NBER WORKING PAPER SERIES

SUNK-COST HYSTERESIS

Richard Baldwin

Working Paper No. 2911

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 1989

This paper is part of NBER's research program in International Studies. Any opinions expressed are those of the author not those of the National Bureau of Economic Research.

NBER Working Paper #2911
March 1989

SUNK-COST HYSTERESIS

ABSTRACT

Despite its important theoretical, empirical and policy implications, sunk–cost hysteresis has not been characterized for the case of model consistent, or rational expectations (previous studies assume that firms believe the forcing variable is generated by some ad hoc, time invariant process such as an iid or Brownian motion process). This omission is significant since if firms do have forward–looking expectations, the existing characterizations cannot be used for empirical testing, or as a guide in developing appropriate econometric techniques. Furthermore, policy conclusions based on such characterizations may be misleading.

This paper demonstrates the possibility and characterizes the nature of sunk–cost hysteresis for a broad class of assumptions on the forcing variable process. Most notably this class includes rational or model consistent expectations. Specifically, we show that the firm's problem with a quite general forcing variable process can be reduced to be formally identical to the iid case. Additionally we analytically show that (i) the hysteresis band tends to widen with greater sunk costs, (ii) the effect of greater volatility on the band width depends upon the specific nature of the process generating the uncertainty, and (iii) greater persistence in the shocks has the effect of making well–entrenched firms more likely to exit and of narrowing the band for marginal firms. Lastly we show that the possibility of sunk–cost hysteresis is robust to a number of modifications of the basic sunk cost model.

Richard Baldwin
611 Uris Hall
Columbia University
New York, NY 10027

Sunk–Cost Hysteresis

Richard Baldwin*
Columbia University and NBER

Despite its important theoretical, empirical and policy implications, sunk–cost hysteresis has not been characterized for the case of model consistent, or rational expectations (previous studies assume that firms believe the forcing variable is generated by some ad hoc, time invariant process such as an iid or Brownian motion process). This omission is significant since if firms do have forward–looking expectations, the existing characterizations cannot be used for empirical testing, or as a guide in developing appropriate econometric techniques. This paper demonstrates the possibility and characterizes the nature of sunk–cost hysteresis for a broad class of assumptions concerning firms' beliefs about the process generating the forcing variable. Most notably this class includes rational or model consistent expectations.

Hysteresis can occur in any dynamic system which has multiple steady–state equilibria since an exogenous shock may knock the system from one steady–state equilibrium to another.[1] The notion of hysteresis in economics dates back at least to Phelps (1972) and is somewhat related to well–known effects such as irreversibilities, ratchet effects and path–dependencies.[2] However, because hysteresis is merely a property of a model, it may arise in different models for entirely different economic reasons. For instance, hysteresis can occur in models where an exogenous shock can lead to an irreversible change in the employability of the workers (as in Phelps 1972), or in union membership (as in Blanchard and Summers 1986), or in the international distribution of factor endowments (as in Kemp and Wan 1974). This paper examines the type of hysteresis first demonstrated by Baldwin (1986, 1988a). The logic of this latter type of hysteresis is simple. In the presence of sunk market entry costs, a firm's entry and exit conditions are asymmetric so a temporary shock can lead to a hysteretic change in market structure and thereby induce hysteresis in prices and quantities.

The possibility of any type of hysteresis has important theoretical, empirical and policy implications. For instance comparative static results, which formally hold only for small changes, are frequently used to analyze the effects of non–infinitesimal shocks or policy changes. Such results may be misleading if the economy is subject to hysteresis since the equilibrium conditions on which the comparative static analysis are based may be hysteretically altered by the shock itself. A recent example of a policy implication of hysteresis is found in the debate on how far the US dollar must fall to be consistent with balanced trade. One approach to this issue uses time series estimates of exchange rate pass–through elasticities, and import and export price elasticities to determine the balanced–trade level of the US dollar (e.g., Bryant, Holtham and Hooper 1988). However, given the unprecedented magnitude of the 1980s dollar cycle, it is at least theoretically possible that the relationship between the US dollar and the trade balance has hysteretically shifted. Indeed, Baldwin (1988a) presents rudimentary evidence that, due to sunk–cost hysteresis, US import prices in the 1980s were lower than expected given the real exchange rate. If hysteresis did in fact occur, the parameters estimated using pre–1980 data overstate the actual response of the trade balance, implying that the dollar must fall further than the Bryant–Holtham–Hooper calculations would suggest. It is easy to see that such miscalculations could occur in any situation involving large shocks, or large changes in policy where hysteresis is a possibility. The empirical implications of hysteresis are perhaps the most important. For instance, estimation equations derived from a model which imposes a unique long–run equilibrium may be misspecified if the model is subject to hysteresis. Furthermore, the use of fixed–coefficient lag structures to capture the effects of history may be inappropriate if hysteresis is a possibility.

The theoretical, empirical and policy implications of the type of hysteresis stemming specifically from sunk costs have been studied by several authors. For instance the possibility of sunk–cost hysteresis has been shown to affect the entry and exit decisions firms facing uncertainty (Dixit 1987a), the pass–through of exchange rates to import prices (Dixit 1987b, Baldwin 1988a, 1988b), the performance of UK exports in the 1980s (Bean 1987), the rise of anti–dumping cases in the US (Dixit 1987a), the dynamic behavior and long–run level of real exchange rates (Baldwin and Lyons 1988a, Baldwin and Krugman 1989, Krugman 1988), and the persistence of the US trade deficit in the 1980s (Krugman and Baldwin 1987). Indeed the possibility of sunk–cost hysteresis is one of the factors that led Krugman (1988) to argue for a return to fixed exchange rates. Bertola (1987) uses a related model to study investment behavior.

Yet despite its broad implications, sunk–cost hysteresis has been formally characterized for only three highly special assumptions on firms beliefs about the process governing future uncertainty. This lack of theoretical generality is a significant omission since it hinders efforts to empirical test for hysteresis as well as the development of econometric techniques to estimate models that allow for sunk costs. Moreover, it hinders the integration of the partial equilibrium sunk cost model into a general equilibrium macro model. Sunk–cost hysteresis was first characterized for assuming that firms had perfect foresight on the future values of the forcing variable (Baldwin 1986, 1988a). Two subsequent studies characterized sunk–cost hysteresis for the case where firms act as if the forcing variable is generated by a stochastic process (identically and independently distributed in Baldwin and Krugman 1989, and Brownian motion in Dixit 1987a, b). The perfect foresight result showed that sunk–cost hysteresis has nothing to do with uncertainty but it and the identically and independently distributed (iid) assumption are obviously unrealistic in many applications. Dixit (1987a, b) takes an important step toward realism by using the techniques of continuous–time dynamic programming to solve for the optimal entry–exit strategy when the uncertainty is expected to evolve according to Brownian motion. Dixit is able to show analytically that sunk–cost hysteresis is a possibility. His numerical solution also allows a characterization of how the position and width of the no–entry–no–exit band changes with various underlying parameters. Most notably he shows that the band widens with greater sunk costs and with greater volatility of the uncertainty.

Brownian motion is a natural way of modeling expectations about uncertainty in many instances. Yet there are many situations when the Brownian motion assumption is unacceptable. For instance, if the stochastic forcing variable is the endogenous outcome of more fundamental factors such as monetary and fiscal policy (as in the case of macro uncertainty), or supply and demand interactions (as in the case of price uncertainty), then the Brownian motion assumption is equivalent to imposing a rigid form of adaptive expectations. Clearly in such cases the Brownian–motion characterization of the entry–exit strategy cannot be used for empirical testing (since agents may well have had forward–looking expectations).[3] Additionally the

Brownian motion assumption is inappropriate when the entry–exit decision feeds back into the stochastic process. For instance unless the firm is atomistic, its own entry into the market will affect the process generating the uncertain return on its investment. In the context of international economics, foreign firms entering the home country market could affect the process generating the exchange rate as shown by Baldwin and Lyons (1988b).

This paper attempts to correct partially this lack of generality by using the techniques of discrete–time dynamic programming. Specifically we solve for the optimal entry–exit strategy and demonstrate the possibility of hysteresis in the presence of sunk entry costs allowing for a broad class of processes generating the forcing variable. The solution involves the use of a fairly standard mathematical technique (Bellman 1957) which reduces the problem with a general forcing variable process to be formally equivalent to the iid case. We also analytically characterize how the width and position of the hysteresis band is affected by changes in (i) the size of the sunk costs, (ii) the volatility of the uncertainty, and (iii) the degree of persistence of the forcing variable process. Lastly we show that the possibility of hysteresis is robust to modifications of several of the assumptions of the basic sunk cost model.

The intellectual antecedents of sunk–cost hysteresis are manifold. The economics are related to the putty–putty versus putty–clay investment distinction (e.g., McDonald and Siegel 1985), and the criticisms of the contestability literature (e.g., Stiglitz 1987) in the industrial organization literature. Williamson (1963) discusses the importance of selling costs as a barrier to entry. Our model follows the basic industrial organization approach of Dixit (1979, 1980) and Eaton and Lipsey (1980, 1981). In the empirical trade literature, Orcutt (1950) discusses and tests for a quantum effect of large exchange rate changes without theoretical explanation.

The paper has six sections. The first two present the basic model and solve for the optimal entry–entry strategy for a broad class of expectations. The third demonstrates the possibility of hysteresis. The fourth section characterizes the nature of hysteresis. The fifth considers modifications of the basic model. The concluding remarks are in the last section.

I. The Basic Model

Consider a firm which can sell in a market by expending (in addition to variable production costs) a fixed and sunk cost, F. To sell in subsequent periods the firm must incur a fixed maintenance cost M, such that F > M (F and M are finite and M may be zero). A binary state variable, A, keeps track of the firm's history ($A_t = 1$ if it was active last period, and $A_t = 0$ otherwise). The firm has two decisions in each period: Whether to be active in the market; and if it is active what level of sales or price (depending on the specific strategic assumptions) to choose. Since the exact nature of the second decision is tangential to our purposes, its details are submerged in a function which gives the level of its operating profit, π_t, as a function of the realization of the stochastic forcing variable, s, (for example s can be thought of as marginal cost, or a variable negatively affecting demand) and a vector of state variables, x_t:

(1.1) $$\pi_t = \pi[s_t, x_t], \text{ where } \frac{\partial \pi[s_t, x_t]}{\partial s_t} < 0.$$

The support of s is the real line and x_t is defined according to the dictates of the specific application. The function π is continuous in all arguments and bounded above and below. The firm chooses an entry–exit strategy to maximize its expected discounted cash flows. The timing of the decision is: The firm observes s_t, then decides to be active or inactive, and then earns $\pi_t - F - DA_t$ (if it is active), or zero (otherwise), where $D \equiv F - M$. Section IV considers a number of variations on these assumptions.

II. The Optimal Entry–Exit Strategy

To provide intuition for the optimal entry–exit strategy, consider the standard problem of a firm with fixed start–up costs in a static setting. In this case, the optimal decision rule (entry condition) is that the firm should be active only if profits can at least cover the fixed costs: $\pi_t \geq G$, where G is the fixed cost (the entry and exit conditions are symmetric). Allowing for multiple periods and sunk costs requires a modification of this rule. Sunk costs imply that it is cheaper for the firm to stay "in" than it is to get "in". Consequently, being active this period may provide the firm with an advantage in future periods. We refer to the discounted expected value of this advantage as the incumbency premium and denote it as Ψ_{t+1}. Thus with sunk costs the optimal decision rule becomes: $\pi_t + \delta\Psi_{t+1} \geq F + DA_t$ where δ is the constant discount rate.

