Integration, geography and the burden of history

Gianmarco I.P. Ottaviano*
Università degli Studi di Bologna
CORE, Louvain-la-Neuve, and CEPR, London

ABSTRACT: This paper develops a simple two-region two-sector general equilibrium model of trade and migration where one monopolistically competitive sector generates local pecuniary externalities. The aim is to gain insight on the question whether economic integration can be expected to increase the differences in industrial structure between more and less developed regions. It is shown that a reduction in trade and/or migration costs weakens the lock-in effect of historical events while strengthening the role of expectations.

J.E.L.: F12, F13, R1

* MAILING ADDRESS: Gianmarco I.P. Ottaviano, Università degli Studi di Bologna, Dipartimento di Scienze Economiche, Piazza Scaravilli 2, 40126 Bologna, Italy. Tel. +39-51-258666. Fax. +39-51-221968. E-mail. ottavian@economia.unibo.it.

1. Introduction

This paper addresses the general question of whether economic integration can be expected to increase the differences in industrial structure between more and less developed regions. In particular, its aim is to gain insight on how trade liberalisation and factor mobility can be expected to affect the relative importance of history and expectations in determining the international distribution of economic activities.

The point of view is the so-called «new economic geography» (Krugman (1991a)) that explains the spatial distribution of economic activities as the result of «macroeconomic complementarities» due to the interaction between competition and market size effects. On the one side, competition for local factors and consumers discourages the concentration of many firms in a single region. On the other, in the presence of increasing returns and trade costs, the concentration of consumers and firms in a single location generates pecuniary externalities that favour agglomeration.

The crucial role of the level of trade costs for the balance between competition and market size effects has been stressed at length in static models where it is shown that lower trade costs encourage agglomeration by weakening the competition effect more than the market size effect. However, the importance of trade costs for the dynamic properties of the economy has been so far left unexplored due to the high non-linearity induced by pecuniary externalities. Thus, all existing models either confine themselves to *ad hoc* dynamic arguments that are not consistent with rational expectations and forward looking optimising behaviour (e.g. Krugman (1992), Fujita, Krugman, and Venables (1998)) or give up pecuniary externalities in favour of more manageable technological spillovers (e.g. Matsuyama (1991), Krugman (1991b)). The problem with the first solution is logical coherence; the problem with the second is that the microeconomic origin of macroeconomic externalities is left unexplained.

This paper illustrates how both problems can be tackled in a parsimonious

model based on Krugman (1991a). As in Krugman's model, agglomeration arises from labour migration in the presence of pecuniary externalities. However, differently from Krugman's model, migration is costly and agents take forward-looking decisions.

The paper studies the dynamic adjustment triggered by the liberalisation of labour movements between two regions with initially different sizes. It asks under what circumstances expectations can reverse the lock-in effect of the historically inherited size advantage of the bigger region. It shows that this can happen only if the initial advantage of the leading region is not too big and if the trade and / or migration costs are low enough.

The remaining part of the paper is organized as follows. Section 2 presents the model. Section 3 solves the model. Section 4 deals with the relative importance of history and expectations. Section 5 concludes.

2. The model

Non-linearity is the barrier that discourages full-fledged dynamic analysis in location models with pecuniary externalities *à la* Krugman. However, on the one side, as argued by Fujita, Krugman, and Venables (1998), non-linearity has little bearing on their fundamental insights. On the other side, as shown by Krugman (1991b), subtle dynamic issues, such as the relative importance of history and expectations, can be satisfactorily studied in linear models. This section presents a «linear» location model *à la* Krugman that allows full-fledged dynamic analysis at the cost of two main simplifying assumptions on, respectively, preferences and factor mobility. Both assumptions will be discussed in due course.

The economic framework is the monopolistic competition model of Dixit and Stiglitz (1977). The economy consists of two regions, *A* and *B*. There is only one input, labour, whose total endowment in the economy is set to one by choice of units so that $L \in [0,1]$ workers are in *A* and $(1-L)$ workers are in *B*.

Workers are infinitely lived with rate of time preference $r \in (0,+\infty)$. However, following Matsuyama (1991), Krugman (1991b), and Galì (1995), consumption smoothing is inhibited by ruling out any form of

intertemporal trade. Consequently, at any point in time, expenditures equal income. For notational convenience, the dependence of variables upon time will be omitted when it does not generate confusion.

