G_2^{\beta-1} \pi Y_2 .$$

As eqns. (10) and (11) show, the price elasticity of demand is the same in both markets, implying that it is optimal for two firms in different regions to activate according to the same markup, with a corresponding discount factor. For this reason, the allocation of resources between the two sectors, patenting and production, will be the same in both regions. It follows that the total rate of entry and activation is independent of the distribution of firms, so eqn. (9) also applies in this two-region setting.

Analogous to Krugman (1991), we can use this to describe two types of equilibria: a short-run equilibrium with an arbitrary division of the manufacturing labor force, but no labor mobility, and a long-run equilibrium in which the workers are attracted to the region that offers the higher real wage. The discussion below focuses on the long-run equilibrium, while the other equilibrium is described briefly in an appendix.

[1] The actual delay is $\ln(1/\tau)/\rho$; see *Chapter 5*.

Stability of agglomeration

Assume that all manufacturing is gathered in a core, which we take as region 1. Below, we determine whether this is a stable equilibrium. With all firms in region 1, the price of manufactured goods will be a factor $1/\tau$ higher in region 2, as sales are split according to one producer price. Then $G_2 = G_1/\tau$, and eqn. (10) can be rewritten as

$$(12) \quad \tilde{Q}_1 = P_1^{-\beta} G_1^{\beta-1} \pi,$$

since $Y_1 + Y_2 = 1$. Similarly, eqn. (11) becomes:

$$(13) \quad \tilde{Q}_2 = P_2^{-\beta} G_1^{\beta-1} \pi \left(Y_1 \tau^{\beta-1} + Y_2 \tau^{-(\beta-1)} \right).$$

If new firms consistently enter in the core but gain no profit, we have $(P_1 - C_1)\tilde{Q}_1 - A_1 = 0$ in equilibrium. One of the firms would defect by choosing the periphery instead, if its profit were positive by doing so. Thus, agglomeration will be stable as long as

$$(14) \quad (P_2 - C_2)\tilde{Q}_2 - A_2 < 0.$$

The optimal price is a fixed markup over the cost of activation, which is proportional to the wage rate. Therefore we have $P_2 = (w_2 / w_1)P_1$, and $C_2 = (w_2 / w_1)C_1$. Similarly, if the patent of the defecting firm also has to be acquired in the periphery, we have $A_2 = (w_2 / w_1)A_1$. Then it follows from eqn. (14) that agglomeration will be stable as long as $\tilde{Q}_2 / \tilde{Q}_1 < 1$. Dividing (13) by (12), and inserting for P_2/P_1, this relative demand can be written as:

$$(15) \quad \tilde{Q}_2 / \tilde{Q}_1 = \left(\frac{w_2}{w_1} \right)^{-\beta} \left(Y_1 \tau^{\beta-1} + Y_2 \tau^{-(\beta-1)} \right).$$

Half of the farmers live in the periphery, and the agricultural consumption share equals (1 - π). Therefore $Y_1 = \frac{1}{2}(1+\pi)$, and $Y_2 = \frac{1}{2}(1-\pi)$. Further, a defecting firm must attract workers to the periphery for one period by a wage rate that is high enough. The workers *qua consumers* will have to purchase all their consumption goods in the periphery during this period, but the prices of manufactures are a factor $1/\tau$ higher in the periphery as long as all other firms are

located in the core. This must be compensated by a higher wage rate. However, as manufactured goods only represent a share π of total consumption, and the price of agriculture is the same in both regions, it suffices to pay a wage w_2 that is a factor τ^{π} higher than w_1. In the limit, this yields $w_2/w_1 = \tau^{\pi}$.

Inserting these results into eqn. (15), we find that agglomeration is stable if K defined by the following expression is less than one:

$$(16) \qquad K = \tau^{\pi\beta}\left\{\frac{1+\pi}{2}\tau^{\beta-1} + \frac{1-\pi}{2}\tau^{-(\beta-1)}\right\}.$$

We always have $K(\tau = 1) = 1$; i.e., location is irrelevant with no trade cost. Furthermore, it can be shown that K is U-shaped in τ if $\beta(1-\pi) > 1$. Figure 1 shows how K typically depends on τ in this case. Agglomeration is stable if the trade cost is smaller than τ^K. It is most stable for a trade cost τ^* further below. (The S-curve is explained later.)

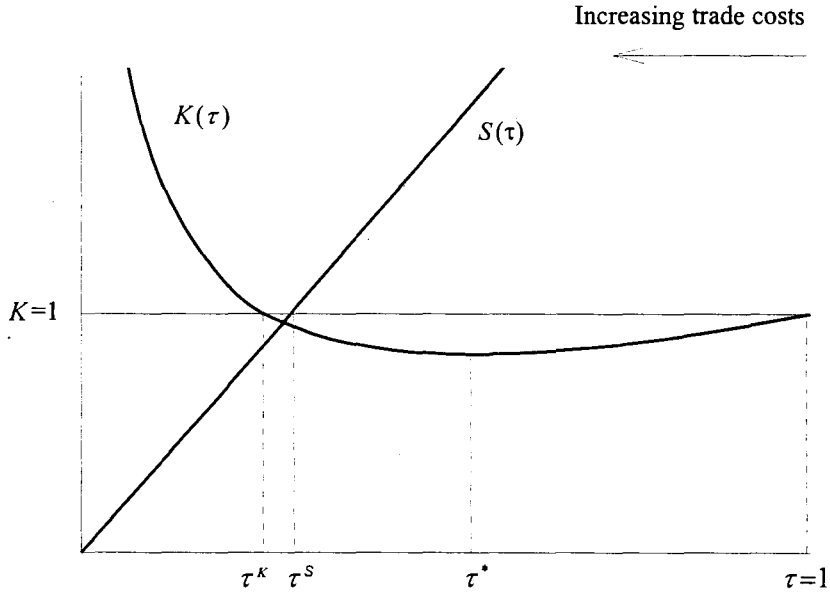

Figure 1. Stability of Agglomeration.

If $\beta(1-\pi) < 1$, economies of scale and the manufacturing share are so large that agglomeration is always stable. Then K is strictly increasing in τ, and $K(\tau) \to 0$ as $\tau \to 0$. (This case is not shown in Figure 1.)

In the static model there is no good reason why the two types of costs should be geographically separated. In the dynamic model, these costs represent quite different investments. In general, they are also separated in time. Thus, we should also consider the case where the defecting firm was allowed to acquire the patent in the core, but activated in the periphery, as this makes defection more attractive. L_A workers are needed for one period to acquire a patent with both sets of assumptions, but if these workers could remain in the core, they would accept a wage rate which is a factor $\tau^{-\pi}$ lower than in the periphery. Then we have $A_2 = A_1$ (as opposed to $A_2 = \tau^{-\pi}A_1$ in the calculations above), and the new boundary follows by multiplying K by the factor $\tau^{-\pi}$, yielding:

$$(17) \qquad \tilde{K} = \tau^{\pi(\beta-1)}\left\{\frac{1+\pi}{2}\tau^{\beta-1} + \frac{1-\pi}{2}\tau^{-(\beta-1)}\right\}.$$

Both of these alternative criteria for stability, $K<1$ and $\tilde{K}<1$, are stated by Krugman for the analogous static model, but he only explains and discusses the first one.

The $\tilde{K}$-factor has an intuitive interpretation in the dynamic model, by considering the entry cost in the core as an option investment that yields a return. Then $\tilde{K}$ is the return from the option to activate in the periphery relative to the return from the option to activate in the core. Thus, if $\tilde{K}>1$, it is optimal to defect. However, if the option investment must be increased by a factor $\tau^{-\pi}$ for the defecting firm (as entry must also take place in the periphery), but without changing the value of the option, the relative return becomes accordingly lower, and K appears.[2]

The shape of $\tilde{K}(\tau)$ is quite similar to $K(\tau)$ as in Figure 1, but the curve shifts upward, which makes agglomeration less stable. However, there is another, more significant difference, as it is easy to show that $\tilde{K}(\tau)$ is always U-shaped like this. Hence, if a defecting firm is allowed

[2] The analogous static model can be interpreted similarly, although we do not usually regard a fixed cost as an option investment in a static model. At least in this case, it actually is: It gives the firm the option to increase the price above the marginal cost.

to split its investments, we obtain the same result as in both models in *Chapter 5*, where agglomeration was never stable for a sufficiently high trade cost.

More importantly for our discussion below, it can be shown that K (and $\tilde{K}$) are increasing functions of β along the boundary that makes agglomeration stable; i.e., close to $K = 1$ (or $\tilde{K} = 1$). This implies that agglomeration becomes less likely if β increases. It can also be shown that K and $\tilde{K}$ are decreasing in π, so agglomeration becomes more likely the larger the manufacturing share.

Stability of a symmetric equilibrium

Krugman and Venables (1995) also find the boundary for stability of the symmetric equilibrium in the static model; i.e., where an equilibrium with half of the firms in each region, is replaced by agglomeration. We do not derive this criterion, but it obviously applies in the dynamic setting as well. It states that the symmetric equilibrium is stable if

$$(18) \qquad S = \tau^{\beta-1}\frac{(1+\pi)(\beta(1+\pi)-1)}{(1-\pi)(\beta(1-\pi)-1)}$$

is smaller than one.[3] We have $dS/d\pi > 0$, so a symmetric equilibrium becomes less likely by increasing the manufacturing share. Furthermore, $dS/d\tau > 0$, as also shown in Figure 1. Thus, the symmetric equilibrium is stable for a trade cost higher than τ^S, so for trade costs between τ^S and τ^K, there are three stable equilibria: one symmetric, and two with agglomeration. Finally, $dS/d\beta < 0$, so spread becomes more likely if β increases.