A. The I.I.D. Finite Horizon Example

To fix ideas, we first find the optimal entry–exit strategy for the simple case of an iid forcing variable and a finite horizon. Although we spend a good deal of time on this example it is well worthwhile. Subsections II.B and II.C show that the solution allowing for a broad class of processes and an infinite horizon is quite similar the simple case considered here.

Consider the discrete–time dynamic system which describes the evolution of x and A:

$$(2.1) \qquad x_{t+1} = k[s_t, x_t, U_t], \quad A_{t+1} = U_t$$

where U is the control variable ($U_t = 1$ if the firm chooses to be active in period t, $U_t = 0$ otherwise). The triplet (s_t, x_t, A_t) is an element of the state space S. The forcing variable, s, is iid across periods with the cumulative distribution function (cdf) $P[s]$. The problem is to find a control policy (a sequence of functions) $\phi = \{\mu_0, \ldots, \mu_{T-1}\}$, where each μ_t $(t=0,\ldots,T-1)$ maps the state space S into the set $\{1, 0\}$, and maximizes:

$$(2.2) \qquad V_0^{\phi}[s_0, x_0, A_0] = E\{\sum_{t=0}^{T-1} \delta^t g[s_t, x_t, A_t, \mu_t[s_t, x_t, A_t]]\} + g_T, \quad \text{such that } 0 < \delta < 1 \text{ and}$$

$$x_{t+1} = k\left[s_t, x_t, \mu_t[s_t, x_t, A_t]\right],$$

$$g[s_t, x_t, A_t, U_t] = \begin{cases} \pi[s_t, x_t] - F - DA_t, & \text{if } U_t = 1 \\ 0, & \text{if } U_t = 0, \quad \text{all t,} \end{cases}$$

and $g_T = 0$. Expectations are over s_t $(t=0,\ldots,T-1)$. We characterize the optimal control policy using Bellman's optimality principle. That is, supposing that the optimal value of the firm in period t (all t) is finite and can be described by the functions $V_t[s_t, x_t, A_t]$ (existence of these functions is addressed below), then, using (2.1), we characterize the control policy by solving the much simpler sub–problem of choosing U_t to maximize:

$$(2.3) \qquad U_t(\pi[s_t, x_t] - F - DA_t) + \delta E\{V_{t+1}\left[s, k[s_t, x_t, U_t], U_t\right]\},$$

where the expectation is over s_{t+1}. (We omit the time subscript on s_{t+1} in (2.3) to indicate that the expectation involves integrating over all possible realizations). The optimizing firm calculates (2.3) for $U_t = 1$ and 0, and opts for the one which yields the highest value. Specifically, it is active only if $\pi[s_t,x_t] - F - DA_t + \delta EV_{t+1}[s,k[s_t,x_t,1],1]]$ is greater than or equal to $\delta EV_{t+1}[s,k[s_t,x_t,0],0]]$.

The optimal entry–exit strategy can be stated in terms of the realization of s_t for which the firm is indifferent to being active and inactive. This critical value, α_t, is defined implicitly by:

(2.4) $$\pi[\alpha_t,x_t] - F - DA_t + EV_{t+1}\left[s,k[\alpha_t,x_t,1],1\right] = EV_{t+1}\left[s,k[\alpha_t,x_t,0],0\right],$$

or more succinctly, $$\pi[\alpha_t,x_t] + \delta\Psi_{t+1}[\alpha_t,x_t] = F + DA_t,$$

where $$\Psi_{t+1}[\alpha_t,x_t] \equiv EV_{t+1}\left[s,k[\alpha_t,x_t,1],1\right] - EV_{t+1}\left[s,k[\alpha_t,x_t,0],0\right].$$

Now assuming that the functions π, V_{t+1} and k are such that,

(2.5) $$\partial(\pi[s_t,x_t] + \delta\Psi_{t+1}[s_t,x_t])/\partial s_t < 0,$$

(more on this condition below), the period t optimal entry–exit strategy is depicted in figure 1. The critical value for the case where $A_t = 0$ is labeled α_t^I, and the critical value for $A_t = 1$ is labeled α_t^O in figure 1. Condition (2.5), which we call the slope condition, ensures that $\pi[s_t,x_t]$ intersects the $F - \delta\Psi_{t+1}[s_t,x_t]$ and $M - \delta\Psi_{t+1}[s_t,x_t]$ lines from above. Clearly, the optimal strategy is: Be active only if $s_t \leq \alpha_t^I$ (if $A_t = 0$), or be active only if $s_t \leq \alpha_t^O$ (if $A_t = 1$). The conditions under which (2.5) holds depend upon the specifics of the functions k and π; in section II.D (2.5) is shown to hold for a broad class of problems. If the inequality in the slope condition is reversed, the inequalities in the decision rule are reversed as shown in figure 2. In Bellman's terminology, α_t partitions the state space into two decision regions.

The hysteresis Band

In each period, the firm faces only α_t^I (if $A_t = 0$) or α_t^O (if $A_t = 1$), not both. Nonetheless, considering the two critical values simultaneously builds understanding of sunk–cost hysteresis. In figure 1, α_t^I and α_t^O divide the realizations of s_t into three regions: For any $s_t \leq \alpha_t^I$ the firm is active regardless of its history; for any $s_t > \alpha_t^O$ the firm is inactive regardless of its history.

However, for any s_t between α_t^I and α_t^O, history matters; if the firm was out last period, it remains out; if it was in last period, it remains in. This is the no–entry–no–exit band referred to in the introduction. We show in section III that its existence implies the possibility of hysteresis (which is why it is sometimes called the hysteresis band). More concisely:

$$\mu_t[s_t,x_t,A_t] = \begin{cases} 1, & \text{if } s_t \leq \alpha_t \\ 0 & \text{otherwise} \end{cases} \tag{2.6}$$

where α_t is implicitly defined by (2.4).

Existence and Optimality

For (2.6) to be optimal, the functions V_t $(t=0,...,T-1)$ and their expectations in (2.4) must exist and be finite. The proof that these conditions are met and that (2.6) is indeed optimal is sketched in appendix 1. Basically, the proof proceeds by solving sub–problems like (2.3), starting with the period T–1, and working backward. In T–1, we know that $EV_T[\cdot,\cdot,\cdot]$ exists and is finite since no matter what the firm does in period T, its value is zero. The solution to (2.3) for T–1 gives us μ_{T-1} and the value function, $V_{T-1}[\cdot,\cdot,\cdot] = \text{Max}\left[\pi[s_{T-1},x_{T-1}]-F-DA_{T-1},\, 0\right]$. Given μ_{T-1} and $V_{T-1}[\cdot,\cdot,\cdot]$ we can form the expectations needed to solve (2.3) for period T–2. Repeated application of this dynamic programming algorithm yields the optimal control policy (2.6), and the value functions: $\text{Max}\left[\pi[s_t]-F-DA_t+\delta EV_{t+1}[s,k[s_t,x_t,1],1],\, \delta EV_{t+1}[s,k[s_t,x_t,0],0]\right]$, $\forall t$.

The incumbency premium $\Psi_{t+1}[s_t,x_t]$, which is the key to characterizing (2.6), is:

$$(F-M)\int_{z=-\infty}^{\alpha_{t+1}^I} dP[z] + \int_{z=\alpha_{t+1}^I}^{\alpha_{t+1}^O} \{\pi[z,k[s_t,x_t,1]] - M + \delta\Psi_{t+2}[z,k[s_t,x_t,1]]\}dP[z]. \tag{2.7}$$

In words, the incumbency premium is the expectation of the difference between $V_{t+1}[\cdot,\cdot,1]$ and $V_{t+1}[\cdot,\cdot,0]$. On one hand, if $s_{t+1} \leq \alpha_{t+1}^I$, the firm will be active whether $A_{t+1} = 1$ or 0, so it is worth $\pi_{t+1} + \delta V_{t+2}[\cdot,\cdot,1]-F-DA_{t+1}$. Clearly for such realizations of s_{t+1}, the difference, $V_{t+1}[\cdot,\cdot,1]-V_{t+1}[\cdot,\cdot,0]$, equals $F-M$. On the other hand, if $s_{t+1} > \alpha_{t+1}^O$, the firm will be inactive for $A_{t+1} = 1$ and 0, and will be worth $V_{t+1}[\cdot,\cdot,0]$. Thus the difference equals zero for such s_{t+1}. Lastly, for $\alpha_{t+1}^I \leq s_{t+1} \leq \alpha_{t+1}^O$, the difference equals $\pi_{t+1} - M + \delta\Psi_{t+2}[\cdot,\cdot]$. Integrating over all possible s_{t+1} (weighting the various outcomes by the pdf of s_{t+1}) gives us $\Psi_{t+1}[s_t,x_t]$.

B. Extension to Infinite Horizon Approximation

The question of whether the world will come to an end is well beyond the scope of this paper. However, in many situations it is unrealistic to assume that firms know exactly when the last period will come. Moreover, the infinite horizon approximation is of special interest since many macro models use this framework. Fortunately the extension to the infinite horizon case is simple. In one sense, most of the work is already done. The existence of the hysteresis band in (2.6) did not depend on the size of T, so making T larger should not affect the existence of the band. Continually making T bigger is essentially what is involved in moving to the infinite horizon approximation. The mathematical problem is to check that the value of the firm remains finite and the expectations remain well–defined as T limits to infinity. Indeed, if for every triplet (s_t,x_t,A_t) V_t converges to a finite number as T goes to infinity, then there exists a time–invariant function V which relates the value of the firm in period t to s_t, x_t and A_t. Given our assumptions, this convergence is a fairly obvious result. π is bounded above and below so g_t is always finite and δ is between 0 and 1 so it is easy to see that the discounted sum of expected g_t's would converge as T goes to infinity. Appendix 2 sketches the proof.

Given $V[s_t,x_t,A_t]$ exists and is finite and the slope condition holds, the stationary entry–exit strategy is: Be active only if:

(2.8) $$s_t \le \alpha_t = a[x_t,A_t], \quad \text{where } a[x_t,A_t] \text{ is defined by:}$$

$$\pi[\alpha_t,x_t] + \delta\Psi[\alpha_t,x_t] = F + DA_t,$$

where, $$\Psi[\alpha_t,x_t] \equiv EV\left[s,k[\alpha_t,x_t,1],1\right] - EV\left[s,k[\alpha_t,x_t,0],0\right]$$

Here expectations are over s_{t+1}. As before, if the inequality in the slope condition is reversed, then the optimal entry–exit strategy would be described by (2.8) with the inequality reversed.

C. General Expectations: *a mathematical sleight–of–hand*

Consider the infinite horizon problem of the firm as outline above with the modification that now the firm believes that the forcing variable evolves according to the very general process:

(2.9) $$s_t = h[\epsilon_t,y_t],$$

where ϵ_t is an iid disturbance and y_t is a vector of state variables governed by the laws of motion:

$$y_{t+1} = y[\epsilon_t, y_t]. \tag{2.10}$$

The function $h[\epsilon,y]$ is assumed to be continuous in all its arguments and increasing in its first.