There are two consumption goods: a homogeneous good M and a horizontally differentiated good C which is appreciated also for its variety (S-D-S «love for variety» (Spence (1976), Dixit and Stiglitz (1977)). Each agent in location i (i=A,B) has instantaneous utility:

$$U_i = B(C_i^{\mu} M_i^{1-\mu})^{1-1/q} \qquad (1)$$

where $B \in (0,+\infty)$ is an arbitrary constant, $\mu \in [0,1]$ is the differentiated good share of expenditures, $q \in (1,+\infty)$ is the elasticity of intertemporal substitution, M_i is the consumption of the homogeneous good, C_i is the C.E.S. quantity index:

$$C_i = \left[n_i c_{ii}^{(\sigma-1)/\sigma} + n_j c_{ji}^{(\sigma-1)/\sigma} \right]^{\sigma/(\sigma-1)} \qquad (2)$$

where c_{ji} is the amount of a typical variety produced in j and consumed in i, $\sigma \in (1,+\infty)$ is the elasticity of substitution between varieties and also the elasticity of demand for each variety. Symmetry among varieties produced at the same location has been considered.

Let us introduce the first of the two simplifying assumptions that will allow closed-form dynamic analysis. It is a restriction on parameters. Due to the absence of intertemporal trade the exact value of the elasticity of intertemporal substitution is immaterial thus adding a degree of freedom in the choice of parameter values. This additional degree of freedom is exploited in order to make the model solvable. Assume therefore $q \equiv \mu/(1+\mu-\sigma)$ which yields a well-defined consumer's problem as long as $1+\mu>\sigma$. As we will see, this restriction will make instantaneous indirect utility flows linear functions of L and will thus allow the evolution of the economy to be described by a system of linear differential equations.

In addition to their consumption, workers also choose in which region to live and whether to be employed in the homogeneous or in the differentiated good sectors. It is assumed that workers are perfectly mobile between sectors in the same region. However, only workers employed in

the differentiated good sector can migrate. In different words, there is no rural-urban inter-regional mobility. This is the second main simplifying assumption that will allow for an analytical characterization of the dynamics. Apart from its technical convenience, this assumption has some tradition in the literature. In many recent developments it is customary to think of the differentiated good sector as being skill-intensive, at least when compared to the homogeneous good sector («labour dualism»). Thus, the assumption on interregional migration captures, in an extreme fashion, the fact that skilled workers are more mobile than unskilled ones (see, e.g., Smith and Zenou (1997) for a recent assessment on «dual labour markets»). This extreme assumption is not unusual in the so-called «new economic geography». For example, the same extreme assumption is adopted by Krugman (1991a). However, while Krugman assumes an exogenous distribution between the two types of workers, here the distribution is endogenously determined.

As in Krugman (1991a), the fact that a fraction of the population is immobile builds a centrifugal force into the model: in the presence of trade costs, firms have an incentive to locate close to the immobile population. However, in the present setting, it will be shown that this force is too weak to prevent agglomeration.

When moving between regions, migrants incur a cost which depends on the rate of migration, $dL/dt=\dot{L}$ (Mussa (1978)). More precisely, a migrant incurs a marginal utility loss equal to $\left|\dot{L}\right|/g$ with $g\in(0,+\infty)$. In other words, each migrant imposes a negative externality on other migrants: the larger the number of migrants, the larger the cost of migration. This assumption may be seen as capturing in a simple way the hardships that people face in reality when taking part in large migration flows.

Consider now the production side of the economy. The differentiated good is produced in a monopolistically competitive increasing-return sector. The related cost function is:

$$w_i l_i = w_i (a + bx_i) \quad a,b>0, \; i=A,B \tag{3}$$

where l_i is the amount of labour required to produce the typical variety, w_i is the wage rate and x_i is the output of the typical variety, x_{ii} of which is produced for the home market and x_{ij} for the foreign one ($x_i=x_{ii}+x_{ij}$). While it is freely traded within regions, the differentiated good can be exchanged between regions only at a cost. Trade costs are modelled as Samuelson's (1952) «iceberg costs»: in order to deliver one unit of any variety from location j to location i ($j\neq i$), t units must be shipped, with $t\in[1,+\infty)$.

The homogeneous good is produced in a perfectly competitive constant-return sector and it is freely traded. It is chosen as the numeraire and its unit input coefficient is set to one by choice of units. It is assumed that in equilibrium the homogenous good is produced in both regions. As it can be easily verified, this is the case if $m<1/2$. Due to the presence of increasing returns in the differentiated good sector, the region with a smaller share of workers eventually specialises in the production of the homogenous good. However, if the expenditures share of this latter good ($1-m$) is larger than $1/2$, the smaller region is not able to supply the whole demand. Together with the assumption of free intra-regional inter-sectoral labour mobility, this ensures factor price equalization, i.e. $w_A=w_B=1$.