Actually, β is just another symbol for the measure of product differentiation, b, so the results above are indeed very similar to those obtained by Krugman. In this dynamic setting, however, the preference for variation will be observed via the growth rate and the development of prices. Before we discuss how to relate these observed variables to agglomeration, it is appropriate to extend the model slightly by including uncertainty.

---

[3] Referring to the discussion above, the symmetric equilibrium only applies if $\beta(1-\pi) > 1$.

Uncertainty

Uncertainty is not a part of Krugman's model, and the discussion above shows that it is not a requirement in this model either. However, it is so simple to extend the dynamic model to a stochastic environment that we briefly describe how it can be done, referring to *Chapter 6* for details.

In broad terms, we can allow the consumer's taste for specific products to fluctuate according to a set of independent, geometric Brownian motions instead of staying fixed. For a particular set of such taste processes, each firm will face an independent, geometric Brownian price process where both the drift μ ($< \mu_{\text{det}}$) and the volatility σ (> 0) are now to be interpreted as taste variables. Furthermore, β ($= b$) becomes the positive solution to the familiar quadratics:

(19) $\quad \frac{1}{2}\sigma^2 x(x-1) + \mu x - \rho = 0 .$

All equations above still apply, but the discount factor must be interpreted as an expectation. The utility function still describes a constant preference for variation (as b is constant), but in a large-group sense, as the taste for specific goods will change.

It is required that the same realization of each price process applies to all consumers. This is a reasonable assumption in the core-periphery model, where labor mobility is a cornerstone. In general, however, it would be interesting to allow for less than perfect correlation of firm-specific prices in different regions, preferably with less correlation the larger the trade cost.

Sensitivity

As noted above, we have $dS/d\beta < 0$, $dK/d\beta > 0$, and $d\tilde{K} / d\beta > 0$. Thus, agglomeration is more likely the smaller β. The relationship to the observed variables ρ, μ and σ follows from eqn. (19):

First, increasing ρ increases β, so a large growth rate is related to smaller product differences for fixed μ and σ. On the producer size, the value of holding patents decreases as a result of rapid depreciation, and agglomeration is less likely. It is misleading, however, to say that

growth causes spread, as μ, σ, and ρ are functionally linked to the specific preference for variation, characterized by the constant b. Thus, if ρ changes, μ or σ might adjust.

Second, increasing μ decreases β, as it counteracts the effect of depreciation embodied in ρ. It follows that agglomeration is more likely if the demand for specific products tends to grow.

Third, increasing σ also decreases β, as the probability of high prices increases. Thus, agglomeration is positively related to firm-specific price uncertainty as well.

The limiting behaviour of β was discussed in *Chapter 3*, and does not need to be repeated. Just note that neither uncertainty ($\sigma > 0$) nor expected price growth ($\mu > 0$) is required for product differentiation, scale economies and agglomeration. However, μ and σ cannot be zero (or very close to zero) at the same time. In that case the symmetric equilibrium is stable even for a very small trade cost, and agglomeration cannot be sustained.

The dynamics of changes

The dynamic interpretation of the core-periphery model gives a natural interpretation of structural changes. Consider a case where the starting point is a symmetric equilibrium, and a high trade cost. Fifty percent of the investments will then take place in each region. Then imagine that the trade cost decreases (τ increases). When τ reaches the level where the symmetric equilibrium becomes unstable, the one region that happens to get a head start will suddenly get all the new firms. A rapid shift towards agglomeration follows.

This argument can be expanded by using the (more realistic) extension that was discussed in *Chapter 3*: We can interpret a firm as an operating factory which faces a probability λ of sudden death each period. Then the expected lifetime of a firm is $1/\lambda$, so the total number of firms in each region is $N/2\lambda$ in the symmetric equilibrium. If the equilibrium becomes unstable, it will now take a long time to reach agglomeration if λ is small. Figure 2 plots a typical pattern for the number of firms during the transition.

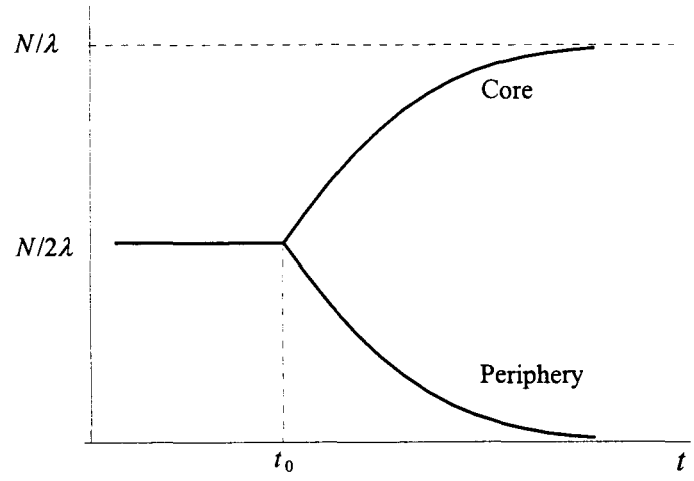

Figure 2. Transition from Symmetric Equilibrium to Agglomeration.

The symmetric equilibrium breaks at time t_0. Thereafter old firms in both regions exit at the same rate as new firms make entry, but all entries take place in the core. This gives a transition towards agglomeration, at a speed which is lower the smaller λ.

It does not seem to be crucial for this argument that the trade cost reductions are unexpected or very slow (and thus to be considered constant over the lifetime of a firm). In principle, we can interpret τ as an expected and discounted value that takes into account the downward trend.

4. Conclusions

Alternative interpretations

In the previous sections, each production unit was interpreted as a firm. This suggests that manufacturing is regarded as one industry consisting of a number of differentiated products. An objection to this might be that linkages created by massive labor mobility are unrealistic, since specific industries do not very often represent a large share of total consumption. However, we can apply a more aggregate interpretation by which a firm represents an industry. The total number then sums up to manufacturing as a whole, or all industries characterized by irreversible investment and specific demand.

Moreover, Krugman and Venables (1996) develop a stability criterion exactly like the K-factor (16), but based on a different interpretation of π. Up until now, π has been the

manufacturing consumption share, which indirectly represents the importance of the mobility of the workers who receive this share as income. In Krugman and Venables' alternative static model, supply and demand linkages are created by intermediate goods, and π represents the share of intermediate goods in production costs. However, as they put it, in both models this share "*...determines the importance of forward and backward linkages and thus of localized external economies*". It seems likely that an alternative interpretation of π also exists for the dynamic model.

Finally, an "alternative interpretation of the alternative model" is possible. That is, each production unit could be regarded as an industry, with π representing the overall production share of intermediate goods from all types of manufacturing.

Final remarks

There are two main conclusions to be drawn: First, we have developed a dynamic model that sheds light on the relationship between agglomeration and irreversible investment. Second, the approach shows how to develop a dynamic "twin" of a typical static Dixit-Stiglitz model.

With respect to the first point, the model has brought forward two sets of insight: Some well-known results from static modelling, like the non-monotonic relationship between trade costs and stability of agglomeration, were confirmed in a dynamic context. In addition, several observed dynamic variables have been related to agglomeration. To sum up, the model shows that agglomeration is more likely if we observe:

- intermediate trade costs (τ)
- large manufacturing shares (π)
- much firm-specific price uncertainty (σ)
- large positive trends in firm-specific prices (μ)
- small growth (ρ)

Krugman (1991) points out that, although the elasticity of substitution in the static model "*...is a parameter of tastes rather than technology, it can be interpreted as an inverse index of equilibrium economies of scale*". As discussed in previous chapters, β similarly acts as an inverse measure of dynamic economies of scale. This is also reflected in the growth rate (ρ),

that tends to decrease product differences in a dynamic respect although it is not a parameter of taste. On the producer side, it shows up as a parameter that decreases the value of patents. Uncertainty and expected growth of demand (σ and μ) show up as price parameters that increase the value of patents; i.e., dynamic market power.

The causality in the dynamic model must be emphasized. The fundamental economic force is the same as in the static model: The preference for variation enables production of goods based on economies of scale. Thus we should not say that agglomeration forces are created by firm-specific uncertainty and growth of demand, or that they are diminished by growth of productivity. However, observed combinations of these variables leading to a small β indicate a preference for variation, and agglomeration becomes more likely.

The approach can be generalized. One response could be to include operating costs, and consider temporary exit when prices are low. This requires a discussion of the negative root of eqn. (19). Dynamic location models with exit options are more realistic, but also more complicated. It is unlikely that an equilibrium model could be analyzed far without turning to simulations. To come around this, we attempt a simplified approach to entry and exit using a firm-level model in *Chapter 10*. Before that, the next chapter discusses agglomeration and endogenous growth by combining the results above with some of the results from *Chapter 4*.

Acknowledgements

I wish to thank Victor Norman and Anthony Venables for comments on this chapter.