Equation (2.9) obviously includes rational or model consistent expectations when there is a single underlying source of iid uncertainty in the model. In any particular example, the laws of motion governing the state variables will be defined by the specifics of the macro or micro model and ϵ will be some transformation of the underlying source of iid uncertainty. As the studies mentioned in the introduction showed, one implication of hysteresis is that a large shock may alter some of the parameters in a macro or micro model. It is important to note that (2.9) is sufficiently general to account for such shifts. As long as firms' expectations — including the possibility of parameter shifts — are based on state variables (that obey well–behaved laws of motion) and an underlying iid shock. This point is made explicitly in the context of exchange rate hysteresis in Baldwin and Lyons (1988b).

Now to reduce the problem to the iid case solved above, we define:

$$\bar{\Pi}[\epsilon_t, w_t] \equiv \pi\Big[h[\epsilon_t, y_t], x_t\Big], \tag{2.11}$$

where $w_t \equiv (x_t; y_t)$, and then treat ϵ <u>exactly</u> as we did s in the iid problem, using $\bar{\Pi}$ instead of π, w and its laws of motion instead of x and expanding the state space to include ϵ_t and y_t. Given the arguments made above, the optimal entry–exit strategy of the firm is given by (2.8).

The AR1 Example of (2.9)

To illuminate this technique, we briefly consider an explicit example of (2.9). Suppose the firm faces an infinite horizon, that π depends only on s and the firm expects s to evolve according to the first–order autoregressive (AR1) process:

$$s_{t+1} = \rho s_t + \epsilon_{t+1}, \quad 0 \le \rho \le 1. \tag{2.12}$$

where ϵ is iid, $E\{\epsilon\} = 0$ and the support of ϵ is the real line. The state variables are ϵ_t, s_{t-1} and A_t (specifically $y_t \equiv s_{t-1}$, and the function y is such that $y_{t+1} = \epsilon_t + \rho y_t$). By the dynamic programming algorithm, $V[\epsilon_t+\rho s_{t-1}, A_t]$ is:

$$(2.13) \qquad \text{Max}\left[\pi[\epsilon_t+\rho s_{t-1}]-F-DA_t+\delta EV[\epsilon+\rho(\epsilon_t+\rho s_{t-1}),1], \delta EV[\epsilon+\rho(\epsilon_t+\rho s_{t-1}),0]\right],$$

where the expectation are over ϵ_{t+1}. In period t, the optimal decision rule is: Be active only if $\rho s_{t-1}+\epsilon_t \leq \tilde{\alpha}_t = \bar{a}[A_t]$, where the function $\bar{a}[A_t]$ is implicitly defined by:

$$(2.14) \qquad \pi[\tilde{\alpha}_t] + \delta\Psi[\tilde{\alpha}_t] = F + DA_t, \quad \text{where}$$

$$\Psi[q] \equiv EV[\epsilon+\rho q,1] - EV[\epsilon+\rho q,0], \quad \text{or equivalently}$$

$$\Psi[q] = (F-M)P[\alpha^I-\rho q] - \int_{z=\alpha^I}^{\alpha^O} \{\pi[z]-M+\delta\Psi[z]\}dP[z-\rho q].$$

Note that the cdf of s_t in the AR1 case is $P[s_t-\rho s_{t-1}]$ where $P[\cdot]$ is the cdf of ϵ. To state the entry–exit strategy more explicitly, if the firm was inactive last period, it should come "in" only if $s_t \leq \alpha^I$; if it was active last period, it should go "out" only if $s_t > \alpha^O$ where:

$$(2.15) \qquad \begin{aligned} \pi[\alpha^I] + \delta\Psi[\alpha^I] &= F \\ \pi[\alpha^O] + \delta\Psi[\alpha^O] &= M. \end{aligned}$$

In the AR1 case the critical values are time invariant, since the current realization of s is a summary statistic for the state of the s process. An equivalent decision rule which treats ϵ exactly like s in the iid case is: Be active only if $\epsilon_t \leq \xi_t = \bar{a}[A_t] - \rho s_{t-1}$.

D. Expectation Processes for which the Slope Condition Holds

The condition that $\partial(\tilde{\Pi}[\epsilon_t, w_t] + \delta\Psi[\epsilon_t, w_t])/\partial\epsilon_t$ be negative is not easy to interpret since the Ψ function depends on the functions V, π, h, y and k. Here we examine the case where $\tilde{\Pi}$ depends only on s in order to identify a class of forcing variable processes that are consistent with the slope condition. In this case, one easily interpretable sufficient condition is that the expected value of incumbency premium, Ψ, be everywhere non–increasing in ϵ_t. In other words, if a higher ϵ this period is bad news for profits in all future periods then the expected premium from being active this period is diminished. This condition implies that the slope condition holds due to (1.1) and the assumption that $\partial h[\cdot,\cdot]/\partial\epsilon_t$ is positive. This sufficient condition is met when the shocks are expected to evolve according to an autoregressive–moving average (ARMA) process which has a moving average (MA) representation with non–negative parameters.

Proposition 1 (Ψ is non–increasing for any ARMA forcing variable process with positive persistence): If the function, h, can be represented by $s_t = \sum_{j=0}^{\infty} \theta_j \epsilon_{t-j}$, such that all θ_j are non–negative, then Ψ is non–increasing in the contemporaneous realization of ϵ.

The outline of the proof is as follows. First, any ARMA process can be represented as an MA process (perhaps of infinite order). In particular, if all the AR and MA coefficients in an ARMA process are non–negative then all coefficients in the MA representation are also be non–negative. Second, we demonstrate that Ψ_{t+1} is non–increasing in ϵ_t (all t) for the T–period finite horizon case. Lastly we argue that Ψ_t converges to a time invariant function as T approaches infinity. The first and last steps involve applications of well–know results so we omit them. Given:

$$s_t = \sum_{i=0}^{\infty} \theta_i \epsilon_{t-i}, \quad \theta_i \geq 0 \text{ all i, } \epsilon \text{ is iid,} \tag{2.16}$$

The optimal strategy is given by (2.6) taking the contemporaneous realization of ϵ as the disturbance and all lagged values of ϵ as the state vector x_t. To verify that $\partial\Psi_{t+1}[...]/\partial\epsilon_t \leq 0$ for an arbitrary period t, note that the incumbency premium, $\Psi_{t+1}[\epsilon_t, \epsilon_{t-1}, ...]$, is:

$$(F-M)P[\alpha^I_{t+1}] + \int_{z=\alpha^O_{t+1}}^{\alpha^I_{t+1}} \{\pi[\theta_0 z + \theta_1 \epsilon_t + \sum_{i=2}^{\infty} \theta_i \epsilon_{t+1-i}] - M + \delta\Psi_{t+2}[z, \epsilon_t, \epsilon_{t-1}, ...]\} dP[z], \tag{2.17}$$

where $P[\cdot]$ is the cdf of ϵ. Since α^I_{t+1} and α^O_{t+1} depend on ϵ_t, ϵ_t enters the integrand and limits of integration in (2.17). The partial derivative of the incumbency premium, $\Psi_{t+1}[...]$, with respect to ϵ_t is (defining $s^I_{t+1} = \theta_0 \alpha^I_{t+1} + \sum_{i=1}^{\infty} \theta_i \epsilon_{t+1-i}$ and $s^O_{t+1} = \theta_0 \alpha^O_{t+1} + \sum_{i=1}^{\infty} \theta_i \epsilon_{t+1-i}$):[4]

$$\left[(F-M)P'[\alpha^I_{t+1}]\frac{\partial\alpha^I_{t+1}}{\partial\epsilon_t}\right] + \left[\frac{\partial\alpha^O_{t+1}}{\partial\epsilon_t}\{\pi[s^O_{t+1}] - M + \delta\Psi_{t+2}[\alpha^O_{t+1}, ...]\}P'[\alpha^O_{t+1}]\right] \tag{2.18}$$

$$- \left[\frac{\partial\alpha^I_{t+1}}{\partial\epsilon_t}\{\pi[s^I_{t+1}] - M + \delta\Psi_{t+2}[\alpha^I_{t+1}, ...]\}P'[\alpha^I_{t+1}]\right] + \left[\int_{z=\alpha^I_{t+1}}^{\alpha^O_{t+1}} \{\theta_1 \pi'[\theta_0 z + \sum_{i=1}^{\infty} \theta_i \epsilon_{t+1-i}] + \delta\frac{\partial\Psi_{t+2}[z, ...]}{\partial\epsilon_t}\} dP[z]\right].$$

where $\pi'[\cdot]$ is the derivative of $\pi[\cdot]$ and $P'[\cdot]$ is the pdf of ϵ. By definition of the critical value α^O_{t+1} the second term in large parentheses is zero since at α^O_{t+1} the sum of profits and incumbency premium exactly equals M. By definition of α^I_{t+1}, the first and third terms in large parentheses cancel since at α^I_{t+1} the sum of profits and incumbency premium exactly equals F. (NB: These cancellations could be thought of an application of the envelope theorem to a discrete choice situation. The change in the value of firm in t+1 with respect to ϵ_t is the same with and without re–optimization on the α_{t+1}'s. Since this is true for $A_t = 1$ and 0, it is also true for Ψ_{t+1}. These cancellations are used extensively in the analysis in section IV.) Clearly, the partial derivative depends only on the last term in large parentheses. Since θ_1 is non–negative and $\pi'[\cdot]$ is negative, the term will be negative if the partial of Ψ_{t+2} with respect to ϵ_t is non–positive. The formula for this latter partial (using the envelope cancellations) is similar (2.18):

$$\frac{\partial \Psi_{t+2}[z,\ldots]}{\partial \epsilon_t} = \left[\int_{z=\alpha^I_{t+2}}^{\alpha^O_{t+2}} \{\theta_2 \pi'[\theta_0 z + \sum_{i=1}^{\infty} \theta_i \epsilon_{t+2-i}] + \delta \frac{\partial \Psi_{t+3}[z,\ldots]}{\partial \epsilon_t} \} dP[z] \right]. \tag{2.19}$$

Since θ_2 is non–negative and $\pi'[\cdot]$ is negative, if $\partial\Psi_{t+3}[\ldots]/\partial\epsilon_t$ is negative then (2.19) is negative. Repeating the argument we would eventually reach period T–1 in which the incumbency premium is zero regardless of the realization of ϵ_t (so $\partial\Psi_T[\ldots]/\partial\epsilon_t = 0$). Thus $\partial\Psi_{t+1}[\ldots]/\partial\epsilon_t$ is non–positive since $\pi'[\cdot]$ is everywhere negative and all θ_i's are non–negative. If $\partial\Psi_{t+1}[\ldots]/\partial\epsilon_t$ is non–positive for the T–period finite horizon, and $\Psi_{t+1}[\ldots]$ converges to Ψ then it must be that $\partial\Psi[\ldots]/\partial\epsilon_t$ is also non–positive so the sufficient condition for the slope condition to hold is met.

III. Existence of Sunk Cost Hysteresis

To demonstrate the possibility of hysteresis we first demonstrate the existence of a no–entry–no–exit band:

Proposition 2 (Existence of the hysteresis band): Assuming condition (2.5) holds, the optimal entry–exit strategy described by (2.6) is such that $\alpha^I_t < \alpha^O_t$.

The proof is simple. If they are equal then by (2.4) F = M. Since this is a contradiction of an assumption, it must be that they are not equal. Also supposing that $\alpha^I_t > \alpha^O_t$ then by the slope condition, α^I_t is not the optimal decision rule for $A_t = 0$. Since we showed that α^I_t was indeed the optimal decision rule, this supposition must be false.

Next we show that the existence of the no–entry–no–exit band implies that hysteresis is a possibility. Note that we demonstrate that hysteresis is a possibility — not an inevitability.