3. The spatial evolution of the economy

Without intertemporal trade, a worker faces a static consumption decision and a dynamic migration decision. Considering the former, the solution of the model is standard (e.g., Krugman (1993)). The total number of firms in the economy is constant, $(n_A+n_B)=m/(\alpha s)$. The pattern of production is:

$$n=\frac{1+f}{1-f}L-\frac{f}{1-f} \quad when \quad \frac{f}{1+f}<L<\frac{1}{1+f} \tag{4a}$$

$$n=0 \quad when \quad L\leq\frac{f}{1+f}\equiv L_L \tag{4b}$$

$$n=1 \quad when \quad L\geq\frac{1}{1+f}\equiv L_H \tag{4c}$$

where $n\in[0,1]$ is the share of the total number of firms that are located in A and $f\in(0,1]$ is the ratio of total demand by domestic residents for each foreign variety to demand for each domestic variety:

$$f = \frac{x_{ji}}{x_{ii}} = \frac{c_{ji}t}{c_{ii}} = \left(\frac{p_{ii}}{p_{jj}t}\right)^{s} t = t^{1-s} \quad (5)$$

Equations (4) show that, whenever L lies inside the range $[L_L, L_H]$, agglomeration of the differentiated good sector is incomplete but the region with more workers produces a more than proportionate number of varieties. This is often called the «market size effect» (Helpman and Krugman (1985)).

The static consumption decision yields the following instantaneous flows of indirect utility for typical workers in the two regions:

$$U_A = n + f(1-n) \quad (6a)$$

$$U_A = fn + (1-n) \quad (6b)$$

where, without loss of generality, utilities have been scaled by the adequate choice of the arbitrary constant B.

By equations (4) the instantaneous flow indirect utility differential is then:

$$U_A - U_B = 2(1+f)L - (1+f) \text{ when } \frac{f}{1+f} < L < \frac{1}{1+f} \quad (7a)$$

$$U_A - U_B = -(1-f) \text{ when } L \le \frac{f}{1+f} \equiv L_L \quad (7b)$$

$$U_A - U_B = 1-f \text{ when } L \ge \frac{1}{1+f} \equiv L_H \quad (7c)$$

Since the indirect utility is higher where there are more workers, the model always exhibits agglomeration economies. Given the definition of f, those economies are stronger the lower the trade cost t and the elasticity s, which can be interpreted as an inverse index of returns of scale in equilibrium (e.g. Krugman (1991a)). This interpretation of the parameter s is worth discussing. s is both the elasticity of demand and the elasticity of substitution. It is therefore a preference parameter. However, in equilibrium it turns out to be a direct measure of the price distortion and an inverse measure of the quantity distortion due to monopoly power. For this reason, Krugman (1991a) likes to interpret s as an inverse measure of returns to scale that remain unexploited in equilibrium due to monopoly power. Even though such interpretation may seem unpalatable, it is kept in

the present paper to ease the comparison with Krugman's work.

Consider now the intertemporal migration decision. By assumption only workers employed in the monopolistically competitive sector can migrate. This implies that the economy will never move away from the incomplete specialization range $[L_L, L_H]$.

The migration decision is based on a «shadow price» defined as follows. Let $v_A(t)$ and $v_B(t)$ be the expected discounted sums of future utility minus moving costs of an agent currently in A and in B respectively. Let T be the first time the economy hits the boundaries $L=L_L$ or $L=L_H$, then by definition:

$$v_A(t) = \int_t^T U_A(s)\, e^{-r(s-t)} ds + v_A(T)\, e^{-r(T-t)} \qquad (8a)$$

$$v_B(t) = \int_t^T U_B(s)\, e^{-r(s-t)} ds + v_B(T)\, e^{-r(T-t)} \qquad (8b)$$

Moreover, since agents in each location have the option to move to the other location by paying the marginal relocation cost, $|\dot{L}|/g$,

$$v_A(t) \geq v_B(t) - \frac{|\dot{L}(t)|}{g} \quad \textit{with equality if} \quad \dot{L}(t) < 0 \qquad (9a)$$

$$v_B(t) \geq v_A(t) - \frac{|\dot{L}(t)|}{g} \quad \textit{with equality if} \quad \dot{L}(t) > 0 \qquad (9b)$$

The shadow price is then defined as:

$$v(t) \equiv v_A(t) - v_B(t) \qquad (10)$$

This shadow price represents the difference in «private» value between being in region A rather than in region B.