APPENDIX: Equilibrium with no labor mobility

Below, we derive a set of equations that characterize equilibrium if labor mobility in manufacturing is excluded. The exposition is highly inspired by Krugman (1991).

Choose a measure of scale so that the wage rate in agriculture, as well as total income, equals one in all periods. Then the rate of total farmer income is $(1 - \pi)/2$ in each region, as noted in the text. The manufacturing wage rate in region j is w_j. Similarly, the number of workers is L_j, so the total income rate in each region becomes:

$$\text{(A.1)} \quad Y_j = \frac{1-\pi}{2} + w_j L_j, \quad j = 1,2.$$

As also discussed in the text, the relative rate of entry in manufacturing, N_j/N_k, is equal to the relative number of workers, L_j/L_k, in equilibrium. This result can be used to find how total income is spread, by considering the value of sales of a representative firm in each region.

The cost of activation is incurred in labor, so the optimal markup price is proportional to the wage rate. Further, the consumer price of a typical import good relative to a domestic good is a factor $1/\tau$ higher than the producer price. Thus, if the region j consumption share of a typical product from region k is denoted by q_{jk}, the demand functions (3) yield:

$$q_{jj} / q_{jk} = (P_j \tau / P_k)^{-\beta} = (w_j \tau / w_k)^{-\beta}, \; j \neq k;$$

i.e., relative consumption is isoelastic in the relative wage. To obtain the expenditure on a typical domestic good relative to an import good, this must be multiplied by the relative consumer price, which is :

$$(P_j \tau / P_k) = (w_j \tau / w_k), \quad j \neq k.$$

To obtain the total value of the region j rate of consumption on domestic goods relative to import goods, we must also multiply by the relative rates of entry:

$$N_j/N_k = L_j/L_k, \quad j \neq k.$$

Denoting the product of the three expressions above by Z_j, we find:

$$\text{(A.2)} \quad Z_j = \frac{L_j}{L_k}\left(\frac{w_j}{w_k}\tau\right)^{-(\beta-1)}, \quad j \neq k.$$

The total rate of worker income in each region equals the total expenditure rate on all manufactured goods produced there, as trade costs are embodied in the goods; thus:

$$\text{(A.3)} \quad w_j L_j = \pi\left(\frac{Z_j}{Z_j+1}Y_j + \frac{1}{Z_k+1}Y_k\right), j \neq k.$$

The six equations in (A.1), (A.2) and (A.3) determine the six variables w_j, Z_j, Y_j, j = 1,2, and describe an equilibrium analogous to the short-run equilibrium studied by Krugman in the static model. By the new interpretation, the transition to a long-run equilibrium with labor mobility follows directly from the dynamic setup.

References

Baldwin, Richard E. and Rikard Forslid, 1997, "The Core-Periphery Model and Endogeneous Growth", *CEPR Discussion Paper* 1749

Dixit, Avinash K. and Joseph E. Stiglitz, 1977, "Monopolistic Competition and Optimum Product Diversity", *American Economic Review* 67, 297-308

Fujita, Masahisa, Paul Krugman and Anthony J. Venables, 1998, *The Spatial Economy,* book manuscript

Krugman, Paul R., 1991, "Increasing Returns and Economic Geography", *Journal of Political Economy* 99, 3:483-499

Krugman, Paul R. and Anthony J. Venables, 1996, "Integration, specialization, and adjustment", *European Economic Review*, Papers and Proceedings 40, 959-967

Krugman, Paul and Anthony J. Venables, 1995, "Globalization and the Inequality of Nations", *The Quarterly Journal of Economics,* vol. CX, 4:857-880

Puga, Diego, 1996, "Urbanization patterns: European vs. less developed countries", *CEP Discussion Paper* 305, forthcoming in the *Journal of Regional Science*

PART IV

Miscellaneous

CHAPTER 8

Agglomeration and Growth

1. Introduction

This dissertation has been built on one simple observation that has proved to be important: The analogy between a discount factor and a static quantity measure. *Chapter 1* described the analogy at the firm-level, based on a standard investment problem. The idea has been explored in a number of directions, and the methodology does not seem to be exhausted by the presented applications. The objective of the final part, which is initiated by this chapter, is to indicate some extensions that can form the basis for future research.

In this chapter we discuss how it is possible to combine results by increasing the number of endogenous variables in one particular case, just noting that a number of similar extensions seem possible. More precisely, we will extend the core-periphery model in *Chapter 7* by endogenizing the growth rate as in *Chapter 4*.

It must be emphasized that the arguments in this chapter will not be rigid and formal, but intuitive and suggestive. No clear-cut conclusions are stated. This is partly because a more realistic description of the combined effects of agglomeration and growth would have to consider issues that are difficult to handle. In particular, the fundamental assumption on rational expectations can create problems. Nevertheless, some educated reasoning about the relationship between agglomeration and endogenous growth is attempted; partly because it is interesting in itself, and partly as it will illustrate the potential of the approach.

2. Agglomeration and growth

Several sets of assumptions were applied in *Chapter 4*, but here we shall just need the conclusions from the setting with geometric Brownian prices, and where ρ is interpreted as a growth rate that is positively related to the size of the industry.

In *Chapter 4* we were not concerned about whether product differences were just a matter of dynamics (as in *Chapter 3*), or whether the goods were true differentiated goods (as in *Chapter 6*). However, all equilibrium formulas are the same, and the growth model obviously applies in both cases. Thus, as long as we restrict the discussion to a uni-location economy, the only difference relates to individual welfare. With true product differentiation, there are two gains from increasing the size of the economy: A higher growth rate as well as more variation for the individual consumer. If the goods are homogeneous as in *Chapter 3*, the size of the economy has only growth effects.

Somewhat similarly, the growth rate in the dynamic core-periphery model could be exogenous as assumed in *Chapter 7*, but we might also consider it to be endogenous as in the *Chapter 4* growth model. The difference is mainly that β becomes endogenous if ρ is endogenous. However, since firms are small and hold ρ for fixed in any case, this does not affect their decisions, and the same equilibrium formulas apply.

Thus, let us interpret ρ in the core-periphery model as an endogenous growth rate that depends on the size of manufacturing. Further, assume that the spillovers that make up this growth rate do not cross the border between the two regions, and that a symmetric equilibrium has been established. Then the growth rate will be the same in both regions, but as manufacturing in each region is fairly small, the growth rate will also be small.

If the symmetric equilibrium becomes unstable as trade costs fall, we may see that suddenly all new firms enter in the one region that happens to get a head start. What can be expected to happen to the growth rate? It will increase in the core region, as the size of manufacturing increases, and, similarly, it will drop in the other region. Thus, we should expect a development as shown in Figure 1.

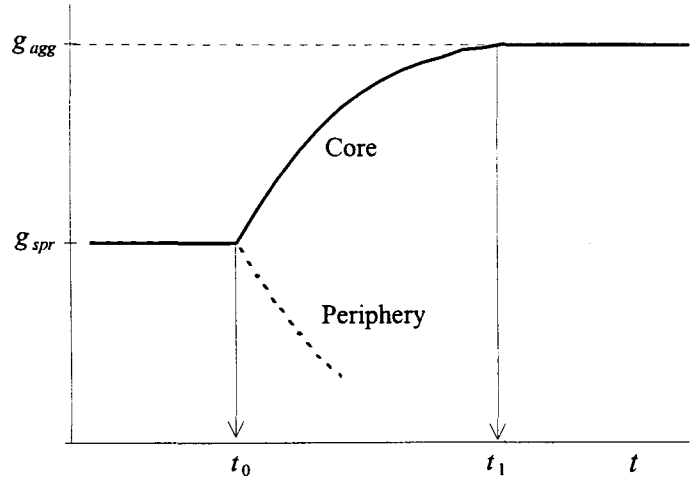

Figure 1. Growth Rates in a Transition from Symmetric Equilibrium to Agglomeration.

At time t_0, the symmetric equilibrium breaks, and the growth "takes off", based on agglomeration in one of the regions. Most likely, this will happen through a process that lasts for a long time, since firms are normally operating factories that are tied to specific locations by irreversible investments.[1] The rate of knowledge accumulation that follows from agglomerated production will grow according to the speed towards complete agglomeration. The transition will take place as firms dying in one region are replaced by new firms in the other region. The growth rate will increase from the initial level g_{spr}, and settle (almost asymptotically) at a higher level g_{agg} in one of the regions.

Is this the only possible outcome? That may depend on some philosophical considerations on how the consumers respond to increased growth. In our framework of product differentiation, the parameters of the price processes are defined conditional on a specific growth rate (ρ), and according to a specific taste for variation. If ρ increases due to an accidental event, a consumer will get access to a smaller number of product categories each period.[2] If he prefers the same number of categories as before, a parameter like μ must be adjusted upwards. However, the growth rate also increases, so in any case the consumer will experience a

[1] See *Chapter 3*, Section 5.

[2] See *Chapter 7*, eqn. (9). As ρ increases, β will also increase, so N decreases.

different development of welfare. Hence, other parameters might happen not be adjusted exactly like this.