Proposition 3 (Possibility of hysteresis): Given that the firm's optimal entry–exit strategy is marked by a no–entry–no–exit band and that the firm is not atomistic in its own market, a sufficiently large, temporary shock could hysteresis in quantities and prices.

The proof is by construction of a simple example. More complicated examples may be interesting but are not necessary to demonstrate the possibility of hysteresis. Suppose that π depends only on the current s and the firm was inactive in period t–1 so $A_t = 0$. Consider the following historical sequence of realizations of ϵ. In period t the value of ϵ was ϵ_t^o, where ϵ_t^o lies in the no–entry–no–exit band. In period t+1, the value of the ϵ decreased to ϵ''_{t+1}, where ϵ''_{t+1} is less than α_{t+1}^I. Lastly for the next N–1 periods, ϵ_{t+i} (i=2,...N) was within the no–entry–no–exit band. For comparison consider another historical sequence of ϵ_t (t=1,...,N). This second sequence is identical to the first except that ϵ_{t+1} lies within the band rather than below it. The probabilities of such sequences could be calculated using the pdf of ϵ, however it is sufficient to note that they are non–zero. In non–hysteretic models, the effects of ϵ''_{t+1} would eventually fade out in the sense that price and quantity in period N would be the same for both sequences (for sufficiently large N). However, due to the asymmetric entry and exit conditions created by the sunk costs, in the first sequence the firm switches from being inactive in t to being active in t+1 in the first sequence, and more importantly, remains active during the next N–1 periods. Comparing this to the second sequence (where the firm does not enter), we see that hysteresis in market structure has occurred. As long as the firm is not atomistic in its own market, its presence will affect the quantity and/or price in its own market. As a result, past shocks affect real variables in a manner that cannot be captured with the simple fixed lag structures commonly employed in the empirical work. In a time series estimation that did not include the number of firms as an explanatory variable, hysteresis would appear as a structural break. Obviously since home imports are foreign exports, this demonstrates the possibility of hysteresis in exports as well.

Note that hysteresis differs from irreversibilities since the system could be restored to its period t state by a corrective shock.

IV. Characterizing the Nature of Sunk Cost Hysteresis

This section characterizes how the width and position of the band is affected by changes in underlying parameters such as the size of the sunk costs, the degree of persistence in the s process, and the degree of volatility in the s process. We restrict our attention to the infinite horizon case, where π_t depends only on the current s_t, and s is expected to evolve according to the AR1 process described in (2.12). The AR1 process reflects a large class of expectations such as the random walk and iid. Generalizing to other specific ARMA processes would involve similar analytic techniques. These comparative static exercise involve totally differentiating (2.15) with respect to α^I, α^O and either F, ρ or an index of volatility, r. Clearly Ψ depends on these parameters, and to be entirely correct, we should have denoted it as $\Psi[s_t,x_t;\Delta]$ where Δ is the vector of parameters. However for notational simplicity we explicitly include only F, ρ or r as a variable in Ψ as needed.

A. The Band and the Size of F

In this sub–section, we prove that under certain conditions an increase in the size of the sunk costs, F, widens the band by increasing α^O and decreasing α^I. Heuristically, figure 1 shows that α^O increases with F only if the incumbency premium (evaluated at α^O) increases with F. To see why Ψ increases with F, recall that Ψ is positive because it is cheaper to stay in than to get in (i.e., F > M). As this difference increases the expected value of the incumbency advantage increases. The effect of a higher F on α^I is slightly more complicated because a higher F increases the cost as well as the benefit of becoming active. More formally, treating F as a variable, the definition of $\Psi[s_t,F]$ evaluated at $s_t = q$ is:

$$\text{(4.1)} \qquad (F-M)P[\alpha^I - \rho q] + \int_{z=\alpha^I}^{\alpha^O} \{\pi[z] - M + \delta\Psi[z,F]\}dP[z-\rho q]$$

where $P[\cdot]$ is the cdf of ϵ in (2.12). Totally differentiating (2.15), using (4.1), we get:

$$\text{(4.2)} \qquad \frac{d\alpha^O}{dF} = \frac{\delta\Psi_F[\alpha^O, F]}{-\pi'[\alpha^O] - \delta\Psi_1[\alpha^O, F]}, \text{ and } \frac{d\alpha^I}{dF} = \frac{\delta\Psi_F[\alpha^I, F] - 1}{-\pi'[\alpha^I] - \delta\Psi_1[\alpha^I, F]},$$

where $\Psi_1[\cdot,F]$ is the partial of Ψ with respect to its first argument. By assumption $\pi'[\cdot]$ is negative and differentiating (4.1), we have that $\Psi_1[q,F]$ equals:

$$(4.3) \qquad -\rho(F-M)P'[\alpha^I-\rho q] - \rho\int_{z=\alpha^I}^{\alpha^O}\{\pi[z]-M+\delta\Psi[z,F]\}P''[z-\rho q]dz.$$

To show that this is non–positive, we integrate the second term by parts:

$$(4.4) \qquad \Psi_1[q,F] = (-\rho)(F-M)P'[\alpha^I-\rho q]$$

$$-\rho\{\pi[\alpha^O]-M+\delta\Psi[\alpha^O,F]\}P'[\alpha^O-\rho q] + \rho\{\pi[\alpha^I]-M+\delta\Psi[\alpha^I,F]\}P'[\alpha^I-\rho q]$$

$$+\rho\int_{z=\alpha^I}^{\alpha^O}\{\pi'[z]+\delta\Psi_1[z,F]\}dP[z-\rho q].$$

By the envelope cancellations the first three terms cancel out, leaving only the integral term which involves the partial that we are trying to sign. Repeatedly substituting $\Psi_1[\cdot,\rho]$ into itself shows that the partial is non–positive since it depends only on terms involving $\rho\pi'[\cdot]dP[\cdot]$. Thus from (4.3) we have that the sign of $d\alpha^O/dF$ depends on the sign of Ψ_F and the sign of $d\alpha^I/dF$ depend on the sign of Ψ_F-1. Differentiating with respect to F and using the envelope cancellations described above:

$$(4.5) \qquad \Psi_F[q,F] = P[\alpha^I-\rho q] + \int_{\alpha^I}^{\alpha^O}\{\delta\Psi_F[z,F]\}dP[z-\rho q].$$

To build intuition, we first sign (4.5) for the iid case ($\rho = 0$). With $\rho = 0$:

$$(4.6) \qquad \Psi_F[q,F] = P[\alpha^I] + \left[\int_{\alpha^I}^{\alpha^O}\delta\Psi_F[z,F]dP[z]\right].$$

Using the fact that $\Psi_F[z,F]$ is independent of the realization of z_{t+1} when the process is iid, $\Psi_F[\cdot,F]$ equals $P[\alpha^I]+\delta\Psi_F[q,F](P[\alpha^O]-P[\alpha^I])$, so:

$$(4.7) \qquad \Psi_F[q,F] = \frac{P[\alpha^I]}{1-\delta(P[\alpha^O]-P[\alpha^I])}, \quad \text{for all q.}$$

Standard properties of cdf's and the fact that $\alpha^I < \alpha^O$ can be used to show that $\Psi_F[\cdot,F]$ lies between zero and unity, as long as the discount factor, δ, is not too small. By (4.3) then, it is clear that $d\alpha^I/dF$ is negative and $d\alpha^O/dF$ is positive. Consequently a higher F leads to a wider band.

General Rho

The demonstration is more intricate for $\rho \neq 0$. In appendix 3, we show that Ψ_F is still positive in this case so $d\alpha^O/dF$ is positive. By a series of substitutions (detailed in appendix 3):

(4.8) $$\Psi_F[\alpha^I,F] \leq C \equiv P[\alpha^I - \rho\alpha^I]\,(1 + \frac{\delta\beta}{1-\delta\varphi}), \text{ where}$$

$$\varphi = (P[\alpha^O - \rho\alpha^I] - P[\alpha^I - \rho\alpha^O]),\ \beta = (P[\alpha^O - \rho\alpha^I] - P[\alpha^I - \rho\alpha^I]).$$

Given this, it is possible to show that as long as difference between F and M is not too great, then C (and consequently $\Psi_F[\alpha^I,F]$) is less than unity. In particular if we evaluate $d\alpha^I/dF$ at the point where F=M, then we can use the fact that at that this value of F, $\alpha^O = \alpha^I$. Thus at this value of F, β and φ are zero, so C equals $P[\alpha^I(1-\rho)]$ which is obviously strictly less than unity so $d\alpha^I/dF$ is negative. Consequently, we know that at least for F sufficiently close to M, greater sunk costs lead to a wider band.

B. The Band and The Degree of Persistence in the s Process

Next we address the link between ρ and the band. There are several economic interpretations of changing rho. First as rho decreases from unity the process tends to revert to its mean faster, reducing the degree of persistence. One instance where this has a concrete economic meaning is the case of the standard linear stochastic version of the sticky price monetary model of exchange rate determination. In this model it is well know that a lower rho indicates faster adjustment of the domestic price level to changes in aggregate demand. Changes in this speed of adjustment can come from faster pass–through of exchange rate changes to traded goods prices. This latter source of changes is emphasized in Baldwin and Lyons (1988a).

Treating ρ as a variable rather than a parameter, α^I and α^O are defined by: $\pi[\alpha^I] + \delta\Psi[\alpha^I,\rho] = F$ and $\pi[\alpha^O] + \delta\Psi[\alpha^O,\rho] = M$. Totally differentiating these:

(4.9) $$\frac{d\alpha^O}{d\rho} = \frac{\delta\Psi_\rho[\alpha^O,\rho]}{-\pi'[\alpha^O]-\delta\Psi_1[\alpha^O,\rho]}, \text{ and } \frac{d\alpha^I}{d\rho} = \frac{\delta\Psi_\rho[\alpha^I,\rho]}{-\pi'[\alpha^I]-\delta\Psi_1[\alpha^I,\rho]}.$$

Since $\pi'[\cdot]$ is negative and $\delta\Psi_1[\cdot,\rho]$ can be shown to be negative (using a (4.4) type of argument), the derivatives have the same sign as Ψ_ρ. The definition of $\Psi[q,\rho]$ is:

$$(4.10) \qquad (F-M)P[\alpha^I - \rho q] + \int_{z=\alpha^I}^{\alpha^O} \{\pi[z] - M + \delta\Psi[z,\rho]\} dP[z-\rho q],$$

so $\Psi_\rho[q,\rho]$ equals:

$$(4.11)(F-M)P'[\alpha^I-\rho q](\frac{d\alpha^I}{d\rho} - q) + \int_{\alpha^I}^{\alpha^O} (-q)\{\pi[z]-M+\delta\Psi[z,\rho]\}\}P''[z-\rho q]dz + \int_{\alpha^I}^{\alpha^O} \delta\Psi_\rho[z,\rho]dP[z-\rho q]$$

$$+ \left[(\pi[\alpha^O]-M+\delta\Psi[\alpha^O,\rho])P'[\alpha^O-\rho q]\frac{d\alpha^O}{d\rho}\right] + \left[(\pi[\alpha^I]-M+\delta\Psi[\alpha^I,\rho])P'[\alpha^I-\rho q]\frac{d\alpha^I}{d\rho}\right],$$

where $P''[\cdot]$ is derivative of the pdf of ϵ. With the envelope theorem cancellations $\Psi_\rho[q,\rho]$ becomes:

$$(4.12) \left[(F-M)P'[\alpha^I-\rho q](-q) + \int_{\alpha^I}^{\alpha^O} (-q)\{\pi[z]-M+\delta\Psi[z,\rho]\}P''[z-\rho q]dz\right] + \int_{z=\alpha^I}^{\alpha^O} \delta\Psi_\rho[z,\rho]dP[z-\rho q].$$