Equations (9) and (8) can be used to derive the economy laws of motion. They imply respectively:

$$\dot{L}(t) = g\, v(t) \qquad (11a)$$

$$\dot{v}(t) = r v(t) - [U_A(t) - U_B(t)] = r v(t) - 2(1+s)L(t) + (1+s) \qquad (11b)$$

Equations (11) have intuitive appeal. Equation (11a) states that the private marginal benefit of migration equals its private marginal cost. Equation (11b) states that the «annuity value» of being in A rather than in B, $rv(t)$, equals the «dividend», $[U_A(t)-U_B(t)]$, plus the «capital gain», $\dot{v}(t)$.

Finally, the terminal conditions can be determined following Fukao and Benabou (1993, Proposition 2). They are either (L_L,0) or (L_H,0). A zero value of $v(T)$ is required because the system hits a boundary, $L=L_L$ or $L=L_H$, in finite time T.

4. History versus expectations

Two roots correspond to system (11) and are defined by:

$$l = \frac{r \pm \sqrt{r^2 - 8(1+f)g}}{2} \qquad (12)$$

If $r^2>8(1+f)g$ there are two real positive roots, the system is unstable and it steadily diverges from (0.5,0). On the contrary, if $r^2<8(1+f)g$ the two roots are complex with positive real part, the system is still unstable, but it diverges from the centre in expanding oscillations (Figure 1). This entails that, while in the first case for any initial value of L in (L_L,L_H) there is only one optimal path and it leads to the closer endpoint, in the second case there is a subset of initial values of L, that support two optimal paths going in opposite directions. In the first case, history alone determines the long-run equilibrium. If $L<0.5$ [$L>0.5$] the economy will eventually reach (L_L,0) [(L_H,0)]. If $L=0.5$ the economy will stay there forever. The situation is different in the second case. If the economy starts anywhere between L_{LP} and L_{HP} (the «overlap» (Krugman (1991b)), both (L_L,0) and (L_H,0) are possible outcomes of self-fulfilling expectations. For any initial value of L in the overlap there are two optimal spiral paths: one leading to (L_L,0), the other to (L_H,0). Expectations decide along which spiral path the economy is going to move: the expected path will turn out to be the true path («self-fulfilling prophecy»). Outside the overlap history alone matters. So, historically inherited spatial distributions of economic activity can be changed by expectations only if L belongs to the overlap.

The width of the overlap can be found following Fukao and Benabou (1993). Starting from (L_F,0) and (L_H,0), the system is solved backwards in time. Let $\{L(t),v(t)\}$ be the trajectory that solves (11) with initial condition (L_F,0). Call t^* the first time $v(t)=0$ along the trajectory, then $L_{HP}=L(t^*)$. The same procedure with initial condition (L_H,0) can be used to determine L_{LP}.

To assess the relative importance of expectations with respect to history, what matters is the relative width of the overlap [L_{LP}, L_{HP}] with respect to the range [L_L,L_H]. The expression of the relative width of the overlap, say Λ, is:

$$\Lambda = e^{-rp\left[8(1+f)g - r^2\right]^{-1/2}} \quad (13)$$

Thus, four parameters affect the existence and the relative width of the overlap: the rate of time preference r, the inverse index of migration costs g, the transport cost parameter t and the elasticity of substitution s. More precisely, the overlap exists if the rate of time preference, migration costs, trade costs and the elasticity of substitution are low. Moreover its relative width Λ is decreasing in r, in t and in s, while it is increasing in the speed of adjustment (i.e. increasing in g). The intuition behind these results is the following. If the future is heavily discounted, workers do not care much about other workers' future decisions, so that the possibility of self-fulfilling prophecies is reduced. If the adjustment is slow, the indirect utility differential will remain close to its current level for a long time whatever the expectations, so that households' migration always follows current differentials. Finally, large transport costs t and weak returns to scale (large s) reduce the incentive towards the agglomeration of economic activities: there is weak interdependence among workers' location decisions, workers do not care much about other workers' future decisions, so that there is little ground for self-fulfilling expectations.