In fact, if none of the parameters of the demand processes adjust (or if there is a time lag), the new equilibrium may be undermined even before it gets established. This is most easily observed if we apply the simple definition of the firm; i.e., a firm that disappears right after activation. In that case a symmetric equilibrium that breaks ought to be replaced very quickly by an equilibrium with complete agglomeration. However, in the new equilibrium the growth rate will be higher, which means that β will also be higher. For various combinations of parameters, we may then characterize the stability of equilibria by plotting the functions $K(\tau)$ and $S(\tau)$ as in *Chapter 7*. This is done in Figure 2 for two values of β. The resulting function pairs are (K1,S1) and (K2,S2).[3]

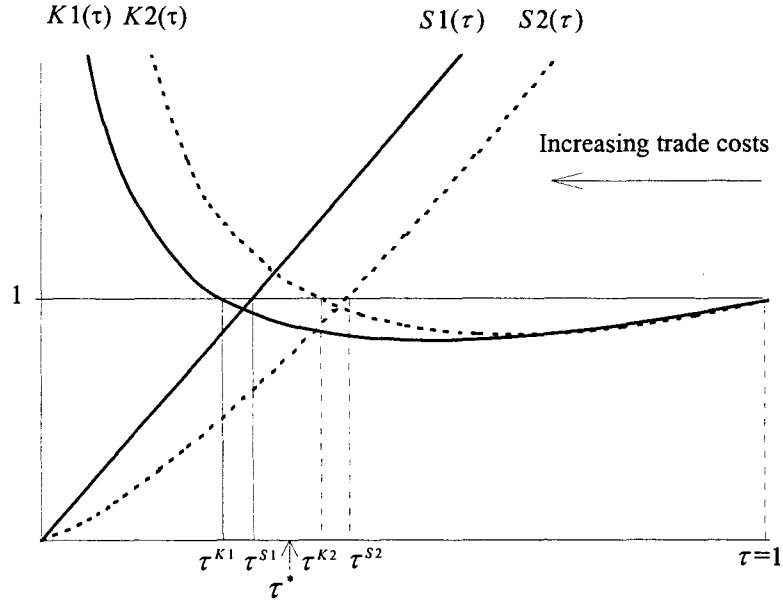

Figure 2. Instable Equilibria with Agglomeration and Growth.

Consider what will happen if the initial point is a symmetric equilibrium that corresponds to the first parameter set (K1,S1). This equilibrium will be stable as long as $\tau < \tau^{S1}$. Let the trade cost shift down to τ^*, e.g. by a technological shock or an unexpected political event.

[3] The numbers can be stated for reference: $\beta_1 = 2.0$, $\beta_2 = 2.3$, and $\pi = 0.2$.

Then the symmetric equilibrium breaks, and it should be replaced by an equilibrium with agglomeration. However, with agglomeration the growth rate increases, which implies that β increases. That leads K1 and S2 to shift to K2 and S2, respectively. As observed from the figure, agglomeration is no longer stable, and we may get a transition back to symmetry etc. etc.

Of course, this is a stylized story. It is highly unrealistic that firms and workers will jump back and forth in this manner. Nevertheless, the basic argument seems reasonable. The combination of agglomeration and growth forces may create industrial production patterns that are consistently changing, and that may be very difficult to predict. In particular, this will be true if trade costs, consumption shares, and growth rates are changing at the same time.

3. Final remarks

The results obtained by the extended core-periphery in this chapter cannot be taken literally due to the highly simplified assumptions. However, they indicate that growth may be induced by agglomeration, while growth at the same time may also discourage agglomeration. Thus, some counteracting forces seem to appear when combining agglomeration and endogenous growth.

Like the model with pecuniary externalities that has been discussed here and in *Chapter 7*, the *Chapter 5* model with true externalities gave the result that agglomeration typically is discouraged by growth. As a matter of fact, the main conclusions on agglomeration effects for the two types of externalities seem to coincide in all respects: The higher the firm-specific uncertainty (σ) and growth of demand (μ), the greater the dynamic economies of scale, and the more likely is agglomeration. The more growth (ρ) there is, however, the less likely is agglomeration. Furthermore, agglomeration becomes more likely in all models the larger the consumption share on goods characterized by dynamic economies of scale (π). Finally, all models confirm the well-known result that agglomeration is more likely for intermediate trade costs.

In *Chapter 3*, the patent cost was endogenized in several ways, e.g. by assuming that μ could be increased by increasing A. The main result obtained by these extensions can also be related

to agglomeration, as the efficiency of such efforts is measured by their effect on β. The more effective they are, the smaller β and the greater the economies of scale. Hence, agglomeration and possibly also growth might be encouraged, although the warnings with respect to consumer response to increased growth apply to this argument as well.

Policy implications have barely been discussed, and this chapter may have contributed to an explanation: In a setting with agglomeration forces, the optimal policy depends on the specific assumptions that are made; whether there are spillovers, where they may be located, whether the economy is close to a border of stability so that available political means have significant effects, etc.

Nonetheless, one result seems clear in this two-region model, when combining agglomeration and growth forces: If both regions hold on to manufacturing instead of letting it agglomerate, they are going to lose. Each region may be satisfied by being just as well off as the other one, but both regions will suffer from a low growth rate. If one of the regions gets all manufacturing, it will have higher wages due to trade costs, but this is just a difference in the level of income. Even the farmers in the periphery will gain from increased growth in the long run: they will just reach each specific level of real income a bit later than their colleagues in the core.

CHAPTER 9

Time Differentiation

1. Introduction

The equilibrium model with firm-specific demand and constant costs in *Chapter 3* was based on homogeneous goods, although some of its characteristics are fairly similar to those of static models with product differentiation and monopolistic competition. However, it was concluded that intra-industry trade and gains from scale, which are typical feature of models with monopolistic competition, will not occur. This conclusion was turned around in *Chapter 6*, where a minor technical change in the utility function brought about a true dynamic interpretation of the Dixit-Stiglitz model, preserving all scale effects of that model.

The two models look so similar that one has to ask: Can the *Chapter 3* model be adjusted and give rise to intra-industry trade without departing from homogeneous goods? This would yield a quite interesting case, as most trade models break one of the assumptions. There may be trade in homogeneous goods if there is imperfect competition and entry barriers, e.g. as in Cournot-oligopoly. There may also be trade in differentiated products with free entry as in monopolistic competition. However, the combination of intra-industry trade in homogeneous goods and perfect competition in the sense of free entry and zero profit is not usual.

In fact, this is partly because most trade models are static models. In a dynamic model such trade is not mysterious at all, and the objective of the current chapter is to describe how it can arise. We discuss this in two steps: First, the basic arguments are presented somewhat simplified. Then we show how the *Chapter 3* model with price variation can be adjusted to obtain the same effects in a formal equilibrium framework. Starting with the *Chapter 3* model, some convexities must be introduced, but a strictly concave periodic utility function will suffice. Since the number of goods is finite, intra-industry trade can arise as the goods are available at different times. "Time differentiation" seems to be a proper term.

2. A simplified discussion

The easiest way to demonstrate gains from intra-industry trade in homogeneous goods is by considering a simple two-region, one-sector model with a patent cost and a production cost, as usual. Suppose that it takes exactly one period before the patent can and must be used, like a crop that must be harvested. Discounting and all other dynamics are neglected. Furthermore, imagine that patenting must take place at some minimum scale, and that each region is so small that it does not permit more than one of the two activities each period. Then patenting and production will take place every other period in both regions. With n consumers, each one will get a share $1/n$ of the production that period, and nothing in the other period.

If the periodic utility function is strictly concave, it is obvious that the consumers will prefer a share $1/2n$ every period as opposed to $1/n$ every other period. This is exactly what will happen if trade is allowed: When there is patenting in one region there will be production in the other, and homogeneous goods will be traded in both directions, but at different times.

With no trade cost and full symmetry between the regions, half of the production will be traded no matter how concave the utility function. The amount of trade ought to be smaller the larger the trade cost, and the less concave the utility function.

Uncertainty can also create gains from trade in this respect. Consider a consumer with a periodic utility function $u(x) = x^a$, where x is the quantity of a specific product, and a is a positive constant less than one. Furthermore, let it be uncertain when production can take place after the patent has been acquired. However, the goods are perfect substitutes in a static sense; i.e., when they are consumed. Flowers may serve as an example: Two gardens with the same type of flower may flourish at different times even if the bulbs have the same age.

To simplify, consider a world with just three periods, and neglect discounting. All patents are acquired in period one, and each good matures according to an independent stochastic process. This process is very simple: the good must be produced in one of the two remaining periods with equal probability. Then it follows from the concave utility function (and Jensen's inequality) that a risk-neutral consumer will prefer to make contracts with several manufacturers of such goods if the unit price were the same.

For example, if two firms plan to produce a quantity x each, the two lots will be available in the same period or in different periods with equal probability. The expected utility from a consumption right to the entire production of one firm becomes

$$(1) \qquad E_1[u] = \frac{1}{2}x^a + \frac{1}{2}x^a = x^a,$$

while the expected utility from a consumption right to *half* of the production of *both* firms is:

$$(2) \qquad E_2[u] = \frac{1}{2}x^a + \frac{1}{2}\left(\left(\frac{x}{2}\right)^a + \left(\frac{x}{2}\right)^a\right) = \left(\frac{1}{2} + \frac{1}{2^a}\right)x^a.$$

E_2 is larger than E_1 provided $a < 1$. Similar formulas can be developed showing that the consumer is even better off by spreading a fixed total volume over more contracts. The marginal value of an extra contract decreases with the number of contracts, since the variance in the distribution of consumption between the two periods is smaller the more contracts.[1]

These examples illustrate the qualitative character of the argument for gains from scale and intra-industry trade in homogeneous goods. Below, we frame the idea in an equilibrium context with free entry and large-group assumptions as in the *Chapter 3* model with geometric Brownian price variation.