The sign of this depends on the terms in large parenthesis and on the third term which involves $\Psi_\rho[q,\rho]$ itself. Substituting this definition of $\Psi_\rho[q,\rho]$ into itself, we get:

$$(4.13) \qquad \Psi_\rho[q,\rho] = \left[(F-M)P'[\alpha^I-\rho q](-q) + \int_{\alpha^I}^{\alpha^O} (-q)\{\pi[z]-M+\delta\Psi[z,\rho]\}P''[z-\rho q]dz\right]$$

$$+\delta\left[\int_{z=\alpha^I}^{\alpha^O} \left[(-z)(F-M)P'[\alpha^I-\rho z] + \int_{\gamma=\alpha^I}^{\alpha^O} (-z)\{\pi[\gamma]-M+\delta\Psi[\gamma,\rho]\}P''[\gamma-\rho z]d\gamma\right] dP[z-\rho q]\right]$$

$$+\delta^2 \int_{z=\alpha^I}^{\alpha^O} \int_{\gamma=\alpha^I}^{\alpha^O} \{\Psi_\rho[\gamma,\rho]\}dP[\gamma-\rho z]dP[z-\rho q].$$

The sign of (4.13) depends on the first two terms in large parentheses and on the sign of Ψ_ρ itself. The second term in large parentheses, however, is similar to the term in large parentheses in (4.12). Repeated substitution produces an infinite, discounted sum of terms similar to the term in large parentheses in (4.12). Thus, if we can show that the term in large parentheses in (4.12) is always of one sign then we can sign (4.12).

Intuition for the sign of (4.12) is provided in figure 3. Essentially increasing ρ has the effect of shifting the pdf of s_{t+1} to the right if q is positive (e.g., s''_t), and to the left if q is negative (e.g., s_t'). Since the difference $V[s_{t+1},1] - V[s_{t+1},0]$ is non–increasing, a rightward shift places more weight on the lower values of Ψ, thereby decreasing the conditional expectation. A leftward shift place more weight on higher values, thereby increasing Ψ. This intuition is only suggestive since ρ affects the value of $V[s_{t+1},1] - V[s_{t+1},0]$ in the range $\alpha^I - \alpha^O$. More formally, we sign the term in large parentheses in (4.12) by integrating by parts (using the fact that F–M equals $\pi[\alpha^I]-M+\delta\Psi[\alpha^I,\rho]$) to get:

$$\begin{aligned}(4.14)\qquad \Psi_\rho[q,\rho] &= (-q)(\pi[\alpha^I]-M+\delta\Psi[\alpha^I,\rho])P'[\alpha^I-\rho q] \\ &+ (q)\{\pi[\alpha^I]-M+\delta\Psi[\alpha^I,\rho]\}P'[\alpha^I-\rho q] + (-q)\{\pi[\alpha^O]-M+\delta\Psi[\alpha^O,\rho]\}P'[\alpha^O-\rho q] \\ &+ (q)\int_{\alpha^I}^{\alpha^O} \{\pi'[z]+\Psi_1[z,\rho]\}dP[z-\rho q]\end{aligned}$$

The first three terms cancel out by the envelope cancellations, so the sign of (4.14) depends only on $\pi'[\cdot]$ and Ψ_1 (which are negative) and q (which may have either sign). Clearly then (4.12) has the opposite sign of q.

Interpretation

In the case where both α^O and α^I are negative, $\Psi[\alpha^O,\rho]$ and $\Psi[\alpha^I,\rho]$ increase with ρ so, by inspection of (4.9), the band shifts to the right with increases in the degree of persistence. On the other hand, if α^O and α^I are both positive, $\Psi[\alpha^O,\rho]$ and $\Psi[\alpha^I,\rho]$ decrease with ρ so the band shifts to the left as the degree of persistence increases. If α^I is negative and α^O is positive the band narrows. Loosely speaking, if the demand and cost parameters are such that both α^I and α^O are negative then the firm will not be active at the steady–state level of s, since by definition the unconditional expectation of s is zero (recall $E\{\epsilon\} = 0$). We might refer to such firms as unlikely entrants. If both α^I and α^O are positive, the firm might be thought of as well–entrenched. If α^O is positive and α^I is negative, the firm might be thought of as a marginal firm. Roughly speaking then, we say that an increase in persistence in the exchange rate expectations, narrows the band for marginal firms. For well–entrenched firms, more persistence makes the firms less well entrenched (since the band moves toward zero). For unlikely entrants, more persistence moves the band to the right toward zero.

C. The Band and Volatility

Dixit (1987a) shows that an increase in the variance of the Brownian motion process leads to a widening of the band. This subsection demonstrates that the band widening effect of volatility does not hold in general but does hold for processes that are sufficiently close to a random walk. To demonstrate that band–widening is not a general result, consider a simple counter–example. Let $\rho = 0$ and $\pi[\cdot]$ be linear in s. With iid expectations, the incumbency premium is independent of the current realization s. In particular, the value of the incumbency premium is the same at $s_t = \alpha^I$ and $s_t = \alpha^O$. An increase in volatility changes the value of the incumbency premium, and this would in turn change the position of the band. However the width would not change since $\pi[\cdot]$ is linear. The example is illustrated in figure 4, assuming that the volatility reduces the premium from Ψ_{t+1} to $\tilde{\Psi}_{t+1}$, so that $\hat{\alpha}^I$ and $\hat{\alpha}^O$ shift to $\tilde{\alpha}^I$ and $\tilde{\alpha}^O$.

Random Walk Case

Before turn to the general link between volatility and band width, we show that under assumptions that might be thought of as the discrete time version of those in Dixit (1987a) greater volatility does widen the band. In particular, we assume that $\rho = 1$ and ϵ has a symmetric, single–peaked, mean–zero distribution, and $\pi[\cdot]$ is linear over an interval that includes α^I and α^O. Using the Diamond–Rothschild–Stiglitz notation for mean–preserving spreads (MPS), α^I and α^O are defined by: $\pi[\alpha^I]+\delta\Psi[\alpha^I,r]=F$ and $\pi[\alpha^O]+\delta\Psi[\alpha^O,r]=M$, where r indexes MPS's of the pdf $P[\epsilon,r]$ (higher r indicates more volatility). Totally differentiating the relationships that define the band:

$$(4.15)\qquad \frac{d\alpha^I}{dr} = \frac{\delta\Psi_r[\alpha^I, r]}{-\pi[\alpha^I]-\delta\Psi_1[\alpha^I,r]}, \quad \text{and} \quad \frac{d\alpha^O}{dr} = \frac{\delta\Psi_r[\alpha^O, r]}{-\pi[\alpha^O]-\delta\Psi_1[\alpha^O,r]}.$$

A sufficient condition for the band–widening is that α^I decreases while α^O increases in response to a mean–preserving spread (MPS) of ϵ. By (4.13), this occurs only if $\Psi_r[\alpha^I,r]$ is negative and $\Psi_r[\alpha^O,r]$ is positive. Intuition for why this holds in the special case of a random walk is provided in Figure 5. The heavy line shows the difference $V[s_{t+1},1] - V[s_{t+1},0]$ and the solid light line plots the pdf of s_{t+1} when $s_t = \alpha^I$. The random walk assumption together with the assumption that ϵ is has a symmetric mean–zero distribution implies that the distribution of s_{t+1} is centered exactly on α^I. A MPS of ϵ leads to a MPS of s_{t+1} thus moving weight out toward the tails as shown by the dashed line. The conditional expectation of the difference, $V[s_{t+1},1] - V[s_{t+1},0]$, obviously falls; to the left of α^I, the MPS has no effect since $V[s_{t+1},1]-V[s_{t+1},0]$ is constant. To the right of α^I, however, the MPS puts more probability weight on lower values of $V[s_{t+1},1]-V[s_{t+1},0]$. Using an analogous argument we can see that $\Psi[\alpha^O,r]$ increases with r since for s_t equal to α^O, the distribution of s_{t+1} is centered on α^O. Again figure 5 is only suggestive since it excludes the fact that the MPS alters $V[s_{t+1},1]-V[s_{t+1},0]$ over the α^I–α^O interval as shown by the dotted line. While this shift tends to mitigate the above discussed effects, appendix 4 shows that a MPS does indeed lead to a band widening. Appendix 4 also argues that if ρ is close enough to unity then we would still get the band widening effect.

I.I.D. Case

Returning to a more general $\pi[\cdot]$, if rho is zero, then $\Psi_r[\alpha^I,r] = \Psi_r[\alpha^O,r]$ so the band will widen if $\Psi_r[\alpha^I,r]$ is positive and will narrow if it is negative. This follows from (4.15), the convexity of profit functions and the fact that $\alpha^I < \alpha^O$. Also using a diagram similar to figure 5, it is easy to see that $\Psi_r[\alpha^I,r]$ is positive when both α^I and α^O are negative, and is negative when both are positive. Thus roughly speaking, in the iid case, the band widens for firms that are unlikely entrants and narrows for those that are solidly entrenched firms.

V. Model Modifications

This section shows that the existence of a no–entry–no–exit band is quite robust to the specific assumption made in section I. To be concrete, we focus on the finite horizon case.

A. Ex ante Timing of the Entry–Exit Decision

Consider the following timing: At the beginning of this period the firm decides to be active or inactive this period, based on the previous disturbance ϵ_{t-1}. Subsequently ϵ_t is observed and the cash flow $\pi_t - F - DA_t$ (if the firm is active) or zero (if it is inactive) is realized. To be specific we assume that π_t depends only on s, and that s is AR1 as in (2.12). The state variables in period

T−1 are A_{T-1} and s_{T-2}; the value function in T−1 is:

$$V_{T-1}[s_{T-2},A_{T-1}] = \text{Max}[\{\int_{-\infty}^{\infty}\pi[z+\rho s_{T-2}]dP[z]\} - F - DA_{T-1}, 0] \tag{5.1}$$

The optimal decision rule is:

$$\begin{aligned}\mu_{T-1}[s_{T-2},A_{T-1}] &= 1, \text{ if } s_{T-2} \leq \alpha_{T-2} = a_{T-1}[A_{T-1}] \\ &= 0, \text{ otherwise}\end{aligned} \tag{5.2}$$

where the critical value function, $a_{T-1}[A_{T-1}]$, is implicitly defined by:

$$\{\int_{z=-\infty}^{\infty}\pi[z+\rho\alpha_{T-2}]dP[z]\} = F + DA_{T-1} \tag{5.3}$$