These results can be used to assess the impact of economic integration on the spatial evolution of the economy. Consider an initial situation in which the two regions have different size and trade as well as migration costs are prohibitive. Will integration foster agglomeration in the initially larger region? *Ceteris paribus*, it will depend on the extent of integration. If the reduction in trade and migration costs is small (i.e., such that $r^2>8(1+f)g$), then economic integration will only lock in the initial advantage of the larger region. However, if the fall in trade and migration costs is large enough (i.e., such that $r^2<8(1+f)g$) and the two regions are not too different

(i.e., such that $L \in [L_{LP}, L_{HP}]$), then there is room for self-fulfilling expectations to reverse the historical lead of the larger region.

5. Conclusion

Economic integration removes the obstacles to trade and factor mobility. In the presence of increasing returns in production, it reinforces agglomeration economies arising from agents' converging location decisions. Starting from an initial situation in which there is no trade and no labour mobility, a large enough reduction in trade and migration costs creates room for coordinated migration, led by converging expectations about the future evolution of the economy, to challenge the lock-in effect of the historically inherited spatial pattern of economic activity. The more so the stronger the returns to scale and the more patient agents are.

These results have been derived in a parsimonious general equilibrium model of location *à la* Krugman with forward-looking migration decisions. Analytical solutions have been obtained due to two main simplifying assumptions on preferences and labour mobility. While these assumptions are admittedly restrictive they show that proper and fruitful dynamic analysis is not completely out of reach as previously argued in the literature.

Aknowledgements

I am indebted with Richard Baldwin, Konrad Stahl, Jacques Thisse and two anonymous referees for helpful comments. Financial support from the TMR program of the European Commission and the University of Bologna is gratefully acknowledged.

References

Dixit, A.K. and J. Stiglitz, 1977, Monopolistic competition and optimum product diversity, American Economic Review 67, 297-308.

Fukao, K. and R. Benabou, 1993, History versus expectations: A comment, Quarterly Journal of Economics 108, 535-542.

Fujita, M., Krugman, P., and Venables, A., 1998, The Spatial Economy,

(MIT Press, Cambridge) - forthcoming.

Galì, J., 1995, Expectations-driven spatial fluctuations, Regional Science and Urban Economics, 25, 1-19.

Helpman, E., and P. Krugman, 1985, Market Structure and Foreign Trade, (MIT Press, Cambridge).

Krugman P., 1991a, Increasing returns and economic geography, Journal of Political Economy 99, 483-499.

Krugman, P., 1991b, History versus expectations, Quarterly Journal of Economics 106, 651-667 .

Krugman, P., 1992, A dynamic spatial model, NBER, Working Paper, n. 4219.

Krugman, P., 1993, The hub effect: or, threeness in international trade, in: W.J.Ethier, E.Helpman, J.P.Neary, eds., Theory, policy and dynamics in international trade (Cambridge University Press, Cambridge).

Matsuyama, K., 1991, Increasing returns, industrialization, and indeterminacy of equilibrium, Quarterly Journal of Economics 106, 617-650.

Mussa, M., 1978, Dynamic adjustment in the Heckscher-Ohlin-Samuelson model, Journal of Political Economy 86, 775-791.

Samuelson, P., 1952, Spatial price equilibrium and linear programming, American Economic Review 42, 283-303.

Smith, T.E., and Y. Zenou, Dual labor markets, urban unemployment, and multicentric cities, Journal of Economic Theory 76, 185-214.

Spence, M.A., Product selection, fixed costs, and monopolistic competition, Review of Economic Studies 43, 217-235.

APPENDIX: Equilibrium paths and the width of the overlap

This appendix derives the width of the overlap following Fukao and Benabou (1993). If T=0 is the time when the economy reaches either (L_L,0) or (L_H,0), explicit solutions can be derived from system (11) by working backwards in time. In the complex roots case (i.e. $r^2<8(1+f)g$), using (L_L,0) as initial conditions, the solutions are:

$$L(t) = 0.5-(0.5-L_L)e^{\frac{r}{2}t}\left[\cos\left(2(1+f)g-\frac{r^2}{4}\right)^{\frac{1}{2}}t-\frac{r}{2}\left(2(1+f)g-\frac{r^2}{4}\right)^{-\frac{1}{2}}\sin\left(2(1+f)g-\frac{r^2}{4}\right)^{\frac{1}{2}}t\right]$$

(A.1)

$$v(t) = (1-2L_L)(1+f)\left(2(1+f)g-\frac{r^2}{4}\right)^{-\frac{1}{2}}e^{\frac{r}{2}t}\sin\left(2(1+f)g-\frac{r^2}{4}\right)^{\frac{1}{2}}t$$