3. A formal description

Let periodic utility be as follows:

$$(3) \qquad u_t = \int_{-\infty}^{t}\int_{0}^{N_s}\int_{0}^{M}\left(\int_{0}^{Y_m}\delta_{i,s}(m-m_{i,s})\cdot\delta_{i,s}(t-s-T_{i,s})\cdot c_{i,s}(t)\cdot Z_{i,s}(t-s)dy\right)^{(b-1)/b} dm\,di\,ds.$$

---

[1] The gain from increasing the number of contracts can be illustrated by the function $U(N) \equiv E[u(x/N)]_{x=1}$. This can be calculated for various N using the binomial distribution. Eqn. (1) yields $U(1) = 1$, and $U(2)$ follows from eqn. (2). For example, if $a = 0.5$, we get $U(2) = 1.21$, $U(3) = 1.30$ and $U(4) = 1.34$. Thus, $U(N)$ is increasing and concave as expected.

Most symbols in (3) are defined as in *Chapter 3*, but let us go through the whole list as there are some new ones as well: A large number of patents from previous entry times s may be activated, and we have to integrate (ds) throughout history up to the current time t to include all patents. As before, the rate of entry is N_s, but now it is assumed that each period (a year, say) is divided into a large number of shorter intervals (like 365 days). We keep track not only of the *long* period t at which a specific patent i from the long period s is activated, but also the *short* period $m_{i,s}$ within t at which this occurs. There are M short periods for each long period. For each short period, m, consumption is integrated over the total number Y_m of goods that are consumed. As usual, we neglect discounting within each long period, and, of course, also within each short period.

No good contributes to utility in other short periods than the one in which the patent is activated. Therefore the Dirac functions $\delta_{i,s}(m-m_{i,s})$ are included to ensure that the function of integration is zero unless $m = m_{i,s}$. As in *Chapter 3*, we also need functions $\delta_{i,s}(t-s-T_{i,s})$ to remove similar gains if the long period of activation does not coincide with the current one.

The consumption of each good is $c_{i,s}$. Furthermore, a firm-specific demand process $Z_{i,s}$ applies to each firm. In the following, this is described by a transformed process $X_{i,s} = Z_{i,s}^{(b-1)/b}$ for notational convenience.[2] The argument (t - s) corresponds to the age of the process at time t. Finally, $b > 1$ is a constant. This makes the utility concave in each short period, implying a preference for spreading consumption over many such periods.

Now we make a little trick: Assume that M is very large compared with N. Then the probability of two goods being activated in the same short period can be neglected. It follows that Y_m will be either zero or one.[3]

---

[2] If $Z_{i,s}$ is geometric Brownian, as assumed further below, it can be shown by Ito's lemma that $X_{i,s}$ is also geometric Brownian, but with a different drift and volatility.

[3] Technically, this is not a strong assumption. With stochastic, firm-specific demand processes $X_{i,s}$, no firms will activate simultaneously even if the patents were acquired at the same time. Then the assumption causes no trouble. Nor will there be problems if $X_{i,s}$ grows at a deterministic rate. Then the firms will benefit from acquiring patents at different times. For this reason they will also activate their patents at different times.

Due to the δ-functions, since M is large, and since N is constant in equilibrium, the utility function simplifies to

$$(4) \qquad u_t = \int_0^{N_t} c_i^{(b-1)/b} X_i di ,$$

where X_i is the state of the demand process for those patents that are activated in period t. The major difference from the *Chapter 3* model is the utility measure in terms of quantity, as c_i has been replaced by $c_i^{(b-1)/b}$. It may seem as if the utility is strictly concave in each good, as with product differentiation in *Chapter 6*, but this is not the case. The concave utility in each short period is just reflected in one good, since no more than one good will be consumed.

If X_i is geometric Brownian and the cost of activation is constant, a fictitious aggregate utility function can be found by similar arguments as in *Chapter 3*, yielding:

$$(5) \qquad U = \int_0^{N} c_i^{(b-1)/b} Q_i^{(\beta-1)/\beta} di .$$

In principle, it is possible to proceed from here by optimizing utility in two dimensions (quantity and time). It would be nice if we could be sure that $\beta = b$, but can we though? As in *Chapter 6*, this seems to depend on the interpretation. If ρ is interpreted as a growth rate and preferences are continuously adjusted, then it is hard to see why a parameter $\mu_{\text{det}} = \rho/b$ could not arise in just about the same manner as for differentiated products. Actually, μ_{det} would not be a taste parameter for specific goods, but it would be perceived as this because no more than one good would be available in each short period. This gives value from waiting and markup pricing, and firm-specific uncertainty could also be included as in *Chapter 6*.

As observed, this model also simplifies to something similar to monopolistic competition. However, the reason that dynamic market power is observed is new: There is no preference for different product categories; just for smooth consumption of homogeneous goods.

The relationship between the various models can be expressed more clearly as follows: The preferences for variation in the static model can be stated in compact form by a utility function $U(q) = Nq^{(b-1)/b}$, where b (> 1) reflects a *preference for consumption of different*

products at the same time. In the *Chapter 6* model, such preferences are fulfilled through time, but the consumer has no time preferences except that his perception of utility is always related to the current state of productivity. In the model that has been sketched here, the same utility function applies, but b reflects a *preference for consumption of the same product at different times*. Thus, the basic mathematical description coincides in all these models, so the difference becomes mainly a matter of interpretation.

To conclude, we may see gains from scale and intra-industry trade even in an industry where a large number of firms are producing homogeneous goods, simply as trade can make consumption more smooth. In principle, agglomeration forces might also arise.

4. Final remarks

Is this a realistic model? At first sight, it does not seem quite as interesting as the previous ones. At least I have trouble finding good empirical examples to prove that it could be important for the understanding of phenomena like agglomeration.

However, many of its general characteristics are of interest. For example, dynamic uncertainty is obviously important for a lot of trade in homogeneous goods. At the international level, grain may be transported in either direction between two countries depending on the harvest in different years. At the national level in Norway, there may be two-way trade in hydroelectric power, depending on the development of magazined water, stochastic regional demand, time of the day etc. At the local level, those who love fresh crabs may have to call on the fisherman who got some in his pot recently, even though there may be a more local one, but who has no crabs to sell at the moment; thus trade costs apply.

The examples show how an initial investment (in sowing, power plants, fish pots etc.) followed by time and uncertainty can create gains from intra-industry trade in homogeneous goods. Essentially, they contain no news, as similar examples can be found in any standard textbook in insurance theory. The main contribution of this chapter has been to show how they can be modelled in an equilibrium context by use of a new approach.

Acknowledgements

I wish to thank Victor Norman for helpful suggestions to this chapter.

CHAPTER 10

Geographical Entry-Exit

1. Introduction

Most location decisions are at least partly irreversible. If the analysis of such decisions is to be realistic, it must include issues like dynamic price or cost uncertainty that may affect decisions on whether or where to invest. For location to be a relevant matter, there must also be trade costs. In addition, some kind of scale economies are needed for activities that are not completely tied to immobile production factors.

The equilibrium models in *Chapter 3* were based on homogeneous goods, and they did not embody economic forces that make location a matter of interest. However, later chapters have demonstrated several extensions by which this picture changes. In *Chapter 7* it was also noted that effects from dynamic uncertainty and irreversibility should not be limited to entry decisions. Usually there are operating costs, and firms often have other options than just waiting or investing in one region. In particular, they may be able to produce in more than one region, and establish or shut down plants depending on the development of dynamic variables.

Equilibrium analysis of such issues is hard, and will not be attempted here. Instead, we shall develop a simple firm-level model that hopefully may give some ideas for future research. Equilibrium effects will just be addressed informally in the final section.

Furthermore, the analysis is not based on the new methodological approach that has been the cornerstone of the thesis. It might be possible to do so, but here we stick to the standard smooth pasting approach.

The setting is as follows: We consider a firm that is a monopolist in a static as well as a dynamic respect, with the exclusive right to produce and sell a specific product in any quantity. The demand fluctuates randomly, but identically in two regions. (A symmetric extension to an arbitrary number of regions will also be discussed.)

An irreversible entry cost and a fixed operating cost are needed for each plant, while there may also be a constant marginal operating cost. Thus the firm's total production costs are minimized if it produces in just one region. However, trade barriers will also make split production a matter of interest.

All trade barriers are lumped into a net *ad valorem* cost τ between zero and one. If P is the net value of sales in the domestic region (i.e., revenue less marginal operating costs), only τP ($\leq P$) will be obtained from export sale. Since there is a fixed operating cost, spread production will then be profitable if the demand for the product is very high, as in that case the effective trade cost will be correspondingly high.

The following example shows how uncertainty of demand may transfer to fluctuating values of sales in this respect: Assume that inverse demand is given by $p = X \cdot (a - bq)$, where a and b are positive constants, q is the quantity, p is the unit price, and X is the state of a firm-specific demand process. Let marginal operating costs be zero, and exclude all kinds of stocking. The firm will choose a scale of production for which the marginal revenue is zero; i.e., $q = a/2b$. Then it obtains a net value of sales $P \equiv pq = X \cdot (a^2/4b)$, and the flow of optimum sales values will follow from the demand process (except for the constant terms). Thus, uncertainty of demand can be described via P, which for simplicity is denoted as a price.