Obviously since F > M and $\pi'[\cdot]$ is negative, $a_{T-1}[0] < a_{T-1}[1]$. We will have use for the expectation of V_{T-1} conditioned on information available at time T−2 so we write this out:

$$E\{V_{T-1}[s,A_{T-2}] \mid s_{T-3}\} = \int_{z=-\infty}^{\alpha_{T-2}} \{\int_{w=-\infty}^{\infty}\pi[w+\rho(z+\rho s_{T-3})]dP[w]\}\, dP[z] - F - DA_{T-1}. \tag{5.4}$$

where $\alpha_{T-2} = a_{T-1}[A_{T-1}]$. In T−2 the value function, $V_{T-2}[s_{T-3},A_{T-2}]$, is defined by:

$$\text{Max}\left[\{\int_{-\infty}^{\infty}\pi[z+\rho s_{T-3}]dP[z]\} - F - DA_{T-2} + \delta E\{V_{T-1}[s,1] \mid s_{T-3}\},\ \delta E\{V_{T-1}[s,0] \mid s_{T-3}\}\right] \tag{5.5}$$

and decision rule is:

$$\begin{aligned}\mu_{T-2}[s_{T-3},A_{T-2}] &= 1, \text{ if } s_{T-3} \leq \alpha_{T-3} = a_{T-2}[A_{T-2}] \\ &= 0, \text{ otherwise}\end{aligned}$$

where
$$\{\int_{-\infty}^{\infty}\pi[z+\rho\alpha_{T-3}]\}dP[z] + \delta\Psi_{T-1}[\alpha_{T-3}] = F + DA_{T-2},$$

and
$$\Psi_{T-1}[\alpha_{T-3}] \equiv E\{V_{T-1}[s,1] \mid \alpha_{T-3}\} - E\{V_{T-1}[s,0] \mid \alpha_{T-3}\}.$$

Note that, Ψ_{T-1} is decreasing in s_{T-3}, since an increase in s_{T-3} has no effect on $V_{T-1}[s_{T-2},1] - V_{T-1}[s_{T-2},0]$ for those s_{T-2} less than $a_{T-1}[0]$, and for s_{T-2} between $a_{T-1}[0]$ and $a_{T-1}[1]$, an increase in s_{T-3} reduces the difference (since $\pi'[\cdot]$ is negative and ρ is positive). Consequently it must be that $a_{T-2}[0] < a_{T-2}[1]$. Iteratively applying this type of reasoning it is possible to show that the optimal entry–exit strategy is marked by a no–entry–no–exit band for all $t=(0,\ldots,T-1)$.

B. An International Application: Maximization in Foreign Currency

In the single country interpretation of the section I model, the issue of the numeraire of maximization does not arise. However, much of the sunk–cost hysteresis was developed in the context of international trade so the issue of numeraire is important. First it is obvious that if F and M are fixed in foreign numeraire terms (using each country's currency as its numeraire) and the foreign firm maximized its value in terms of that same currency, none of our results would be altered. However, if F and M are fixed in home currency and yet the firm maximizes its value in terms of foreign country currency, the solution is somewhat more involved but is still marked by a hysteresis band. To show this as simply as possible, we assume that π_t (now defined in foreign currency terms) depends only on the level of the iid exchange rate e_t (home price of foreign currency) and F and M are fixed in home currency. In the penultimate period the value function and decision rule are:

$$(5.6) \qquad V_{T-1}[e_{T-1},A_{T-1}] = \mathrm{Max}\left[\pi[e_{T-1}] - (F+DA_{T-1})/e_{T-1},\, 0\right], \quad \text{and}$$

$$\mu_{T-1}[e_{T-1},A_{T-1}] = 1, \quad \text{if } e_{T-1} \leq \alpha_{T-1} = a_{T-1}[A_{T-1}]$$
$$= 0, \quad \text{otherwise}$$

where
$$\pi[\alpha_{T-1}] = (F + DA_{T-1})/\alpha_{T-1}.$$

This assumes that $e_t\pi[e_t]$ is decreasing in e_t which is true if and only if, operating profits measured in home currency are decreasing in the level of the real exchange rate. If instead $e_t\pi[e_t]$ is increasing is e_t then the inequality in (5.6) would be reversed. For an arbitrary period:

$$(5.7) \qquad V_t[e_t,A_t] = \mathrm{Max}\left[\pi[e_t] - (F+DA_t)/e_t + \delta EV_{t+1}[e,1],\, \delta EV_{t+1}[e,0]\right],$$

and the borderline realization of e_t is given by:

$$\pi[\alpha_t] + \delta\Psi_{t+1} = (F + DA_t)/\alpha_t, \quad \text{where}$$

$$\Psi_{t+1} \equiv EV_{t+1}[e,1] - EV_{t+1}[e,0].$$

Here expectations are over e_{t+1}. The decision rule is depicted in Figure 6. In the top part of the figure, we have drawn $\pi[e_t]$ such that it falls more steeply than $F/e_t - \delta\Psi_{t+1}$ and $M/e_t - \delta\Psi_{t+1}$. In this case, $a_t[0] < a_t[1]$ so the decision rule to to be active only if: $e_t \le a_t[A_t]$. As usual there is a hysteresis band. In the bottom part, we have drawn $\pi[e_t]$ has falling less steeply than $F/e_t - \delta\Psi_{t+1}$ and $M/e_t - \delta\Psi_{t+1}$. In this case $a_t[1] < a_t[0]$, however there still is a hysteresis band, since the decision rule is be active only if: $e_t \ge a_t[A_t]$.

C. Demand Side Hysteresis

For many types of goods, consumers find that the cheapest way of judging the quality of a product is to actually try it. In one of the seminal formalizations this idea (Schmalensee 1982), the producer faces a linear demand curve that shifts out by a proportional factor after consumers try the good. Thus profits for a firm that was "in" last period are greater than for a firm that was "out" last period. Plainly this is a situation that is formally quite close to sunk entry–cost problem we have solved. Consider the section I timing with iid shocks. Suppose that consumers have a short memory. If they consumed the product last period the firm would earn operating profits equal to $\pi[s_t](1+\lambda)$ by selling the product this period, where $\lambda > 0$, If they did not consume it last period, the firm's operating profits would be just $\pi[s_t]$. Now in addition to manufacturing the product, the firm must incur some fixed costs, G, (it need not be sunk) in every period it wishes to sell. In the next–to–last period:

(5.8) $$V_{T-1}[s_{T-1},A_{T-1}] = \text{Max}[\pi[s_{T-1}](1 + \lambda A_{T-1}) - G, 0],$$

and be active only if $$s_{T-1} \le \alpha_{T-1} = a_t[A_t], \quad \text{where}$$
$$\pi[\alpha_{T-1}](1 + \lambda A_{T-1}) = G.$$

For an arbitrary period the value function and decision rule are:

(5.9) $$V_t[s_t,A_t] = \text{Max}\left[\pi[s_t](1 + \lambda A_t) - G + \delta EV_{t+1}[s,1],\ \delta EV_{t+1}[s,0]\right],$$

and be active only if: $$s_t \le \alpha_t = a_t[A_t], \quad \text{where}$$
$$\pi[\alpha_t](1 + \lambda A_t) + \Psi_{t+1} = G, \text{ and}$$
$$\Psi_{t+1} \equiv EV_{t+1}[s,1] - EV_{t+1}[s,0].$$

Given the iid assumption the slope condition is met since Ψ_{t+1} is a constant. Obviously hysteresis is a possibility in this model since $a_t[0] < a_t[1]$, for all t.

VI. Concluding Remarks

This paper demonstrates the existence and characterizes the nature of sunk cost hysteresis for for a broad class of processees generating the forcing variable. Most notably this class includes rational or model–generated expectations. Specifically we show that the hysteresis band tends to widen with greater sunk costs, but the effect of greater uncertainty on the band width depends upon the specific nature of the process generating the uncertainty. Lastly we show that greater persistence in the shocks has the effect of making well–entrenched firms more likely to exit, of narrowing the band for marginal firms and of making unlikely entrants more likely to enter.

The paper uses discrete–time dynamic programming techniques to derive the optimal entry–exit strategy. It would appear that these techniques might be useful complement to the very powerful techniques discussed by Dixit (1987a, b). In particular in a wide variety of multi–period problems where it is cheaper to get "in" than to stay "in" (broadly interpreting the meaning of "in"), the analytics in this paper could easily be extended to show the possibility and characterize the nature of hysteresis. These include problems that involve irreversibilities and sunk costs. The finite time dynamic programming algorithm involves mathematics which are simple enough to allow the introduction of more complicated economic considerations. The drawback is that — unlike the continuous–time Brownian motion analytics employed by Dixit (1987a, b) — it is often not possible to explicitly solve for the values of end points of the no–entry–no–exit band. Nonetheless for many analytic and empirical purposes it is not really necessary to find a closed form solution for these critical points. This paper shows that a fairly informative characterization of the band can be derived even without a closed form solution.

APPENDIX 1

Assuming that:

(a1) $\pi[s_t,x_t]$ attains a finite minimum and maximum value and is continuous in all its arguments, and decreasing in its first argument,

(a2) $x_{t+1}= k[s_t,x_t,U_t]$ is a continuous function which is known to the firm,

(a3) the pdf of s, $P'[\cdot]$, is everywhere finite, F and M are finite and

(a4) $EV_T[s_T,x_T,A_T] = c$, for all s_T, x_T and A_T, where c is a finite number,

(a5) the functions π and k are such that $(\pi[\cdot,\cdot]/\partial s_t + \delta\partial\Psi_{t+1}[\cdot,\cdot]/\partial s_t)$ is negative,

then there exists an optimal control policy $\phi = (\mu_0,\mu_1,...,\mu_{T-1})$ where each μ_t maps the state space into the control space, $\{1,0\}$, and has the form (2.6).

The proof is by mathematical induction. Its outline is as follows. In lemma 1, we show by construction that if $EV_{t+1}\left[s,k[s,x_t,U_t],A_{t+1}\right]$ exists and has a finite value and the slope condition in (a5) holds then optimal decision rule in period t is of the nature specified in (2.6). In lemma 2 we show by construction that if $EV_{t+1}\left[s,k[s,x_t,U_t],A_{t+1}\right]$ exists and is finite then the value function for period t, $V_t[s_t,x_t,A_t]$ exists and is finite. In lemma 3 we show that if $V_t[s_t,x_t,A_t]$ exists and is finite and the decision rule has the (2.6) form and then the expectation $EV_t\left[s,k[s,x_{t-1},U_{t-1}],A_t\right]$ exists and is finite. Lastly to start off the chain of mathematical induction, we show by construction in lemma 4 that if $V_T[\cdot,\cdot,\cdot] = c$, then $V_{T-1}[\cdot,\cdot,\cdot]$ exists and is finite.

Proof of the lemmas is as follows.

Lemma 1: If $EV_{t+1}\left[s,k[s_t,x_t,U_t],A_{t+1}\right]$ exists and is finite for all (s_t,x_t), and if $(\pi[\cdot,\cdot] + \delta\partial\Psi_{t+1}[\cdot,\cdot])/\partial s_t$ is well defined and is negative, where $\Psi_{t+1}[s_t,x_t]$ is defined in (2.4), then the optimal decision rule, $\mu_t[s_t,x_t,A_t]$,is: Set $U_t= 1$, iff $s_t \leq \alpha_t$, s.t.:

$$\pi[\alpha_t,x_t] + \delta\Psi_{t+1}[\alpha_t,x_t] = F + DA_t, \quad \text{(A1.1)}$$

The proof is simple. If the firm chooses to be active it is worth: $\pi[s_t,x_t]-F-DA_t+\delta EV_{t+1}[s,k[s_t,x_t,1],1]$; if it chooses to be inactive it is worth: $\delta EV_{t+1}[s,k[s_t,x_t,0],0]$. Since these are both finite numbers, one most be at least as large as the other. The optimal strategy is to choose the action that leads to the highest value. By assumptions (a1), (a5) and (a2) there is only one value of s_t for any (x_t, A_t), call it α_t, for which the two alternatives are equal. Furthermore (a5) implies that for values of s_t less than α_t the activity option is preferable.