(A.2)

From (A.2) $v(t)$=0 for the first time at t=-π[2(1+ϕ)γ-(ρ^2/4)]$^{(-1/2)}$. Substituting this value for t into (A.1) gives a value for L which is the right bound of the overlap, L_{HP}:

$$L_{HP} = 0.5+(0.5-L_L)e^{-rp\left[8(1+f)g-r^2\right]^{-\frac{1}{2}}}$$

(A.3)

On the other hand, using (L_H,0) as initial conditions, the solutions are:

$$L(t) = 0.5+(L_H-0.5)e^{\frac{r}{2}t}\left[\cos\left(2(1+f)g-\frac{r^2}{4}\right)^{\frac{1}{2}}t-\frac{r}{2}\left(2(1+f)g-\frac{r^2}{4}\right)^{-\frac{1}{2}}\sin\left(2(1+f)g-\frac{r^2}{4}\right)^{\frac{1}{2}}t\right]$$

(A.4)

$$v(t) = (1-2L_H)(1+f)\left(2(1+f)g-\frac{r^2}{4}\right)^{-\frac{1}{2}}e^{\frac{r}{2}t}\sin\left(2(1+f)g-\frac{r^2}{4}\right)^{\frac{1}{2}}t$$

(A.5)

Then from (A.5) $v(t)$=0 for the first time at t=-π[2(1+f)g-(r^2/4)]$^{(-1/2)}$ as before, while by (A.4) the left bound of the overlap, L_{LP}, is:

$$L_{LP} = 0.5-(L_H-0.5)e^{-rp\left[8(1+f)g-r^2\right]^{-\frac{1}{2}}}$$

(A.6)

Finally, by (A.3), (A.6) and the definition of L_L and L_H, the width of the overlap, say P=L_{HP}-L_{LP}, is:

$$P = (L_H-L_L)\,e^{-rp\left[8(1+f)g-r^2\right]^{-\frac{1}{2}}}=\frac{1-f}{1+f}e^{-rp\left[8(1+f)g-r^2\right]^{-\frac{1}{2}}}$$

(A.7)

and the overlap is centered around L=0.5.

So, the the relative width of the overlap with respect to [L_L,L_H], labelled Λ, is simply P/(L_H-L_L) that is:

$$\Lambda = e^{-rp\left[8(1+f)g-r^2\right]^{-\frac{1}{2}}}$$

(A.8)

This is equation (13) in the paper.

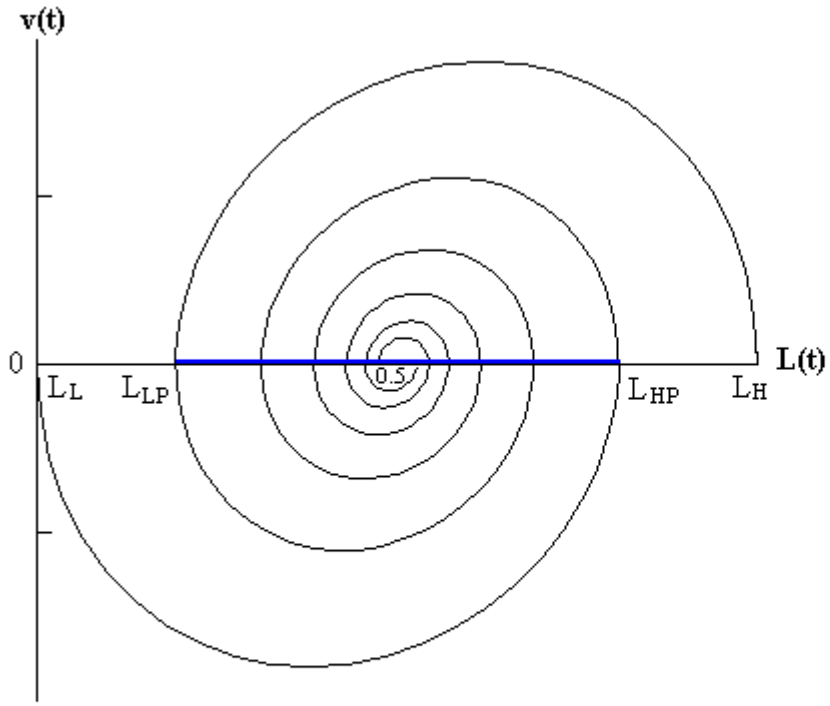

Figure 1 - The "overlap" [L_{LP},L_{HP}]