If the initial situation is high demand and production in one or both regions, the firm can shut down plants in the case of a decline. If demand rises again, it is possible to re-invest. The objective is to study how decisions on when to establish or close production plants are affected by various cost elements and uncertainty of demand.

The approach is highly inspired by two models from Dixit (1988,1989) - the "entry-exit" model and the "entry-exit-scrapping" model. Following the terminology of Mossin (1968), Dixit studies when a ship should be operated, laid-up or scrapped conditional on stochastic operating profits and irreversible transition costs. My model has three states somewhat parallel to those of the entry-exit-scrapping model. They correspond to the number of active plants, as the number of plants will be either zero, one or two if there are two regions.

However, the system of equations will be slightly different, but easier to analyse than in Dixit's model. The reason is that production never starts via the intermediate state in the entry-exit-scrapping model, while it always does so in the model below. In the entry-exit-

scrapping model, a new ship is never built unless operating profits are positive. Thus, disregarding the time it takes to build the ship, it is always optimal to enter the market immediately.

My model looks more like a ladder on which the firm chooses a higher step the larger the demand; it will not produce if demand is very low, it will produce in one region at an intermediate level, and in both regions if demand is high. Starting from below, production will always be initiated from one plant (i.e., the intermediate state). It turns out that the optimal decisions can be described by two sets of solutions to the simpler entry-exit model. This also simplifies the extension to an arbitrary number of regions.

2. The model

The economy consists of two identical regions, and the firm can establish or close production in each of them separately. A number of cost elements are associated with each plant: First, an entry cost A is incurred at time of construction. Second, an exit cost B is necessary to close the plant. Both are fixed irreversible investments. Actually, we might have $B < 0$ as the plant can have a scrap value, but $A + B > 0$ is required to obtain some irreversibility. There is also a fixed operating cost C.[1] In addition, we might envisage a marginal operating cost. However, we simplify as described in the introduction by assuming that the latter cost is embodied in the net value of sales, P, for an optimal scale of production.

If the firm is not producing, no costs are incurred and no revenue obtained. Nonetheless, the firm still has a value due to the option to invest, as production can be profitable in the future. If the firm is producing in one region, the operating cost per period is C. The similar cost is $2C$ if two plants are active; one in each region.

The firm can switch between the three alternative states of production (indexed by 0, 1 and 2) depending on trade costs, costs of switching, and the current price P. The price is geometric Brownian

(1) $dP = \mu P dt + \sigma P dz$,

[1] As in Dixit's model, we might have two operating costs - one (C) that applies when the plant is active, and a smaller one (C_0) when it is not. It can be shown that only the difference matters, so we set $C_0 = 0$.

with the familiar interpretation of variables. By a standard dynamic programming argument, the following differential equations can be shown to hold:

(2) $\frac{1}{2}\sigma^2 P^2 V_0'' + \mu P V_0' - \rho V_0 = 0$

(3) $\frac{1}{2}\sigma^2 P^2 V_1'' + \mu P V_1' - \rho V_1 + P(1+\tau) - C = 0$

(4) $\frac{1}{2}\sigma^2 P^2 V_2'' + \mu P V_2' - \rho V_2 + 2(P - C) = 0$.

Here $V_i = V_i(P)$ is the value of a firm in state i, and ρ is the discount rate. Eqn. (2) holds with no production, eqn. (3) with production in one region, and eqn. (4) with production in two regions. The three leftmost terms represent the homogenous part of any such equation for the value function with geometric Brownian prices. The other terms describe the flow of profits in each state: In state 0, there is no profit. In state 1, the profit is the net value of sales from two regions taking into account trade costs, and less operating costs for one plant. In state 2, there are two plants, each producing for the local market; thus there are no trade costs, but operating costs are twice as large as with concentrated production.[2] The equations have the following solutions:

(5) $V_0(P) = b_0 P^{\beta}$

(6) $V_1(P) = b_1 P^{\beta} + a_1 P^{-\alpha} + \frac{1+\tau}{\rho - \mu} \cdot P - \frac{C}{\rho}$

(7) $V_2(P) = a_2 P^{-\alpha} + \frac{2}{\rho - \mu} \cdot P - \frac{2C}{\rho}$.

Here α and β follow from the common homogeneous part. As in previous chapters, α is the magnitude of the negative root, and β is the positive root (exceeding unity) of the characteristic equation

(8) $\frac{1}{2}\sigma^2 x(x-1) + \mu x - \rho = 0$.

---

[2] In general, optimal production and sale in each region may depend on whether production is concentrated or spread, and the assumption that trade costs can be represented as a percentage reduction in P may be violated. We disregard such complexities (which do not arise if the interpretation of P is as simple as in the introduction).

The constants a_1, a_2, b_0 and b_1 will be determined by value matching and smooth pasting conditions. Two other constants, a_0 and b_2, have been set to zero by limiting arguments considering the value of the firm when P approaches zero in state 0 and infinity in state 2.

Note the interpretation of eqns. (5-7). In eqn. (6), the first term represents the value of the option to invest A, and get better access to the export market (if the price gets high). The second term represents the similar value of the option to exit the market (if the price gets low). The third term is the expected net value of sales by continuing concentrated production forever, and the final term is the cost of doing so. Eqns. (5) and (7) have similar interpretations. The value matching and smooth pasting condition are:

Between state 0 and state 1

(9) $V_0(R_1) = V_1(R_1) - A$

(10) $V_0'(R_1) = V_1'(R_1)$

(11) $V_1(L_1) = V_0(L_1) - B$

(12) $V_1'(L_1) = V_0'(L_1)$

Between state 1 and state 2

(13) $V_1(R_2) = V_2(R_2) - A$

(14) $V_1'(R_2) = V_2'(R_2)$

(15) $V_2(L_2) = V_1(L_2) - B$

(16) $V_2'(L_2) = V_1'(L_2)$.

Here R_i and L_i are the thresholds at which one should enter (up to) or exit (down from) state i. This gives eight equations to determine the constants a_1, a_2, b_0, b_1, L_1, L_2, R_1, and R_2:

(17) $$-bR_1^{\beta} + a_1 R_1^{-\alpha} + \frac{1+\tau}{\rho-\mu} R_1 - \frac{C}{\rho} - A = 0$$

(18) $$-b\beta R_1^{\beta-1} - a_1 \alpha R_1^{-\alpha-1} + \frac{1+\tau}{\rho-\mu} = 0$$

(19) $$-bL_1^{\beta} + a_1 L_1^{-\alpha} + \frac{1+\tau}{\rho-\mu} L_1 - \frac{C}{\rho} + B = 0$$

(20) $$-b\beta L_1^{\beta-1} - a_1 \alpha L_1^{-\alpha-1} + \frac{1+\tau}{\rho-\mu} = 0$$

---

(21) $$-b_1 R_2^{\beta} + aR_2^{-\alpha} + \frac{1-\tau}{\rho-\mu} R_2 - \frac{C}{\rho} - A = 0$$

(22) $$-b_1 \beta R_2^{\beta-1} - a\alpha R_2^{-\alpha-1} + \frac{1-\tau}{\rho-\mu} = 0$$

(23) $$-b_1 L_2^{\beta} + aL_2^{-\alpha} + \frac{1-\tau}{\rho-\mu} L_2 - \frac{C}{\rho} + B = 0$$

(24) $$-b_1 \beta L_2^{\beta-1} - a\alpha L_2^{-\alpha-1} + \frac{1-\tau}{\rho-\mu} = 0 .$$

Here $a = a_2 - a_1$, and $b = b_0 - b_1$. This set of equations consists of two separate, but almost equivalent groups. The first four equations determine b, a_1, L_1 and R_1. The rest determine b_1, a, L_2 and R_2. Also, the two sets are identical to those of the entry-exit model, see Dixit (1988), except for one point: The entry-exit model yields terms with $1/(\rho-\mu)$ whenever the first group has $(1+\tau)/(\rho-\mu)$, and whenever the last group has $(1-\tau)/(\rho-\mu)$. The transformation

$$R = R_1(1+\tau), \quad L = L_1(1+\tau)$$

brings the first system (17-20) back to the standard entry-exit model with new coefficients:

$$\tilde{b} = b(1+\tau)^{-\beta}, \quad \tilde{a} = a_1(1+\tau)^{\alpha} .$$

A similar transformation can be made to eliminate the $(1-\tau)$ terms in the last system, and we get the following general solutions depending on trade costs:

(25) $$L_1(\tau) = \frac{L}{(1+\tau)}, \quad R_1(\tau) = \frac{R}{(1+\tau)}$$

(26) $$L_2(\tau) = \frac{L}{(1-\tau)}, \quad R_2(\tau) = \frac{R}{(1-\tau)} .$$

Here L and R are solutions to the entry-exit model. As for that model, it can be verified that the expressions for entry and exit also hold for Marshallian assumptions, which appear in the limit as $\sigma \to 0$.