Lemma 2: If the expectation, $EV_{t+1}\left[s,k[s_t,x_t,U_t],A_{t+1}\right]$, exists and is finite, then the function $V_t[s_t,x_t,A_t]$ also exist and is finite.

The proof is by construction. $V_t[s_t,x_t,A_t]$ equals:

$$\text{(A1.2)} \qquad \max\left[\pi[s_t,x_t]-F-DA_t+\delta EV_{t+1}[s,k[s_t,x_t,1],1],\delta EV_{t+1}[s,k[s_t,x_t,0],0]\right]$$

By (a1), (a3) and the supposition, V_t is finite.

Lemma 3: If $V_t[s_t,x_t,A_t]$ exists and is finite then $EV_t\left[s,k[s,x_{t-1},U_{t-1}],A_t\right]$ exits and is finite.

Proof is by construction. Using lemma 1 and the supposition implies that the optimal decision rule is defined by α_t so $EV_t\left[s,k[s,x_{t-1},U_{t-1}],A_t\right]$ equals:

$$\text{(A1.3)} \qquad \int_{z=-\infty}^{\alpha_t}\left[\pi[z,k[s_{t-1},x_{t-1},1]]-F-DA_t+\delta E\{V_{t+1}[s,x_{t+1},1]\}\right]dP[z]$$

$$+\int_{z=\alpha_t}^{\infty}\{\delta EV_{t+1}[s,x_{t+1},0]\}dP[z],$$

where for notational convenience, we did not write out:

$$x_{t+1}=k\left[z,k\left[s_{t-1},x_{t-1},U_{t-1}\right],\mu_t\left[z,k[s_{t-1},x_{t-1},U_{t-1}],A_t\right]\right],$$

The expectations here are over s_{t+1}. Given the supposition, (a1), (a2) and (a3), it is obvious that the expectation exist and is a finite number.

Lemma 4: $V_{T-1}[s_{T-1},x_{T-1},A_{T-1}]$ exists and is finite number.

Again proof is by construction and is isomorphic to the proof of lemma 2, noting that $EV_T[\cdot,\cdot,\cdot]=c$.

APPENDIX 2: Convergence and Existence in Infinite Horizon Case.

Our proof technique in this section requires that the cash flow in any period be non–negative. By the boundedness of π_t and the finiteness of F and M, we know that $L \leq \pi[s_t,x_t]-DA_t \leq H$, where

L and H are finite numbers. If we restricted both L and H to be non–negative then the proof would be more direct. However, we do not wish to restricted L to be non–negative since we might very well want to consider cases where a firm can lose money this period in expectation of profits in future periods (of course if H is negative then the firm would never operate so to avoid a vacuous problem we assume that $H > 0$). To accommodate these two requirements we prove convergence for a related infinite horizon problem and then show that our result also implies convergence and optimality for the original problem.

Consider a vector of state variables, $q_t = (s_t;x_t;A_t)$, that is an element of a space S and which obeys the laws of motion implicit in the discrete–time dynamic system:

(A2.1) $$q_{t+1} = f[q_t,u_t]$$

where u_t is an element of a non–empty space C, and s_t is an element of a space D and is characterized by the cdf $P[s]$. Given the initial state q_0 the problem is to find a control policy $\phi = \{\mu_0,\mu_1,...\}$ such that $\mu_k[q_t]:S \rightarrow C, \forall k=0,1,...$ (we call such control policies admissible policies and denote the set of all such policies as Ω) and which maximizes the functional:

(A2.2) $$J_\phi[q_0] = \lim_{T\rightarrow\infty} E\{\sum_{t=0}^{T-1} \delta^t g\left[q_t,\mu_t[q_t]\right]\}, \text{ where } 0 \leq \delta < 1 \text{ and}$$

(A2.3) $$g[q_t,u_t] = \begin{cases} \pi[s_t,x_t] - DA_t + K, & \text{if } u_t=1 \\ K, & \text{if } u_t=0 \end{cases}$$

subject to (A2.1) and such that

(A2.4) $$K = \text{Max}[-L, 0].$$

By adding this appropriately chosen finite constant to each period's cash flow, we have:

(A2.5) $$0 \leq g[q_t,u_t] \leq B, \text{ such that } B = K + H.$$

Introducing more notation, we call the optimal value function J^*, and define it by,

$$J^*[q] = \sup_{\phi \in \Omega} J_\phi[q].$$

Consider the T stage problem derived from (A2.2) by truncation. The optimal value of this sub–problem is: $J^*_T[q] = \sup_{\phi_T \in \Omega_T} J_{\phi_T}[q]$ (where Ω_T is the set of all admissible ϕ_T) and is given by the T–th step of the dynamic programming algorithm as described in appendix 1 (reversing time subscript so that 0 is now the last period and T is the first). To reiterate:

(A2.6) $$J_0[q] = 0, \ \forall q \in S$$

(A2.7) $$J_{t+1}[q] = \sup_{u \in C} \left(g[q,u] + \delta E\{J_t\left[f[q,u]\right]\} \right), \ \forall q \in S, \ t=0,1,\ldots,T-1.$$

where the expectation is with respect to the period t realization of s, and we have dropped the time subscripts on q and u for notational convenience (in (A2.7) the q and u would otherwise all have been q_{t+1} and u_{t+1}).

Proposition: (convergence) The optimal value of (A2.2), J^*, is such that:

(A2.9) $$0 \leq J^*[q] \leq B/(1-\delta), \ \forall q \in S, \text{ and}$$

(A2.10) $$J^*[q] = \lim_{T\to\infty} J_{\phi_T}[q], \ \forall q \in S$$

To show (A2.9) involves an obvious use of (A2.5). To show (A2.10), use the fact that by definition of a supremum:

(A2.11) $$J^*[q] \geq J_{\phi_T}[q] + K(\frac{\delta^T}{1-\delta}).$$

That is, if we used the dynamic programming algorithm to choose the control policy, ϕ_T, for the first T periods and then set $u_t = 0$ for all subsequent periods, the value of the firm would be that given by the right hand side of (A2.11) and obviously J^* is at least as great as this arbitrary control policy. Similarly J^* can be no greater than the value of the firm if it optimized for the first T periods and then received the maximum cash flow for all subsequent periods, i.e.:

(A2.12) $$J^*[q] \leq J_{\phi_T}[q] + B(\frac{\delta^T}{1-\delta}).$$

Plainly then

$$J_{\phi_T}[q] + K(\frac{\delta^T}{1-\delta}) \ \leq \ J^*[q] \ \leq \ J_{\phi_T}[q] + B(\frac{\delta^T}{1-\delta})$$

and since $K(\frac{\delta^T}{1-\delta})$ and $B(\frac{\delta^T}{1-\delta})$ go to zero as T goes to infinity, we have

$$J^*[q] = \lim_{T\to\infty} J_{\phi_T}[q].$$

Now since we know that the value function J_t converges to J^* as $T\to\infty$, we know that the optimal control policy for (A2.2) is stationary since with an infinite horizon the maximization problem is identical in every period. Indeed the optimal policy is given by (skipping over some mathematical details more fully covered in Bertsekas 1972):

$$\mu[q_t] = u_t, \ \forall\, q_t \in S,$$

where the function μ is implicitly defined by:

$$J^*[q_t] = \sup_{u_t \in C} \left(g\left[q_t, u_t\right] + \delta E\{ J^*\left[f[q_t, u_t]\right] \} \right).$$

Lastly we take the fairly obvious step of showing that the optimal policy for the modified problem is also optimal for the original problem. Since we added K to the cash flow in each period, we know that

$$V^*[q] = J^*[q] - K\left(\frac{1}{1-\delta}\right)$$

so the $\mu[q]$ which maximizes expected discounted cash flow plus a constant, as in (A2.2), also maximizes the expected discounted cash flow without the constant.

<u>**APPENDIX 3**</u>: Evaluating $(\partial E\Psi/\partial F)$ for AR1 expectations

By definition $\Psi[s_t, F]$ evaluated at $s_t = q$ is:

(A3.1)
$$(F-M)P[\alpha^I - \rho q] + \int_{z=\alpha^I}^{\alpha^O} \{\pi[z] - M + \delta\Psi[z,F]\} dP[z - \rho q]$$

where $P[\cdot]$ is the cdf of ϵ in (2.12). Differentiating with respect to F using the envelope cancellations:

(A3.2)
$$\Psi_F[q,F] = P[\alpha^I - \rho q] + \int_{\alpha^I}^{\alpha^O} \{\delta\Psi_F[z,F]\}\, dP[z - \rho q]$$

Repeatedly substituting this formula into itself, we have that Ψ_F equals:

(A3.3)
$$P[\alpha^I - \rho q] + \left[\int_{z_{t+1}=\alpha^O}^{\alpha^I} \delta P[\alpha^I - \rho z_{t+1}]\, dP[z_{t+1} - \rho q] \right]$$
$$+ \left[\int_{z_{t+2}=\alpha^O}^{\alpha^I} \int_{z_{t+1}=\alpha^O}^{\alpha^I} \delta^2 P[\alpha^I - \rho z_{t+2}] dP[z_{t+2} - \rho z_{t+1}] dP[z_{t+1} - \rho q] \right]$$
$$+ \left[\int_{z_{t+3}=\alpha^O}^{\alpha^I} \int_{z_{t+2}=\alpha^O}^{\alpha^I} \int_{z_{t+1}=\alpha^O}^{\alpha^I} \delta^3 P[\alpha^I - \rho z_{t+3}] dP[z_{t+3} - \rho z_{t+2}] dP[z_{t+2} - \rho z_{t+1}] dP[z_{t+1} - \rho q] \right] + \ldots$$

Two points are immediately obvious. First this quantity is positive since it is the discounted sum of probabilities. Second, to analytically evaluate this expression is not a easy task since it involves integrals of a cdf.

As explained in the text, the signing of $d\alpha^O/dF$ only requires that the expression is positive. However the sign of $d\alpha^I/dF$ depends on whether the above expression is less than unity. To this end, we make a number of substitutions that simplify the expression. The resulting expression is equal to a number C, and due the nature of the substitutions we know that Ψ_F is less than C. Thus if we can show that C is less than unity, we show that Ψ_F is also less than unity. Specifically, consider the first term in large parentheses in (A3.3). The difficulty here is that the integral of the cdf, $P[\cdot]$, is not an easily interpretable function. However we know that the value of $P[\alpha^I - \rho z_{t+1}]$ over the range of integration is less than or equal to $P[\alpha^I - \rho\alpha^I]$. Thus if we substituted this constant in place of the function the value of the resulting integral is greater than the value of the original integral. With this substitution:

(A3.4) $$\delta P[\alpha^I - \rho\alpha^I](P[\alpha^O - \rho q] - P[\alpha^I - \rho q]) > \left[\int_{\alpha^O}^{\alpha^I} \delta P[\alpha^I - \rho z_{t+1}]\, dP[z_{t+1} - \rho q]\right].$$

Using a similar substitution in the second term in large parentheses is less than:

(A3.5) $$\delta^2 P[\alpha^I - \rho\alpha^I] \int_{\alpha^I}^{\alpha^O} \{P[\alpha^O - \rho z_{t+1}] - P[\alpha^I - \rho z_{t+1}]\} dP[z_{t+1} - \rho q]$$

Making similar substitutions for $P[\alpha^O - \rho z_{t+1}]$ and $P[\alpha^I - \rho z_{t+1}]$, we have that the second term in large parentheses from equation (A3.3) is less than:

(A3.6) $$\delta^2 P[\alpha^I - \rho\alpha^I](P[\alpha^O - \rho\alpha^I] - P[\alpha^I - \rho\alpha^O])(P[\alpha^O - \rho q] - P[\alpha^I - \rho q]).$$

Repeating performing these substitutions gives us an expression, C, which is greater than the original expression (A3.3) and yet involves values which are simple to interpret. Thus C equals:

(A3.7) $$P[\alpha^I - \rho q] + \delta P[\alpha^I - \rho\alpha^I]\beta + \delta^2 P[\alpha^I - \rho\alpha^I]\varphi\beta + \delta^3 P[\alpha^I - \rho\alpha^I]\varphi^2\beta + \ldots,$$

where $\varphi = (P[\alpha^O - \rho\alpha^I] - P[\alpha^I - \rho\alpha^O]$, and $\beta = P[\alpha^O - \rho q] - P[\alpha^I - \rho q]$.