3. Sensitivity

The relationship between the thresholds (L,R) and the exogenous parameters can be studied by some numerical experiments. Figure 1 plots how the thresholds depend on τ if other values are as in Dixit (1988): $\mu = 0$, $\rho = 0.05$, $\sigma = 0.05$, $A = 288$, $B = 0$, $C = 20$.

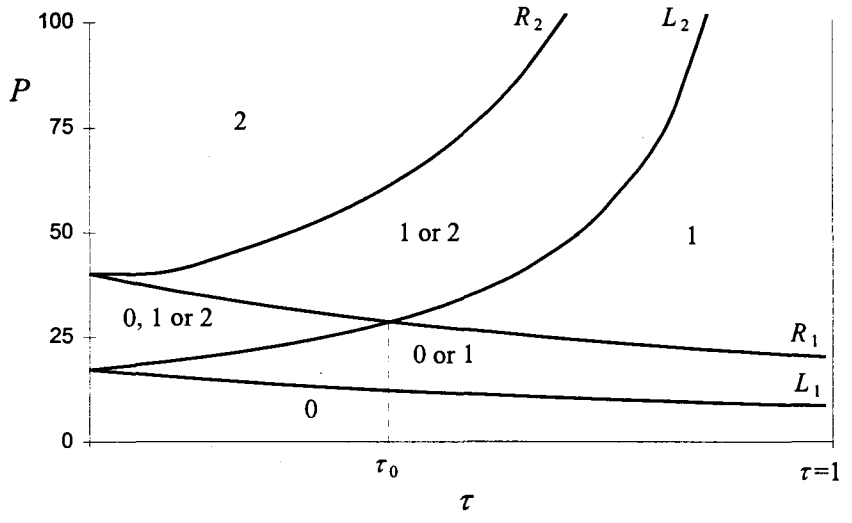

Figure 1. Entry and Exit in a Two-Region Monopoly Model with Trade Costs.

The figure also indicates how many production plants that can occur for specific combinations of price and trade costs. As expected, concentration is more likely the smaller the trade cost, but a unique structure of production only comes out if the price is very low or very high, or if the trade cost is small.

The intersection between L_2 and R_1 may be of some interest. It represents the highest trade cost (τ_0) for which concentration could be unique. From (24) and (25) the following expression for τ_0 is found:

$$(27) \qquad \tau_0 = \frac{(R/L)-1}{(R/L)+1}.$$

Thus, τ_0 is constant for proportional parameter shifts in R and L, while it is an increasing function of R/L.

How does R_i and L_i (and τ_0) depend on exogenous parameters? Higher uncertainty increases R_i and decreases L_i, thereby encouraging status quo and discouraging concentration as a unique solution. Figure 2 shows how the curves shift if σ is increased from 0.05 to 0.20. To avoid too many curves, only R_1 and L_2 are plotted. (R_2 and L_1 shift similarly.)

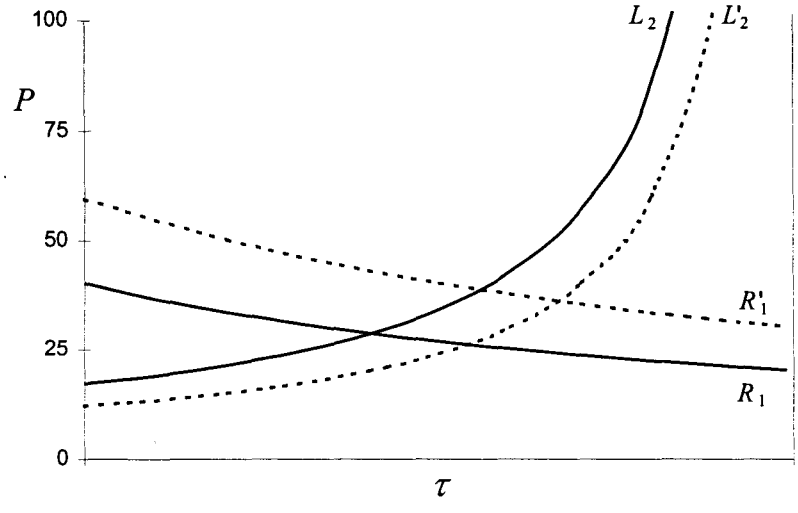

Figure 2. Effect of Increased Uncertainty ($\sigma = 0.05$, $\sigma' = 0.20$).

As the figure shows, higher uncertainty gives more hysteresis. The next figures plot similar curve shifts if the drift (μ) is changed from zero to 0.04 and -0.04, respectively.

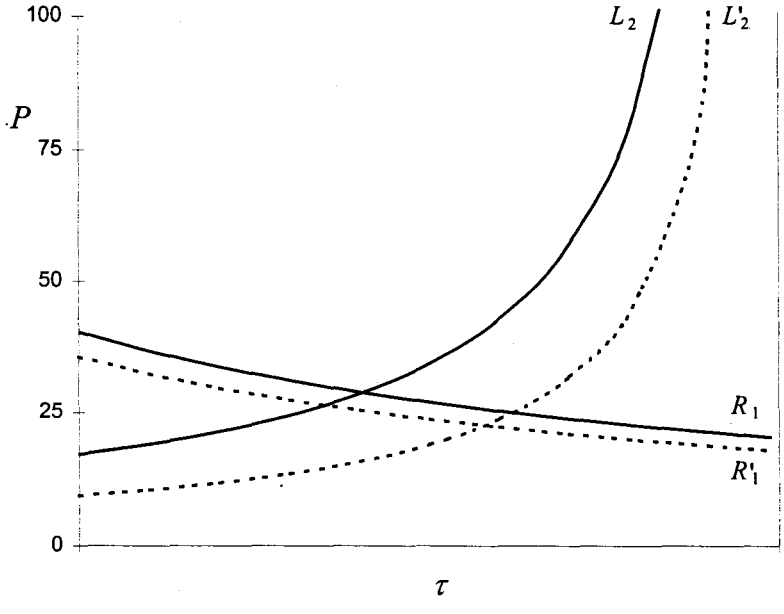

Figure 3. Effects of Expected Price Growth (μ=0, μ'=0.04).

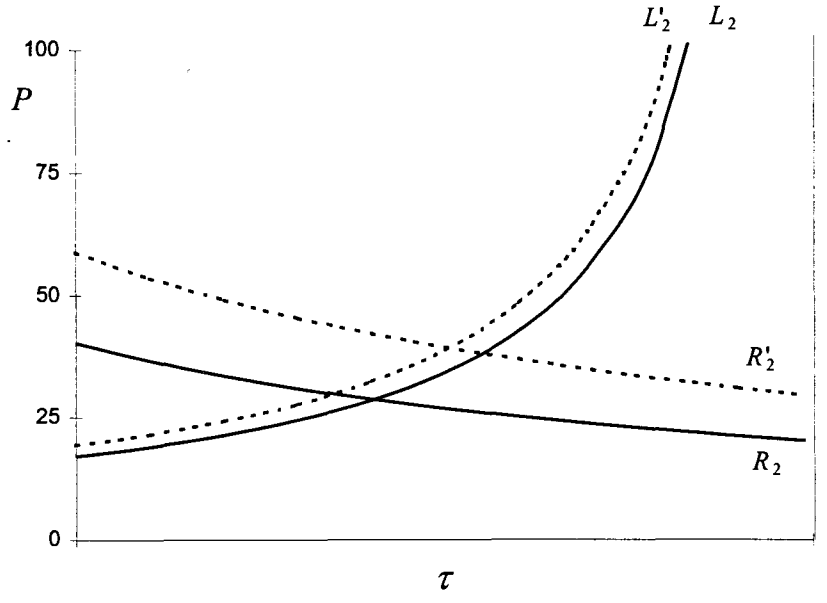

Figure 4. Effects of Expected Price Reductions ($\mu = 0$, $\mu' = -0.04$).

Figures 3 and 4 indicate that τ_0 is U-shaped in μ, and it can be shown that $\tau_0(\mu)$ reaches a minimum value for μ slightly smaller than zero. (The importance of this result is not clear.)

For the entry-exit model, Dixit (1988) derives the following approximation, which holds if σ and $(A + B)$ are small:

$$(28) \qquad \ln\left(\frac{R}{L}\right) = \left(12\sigma^2 \cdot \frac{A+B}{2C+\rho(A-B)}\right)^{1/3}$$

This can be regarded as a measure of hysteresis. The formula shows that the level of hysteresis is increasing in σ and B, and decreasing in C and ρ (as long as $A > B$, which is reasonable). The relationship to A is ambiguous.[3] Referring to eqn. (27), the same results apply to τ_0, as the ln-function is strictly increasing.

[3] More precisely, hysteresis effects increase with A if $B < C/\rho$, and decrease with A if $B > C/\rho$.

4. Many regions

The model can be extended quite easily to a multi-region setting if the trade cost (τ) is the same between each pair of regions. The pattern will be clear by studying the case with three regions, where the characterizing equations are:

(29) $\frac{1}{2}\sigma^2 P^2 V_0'' + \mu P V_0' - \rho V_0 = 0$

(30) $\frac{1}{2}\sigma^2 P^2 V_1'' + \mu P V_1' - \rho V_1 + P(1+2\tau) - C = 0$

(31) $\frac{1}{2}\sigma^2 P^2 V_2'' + \mu P V_2' - \rho V_2 + P(2+\tau) - 2C = 0$

(32) $\frac{1}{2}\sigma^2 P^2 V_3'' + \mu P V_3' - \rho V_3 + 3(P-C) = 0$.

Value matching and smooth pasting give the following equations to determine entry and exit:

(33) $-bR_1^{\beta} + a_1 R_1^{-\alpha} + \frac{1+2\tau}{\rho-\mu} R_1 - \frac{C}{\rho} - A = 0$

(34) $-b\beta R_1^{\beta-1} - a_1 \alpha R_1^{-\alpha-1} + \frac{1+2\tau}{\rho-\mu} = 0$

(35) $-bL_1^{\beta} + a_1 L_1^{-\alpha} + \frac{1+2\tau}{\rho-\mu} L_1 - \frac{C}{\rho} + B = 0$

(36) $-b\beta L_1^{\beta-1} - a_1 \alpha L_1^{-\alpha-1} + \frac{1+2\tau}{\rho-\mu} = 0$

---

(37) $-\bar{b}R_2^{\beta} + aR_2^{-\alpha} + \frac{1-\tau}{\rho-\mu} R_2 - \frac{C}{\rho} - A = 0$

(38) $-\bar{b}\beta R_2^{\beta-1} - a\alpha R_2^{-\alpha-1} + \frac{1-\tau}{\rho-\mu} = 0$

(39) $-\bar{b}L_2^{\beta} + aL_2^{-\alpha} + \frac{1-\tau}{\rho-\mu} L_2 - \frac{C}{\rho} + B = 0$

(40) $-\bar{b}\beta L_2^{\beta-1} - a\alpha L_2^{-\alpha-1} + \frac{1-\tau}{\rho-\mu} = 0$

---

(41) $$-b_2 R_3^{\beta} + \bar{a} R_3^{-\alpha} + \frac{1-\tau}{\rho-\mu} R_3 - \frac{C}{\rho} - A = 0$$

(42) $$-b_2 \beta R_3^{\beta-1} - \bar{a}\alpha R_3^{-\alpha-1} + \frac{1-\tau}{\rho-\mu} = 0$$

(43) $$-b_2 L_3^{\beta} + \bar{a} L_3^{-\alpha} + \frac{1-\tau}{\rho-\mu} L_3 - \frac{C}{\rho} + B = 0$$

(44) $$-b_2 \beta L_3^{\beta-1} - \bar{a}\alpha L_3^{-\alpha-1} + \frac{1-\tau}{\rho-\mu} = 0.$$

Here $b = b_0 - b_1$, $\bar{b} = b_1 - b_2$, $a = a_2 - a_1$, and $\bar{a} = a_3 - a_2$. The last two sets of equations have the same form. They are also like eqns. (21-24), so it is immediately observed that $R_2 = R_3$, and $L_2 = L_3$. The first group is also similar to the former system, except that terms with $1-\tau$ in eqns. (17-20) are replaced by $1-2\tau$. With $n+1$ regions, we get the following general solution:

(45) $$L_1(\tau) = \frac{L}{1+n\tau}, \qquad R_1(\tau) = \frac{R}{1+n\tau}$$

(46) $$L_i(\tau) = \frac{L}{1-\tau}, \qquad R_i(\tau) = \frac{R}{1-\tau}, \qquad i = 2, \ldots, n.$$

A graphical illustration is like Figure 1, except that $R_1(\tau)$ and $L_1(\tau)$ decrease more rapidly with τ the larger n. If n is sufficiently large, these curves can be left out and we get a picture as in Figure 5.

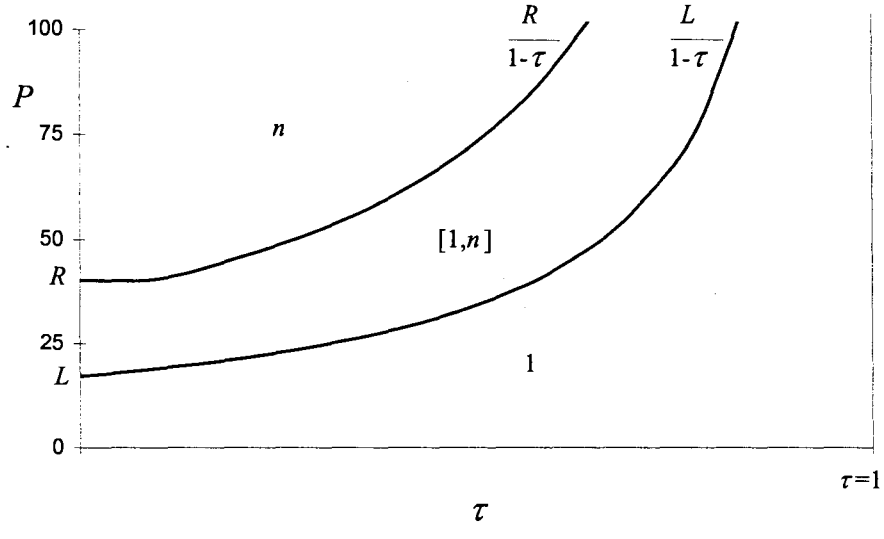

Figure 5. Entry and Exit in a Multi-Region Monopoly Model with Trade Costs.

Inside the hysteresis interval there could be any number of plants, but the most likely result is either complete spread or complete concentration; i.e., either n plants or just a single one. Other numbers will be sustained just until one of the barriers is hit. Hence, if the initial price exceeds $L/(1-\tau)$ and trade costs are decreasing continuously, then one day there will be an abrupt transition to concentration.

To some extent, the obtained results are intuitive: With equal trade costs, it is hard to see why production in all regions is not optimal if it is optimal to produce in more than one, and that concentration is optimal if the trade cost is low. It is less intuitive that the thresholds that determine whether to concentrate or spread are independent of the number of regions. Including more regions does not change the thresholds, it just shifts the option value coefficients (a_1, a_2 etc.) by equal amounts. Dixit (1988) found a similar result for the entry-exit-scrapping model.

5. Leapfrogging

The model can be used to argue how production may jump from one region to another. To see this, assume that the initial point is concentration. If demand increases enough, production may spread to other regions, but a reversal may cause concentration again. Which production centre is going to survive in such cases? This will be a matter of accident. With two regions, production will remain where it started with probability 0.5. For a larger number of regions, the probability will be accordingly smaller.

This can be pushed even further. In principle, it can be argued that the firm in many cases would be better off by shifting centre of production after a period with high demand and spread production. For example, if old factories are more likely to break down by accident than new ones, or if growth of technology is favouring new plants in other ways, the optimal decision is to get rid of old plants first whenever spread production is no longer optimal. Thus we may see that production jumps from one region to another due to temporary changes in dynamic variables.

Although the idea is simple, it seems difficult to model a setting with growth of this type correctly. The main reason is that the value functions, V_i, will depend on a number of time

parameters. This leads to partial differential equations that can only be solved numerically.[4] We would have to keep track of the age of each plant separately, and there would also be more decisions to consider. For example, under some circumstances it may be profitable to replace old plants just because they are no longer cost-effective.

Nevertheless, if the disadvantage of being old is significant, yet fairly small, we should expect solutions that are close to the ones in the previous sections except for one point: Whenever the firm shifts from spread to concentration, it will select a different region than last time.

6. Final remarks

As usual with irreversible investment, history is important also in this model. One implication of this is that monopolistic firms with similar cost and demand characteristics may develop highly different structures of production. Old firms stemming from periods with large trade costs could be based on more spread production patterns than newer ones. (Obviously, the opposite could also be the case, e.g. due to technology improvements decreasing economies of scale, but that is a different story.)

The model predicts abrupt and more or less shockwise structural changes. Is this reasonable if the number of regions is large? No, the approach is stylistic, and we should not expect as extreme shifts in reality. Further, the result depends on the simplifying assumption that trade costs are the same between every pair of regions.

It might be worth asking what the consequences could be if our firm were not a true monopolist, but a small firm in a monopolistic competitive industry. If one such firm concentrates production in one region, others would normally be better off by selecting a different one. Then the industry will spread. However, if there are strong linkages, e.g. due to a large intermediate sector, and especially if trade costs are fairly small, it seems likely that agglomeration could arise as in *Chapter 5* and *Chapter 8*, even if each firm was allowed to split production. In that case, the decision maker in this entry-exit model could represent a cluster.

---

[4] See Dixit and Pindyck (1994:205-206).

If the latter interpretation could be justified by a formal equilibrium model, it would tell a likely story about leapfrogging: In periods with low demand, the total economy may not be large enough for more than one cluster. If demand increases - temporarily, but probably over several years - there could be room for a second one. Then, if the boom stops and demand turns low again, one of the two will break down. As discussed at the firm-level, the cluster most recently established may perfectly well have a better chance of surviving.

Hopefully, the future will prove that this and the other stories in this thesis make sense!

References

Dixit, Avinash K., 1988, "Optimal Lay-up and Scrapping Decisions", Unpublished, Princeton University (July)

Dixit, Avinash K., 1989, "Entry and Exit Decisions under Uncertainty", *Journal of Political Economy* 97, 3:620-638

Dixit, Avinash K. and Robert Pindyck, 1994, *Investment Under Uncertainty*, Princeton University Press

Mossin, Jan, 1968, "An Optimal Policy for Lay-Up Decisions", *Swedish Journal of Economics* 70, 170-177