Solving the infinite sum evaluated at $q = \alpha^I$, we have that:

(A3.8) $$C = P[\alpha^I - \rho q](1 + \frac{\delta\beta}{1 - \delta\varphi}).$$

Both β and φ are fractions which depend on α^I and α^O, and both are zero at $F = M$.

Appendix 4: Band Widening for Rho Close to Unity

Here we show that for the ρ sufficiently close to unity, a symmetric MPS widens the band assuming that s is expected to follow a random walk and $\pi[\cdot]$ is linear. The proof is as follows: We show that the band widens for each period in the T–period finite horizon case, and then apply the usual limiting arguments to assert that the band would widen for the infinite horizon case. In period T–1 the band is unaffected by the MPS since no uncertainty is present when the firm makes its entry–exit decision. In T–2, α^{I}_{T-2} and α^{O}_{T-2} depend on $\Psi^{T-1}[\cdot,r]$ (here we denote the period with superscripts to make room for subscripts denoting partial derivatives). The definition of Ψ^{T-1} evaluated at $s_{T-2}=q$ is:

(A4.1) $$(F-M)P[\alpha^{I}_{T-1}-\rho q,r] + \int_{z=\alpha^{I}_{T-1}}^{\alpha^{O}_{T-1}} \{\pi[z]-M\}dP[z-\rho q,r]$$

its partial with respect to r is (integrating by parts and using the usual cancellations):

(A4.2) $$\Psi^{T-1}_{2}[q,r] = -\int_{\alpha^{I}_{T-1}}^{\alpha^{O}_{T-1}} \pi'[z]\, P_{2}[z-\rho q,r]\, dz$$

where $P_2[\cdot,\cdot]$ denotes the partial of $P[\cdot,\cdot]$ with respect to its second argument. Consider the case where $\rho = 1$. To see that (A4.2) is negative at $q = \alpha^{I}_{T-1}$, and positive at $q = \alpha^{O}_{T-1}$, turn to figure A4.1 which plots $P_2[\cdot,r]$. By definition of a MPS, $P_2[\cdot,r]$ is positive to the left of the mean of ϵ, and negative to the right since ϵ has a symmetric, single–peaked distribution. If $q = \alpha^{I}_{T-1}$, then $z-q$ is always non–negative (zero at $z=\alpha^{I}_{T-1}$, positive everywhere else over the interval of integration), so $P_2[z-q,r]$ is always non–positive (zero at $z=\alpha^{I}_{T-1}$, negative everywhere else over the interval of integration). By inspection of (A4.2), these facts imply that $\Psi^{T-1}_{2}[\alpha^{I}_{T-1},r]$ is negative. Since this is a strict inequality and ρ enters continuously, $\Psi^{T-1}_{2}[\alpha^{I}_{T-1},r]$ is negative for ρ close enough to unity. Similar reasoning shows that $\Psi^{T-1}_{2}[\alpha^{O}_{T-1},r]$ is positive. Also by continuity, $\Psi^{T-1}_{2}[q,r]$ is increasing in q. Turning to period T–2 (integrating by parts and using the cancellations)

$$(A4.3) \qquad \Psi_2^{T-2}[q,r] = -\int_{\alpha^I_{T-2}}^{\alpha^O_{T-2}} \{\pi'[z]+\delta\Psi_1^{T-1}[z,r]\}P_2[z-\rho q,r]dz + \delta\int_{\alpha^I_{T-2}}^{\alpha^O_{T-2}} \{\Psi_2^{T}[z,r]\}dP[z-\rho q,r]$$

By the same type of reasoning used to sign (A4.2), the first term in A4.3 is negative at $q = \alpha^I_{T-2}$ and positive at $q = \alpha^O_{T-2}$. Also since $\Psi_2^{T-1}[z,r]$ is increasing in z for $\rho = 1$, then we know that:

$$(A4.4) \qquad \delta\int_{\alpha^I_{T-1}}^{\alpha^O_{T-1}} \{\Psi_2^{T-1}[z,r]\}dP[z-\rho\alpha^O_{T-2},r] > \delta\int_{\alpha^I_{T-1}}^{\alpha^O_{T-1}} \{\Psi_2^{T-1}[z,r]\}dP[z-\rho\alpha^I_{T-2},r].$$

Clearly then:

$$\Psi_2^{T-1}[\alpha^O_{T-2},r] - \Psi_2^{T-1}[\alpha^I_{T-2},r] > 0$$

This fact together with the linearity of $\pi[\cdot]$ implies that the band widens with a MPS in T−2. Also note that by continuity $\Psi_2^{T-1}[x,r]$ is increasing in x. Repeated applications of these step shows that the band widens in all periods (formally we would use mathematical induction). By the limiting argument in appendix 2, this implies that the band widens for ρ sufficiently close to unity even in the infinite horizon approximation.

FOOTNOTES

* I gratefully acknowledge the helpful comments and suggestions of Rich Lyons, Alberto Giovannini, Ricardo Caballero, Mike Gavin and Aaron Tornell. An early draft of this paper was circulated with the title "Sunk Cost Hysteresis with General Exchange Rate Expectations".

1. American Heritage Dictionary defines hysteresis as "failure of a system changed by an external agent to return to its original value when the cause of the change is removed."

2. Irreversible investment, per se, does not lead to the possibility of hysteresis; the important aspect is that the investment be firm–specific so that the cost is unrecoverable. For instance, if the investment was irreversible but could be resold to another firm at its replacement cost, hysteresis is not a possibility. Ratchet effects differ from hysteresis in that they cannot be reversed by a corrective shock. Path dependencies create the possibility of one type of hysteresis since the long run steady–state depends on the path.

3. For instance in the sticky price monetary model of exchange rate determination (Dornbusch 1976 and extensions), if the UK interest rate is higher than that in the US, then investors must expect the dollar to appreciate if they are to be happy holding dollar assets. This is inconsistent with the random–walk–with–a–drift assumption since stability requires that investors expect the interest rate differential to erode over time.

4. Consider $G_1[z,x] \equiv g[z,x]$, where the subscript indicates the partial with respect to the first argument, and the integral:

$$I[x] = \int_{w=a[x]}^{b[x]} g[w,x]dw = G[b[x],x] - G[a[x],x].$$

Clearly, $$dI[x]/dx = G_1[b[x],x]b'[x] + G_2[b[x],x] - G_1[a[x],x]a'[x] - G_2[a[x],x]$$

which by rearrangement and the fundamental theorem of calculus is equal to:

$$G_1[b[x],x]b'[x] - G_1[a[x],x]a'[x] + \int_{w=a[x]}^{b[x]} g_2[w,x]dw.$$

REFERENCES

Baldwin, Richard. "Hysteresis in Import Prices: The Beachhead Effect." American Economic Review, September 1988a, 78.

_________. "Some Empirical Evidence on Hysteresis in Aggregate U.S. Import Prices." NBER Working Paper No. 2483, January 1988b.

_________. "Hysteresis in Trade." M.I.T. mimeo, prepared for the 1986 NBER Summer Institute, April 1986.

__________, and Paul Krugman. "Persistent Trade Effects of Large Exchange Rate Shocks." Quarterly Journal of Economics, forthcoming, 1989.

__________, and Richard Lyons. "The Mutual Amplification Effect of Exchange Rate Volatility and Unresponsive Trade Prices." NBER Working Paper No. 2677, August 1988a.

__________, and ______________. "Exchange Rate Hysteresis: The Effects of Large vs Small Policy Misalignments." NBER Working Paper No. ????, December, 1988b.

Bean, Charles. "Sterling Misalignment and British Trade Performance," Centre for Labour Economics, LSE, Discussion Paper No. 288, July 1987.

Bellman, Richard. Dynamic Programming. Princeton University Press, 1957.

Bertola, Giuseppe. "Irreversible Investment," MIT Mimeo, October 1987.

Bertsekas, Dimitri P. Dynamic Programming and Stochastic Control. New York: Academic Press, 1976.

Blanchard, Olivier and Lawrence Summers. "Hysteresis in the Unemployment Rate," NBER Macroeconomics Annual, 1986.

Bryant, Ralph, Gerald Holtham and Peter Hooper (eds). External Deficits and the Dollar: the Pit and the Pendulum. Washington, DC, Brookings Institution, 1988.

Dixit, Avinash. "A Model of Duopoly Suggesting a Theory of Entry Barriers." The Bell Journ of Economic, 1979.

_______, "The Role of Investment in Entry–Deterrence." The Economic Journal, 1980.

_______, "Entry and Exit Decisions Under Fluctuating Real Exchange Rates." Princeton Mimeo, September 1987a.

_________. "Hysteresis, Import Penetration, and Exchange Rate Pass–Through." Princeton Mimeo, November 1987b.

Eaton, C and R. Lipsey. "Exit Barriers are Entry Barriers: the Durability of Capital as a Barrier to Entry." The Bell Journal of Economics, 1980.

_______,and _______. "Capital, Commitment and Entry Deterrence." The Bell Journal of Economics, 1981.

Kemp, Murray and Henry Wan. "Hysteresis of Long–Run Equilibrium From Realistic Adjustment Costs," in George Horwich and Paul Samuelson (eds.) Trade, Stability, and Macroeconomics. New York: Academic Press, 1974.

Krugman, Paul. "Exchange Rate Instability," Lord Robbins Lectures, London School of Economics, 1988

__________, and Richard Baldwin. "The Persistence of the US Trade Deficit," Brookings Papers on Economic Activity, 1, 1987.

Mann, Catherine. "Prices, Profit Margins, and Exchange Rates." Federal Reserve Bulletin, June 1986.

__________. "Prices, Profit Margins, and Exchange Rates: After the Fall." DIF Discussion Paper, Federal Reserve Board, July 1987a.

__________. "The Puzzling Behavior of Non–oil Import Prices." Mimeo, November 1987b.

McDonald, Robert and Daniel Siegel. "Investment and the Valuation of Firms when There is an Option to Shut Down." International Economic Review, 26, 1985.

Orcutt, Guy. "Measurement of Price Elasticities in International Trade." Review of Economics and Statistics. 32 (May 1950): 117–132.

Phelps, Edmund, Inflation Policy and Unemployment Theory, W.W.Horton, NY, 1972.

Schmalensee, Richard. "Product Differentiation Advantages of Pioneering Brands." American Economic Review, 72, 3, 1982

Stiglitz, Joseph. "Technological Change, Sunk Costs, and Competition," in Brookings Papers on Economic Activity 3, 1987, Special Issue on Microeconomics, Martin Baily and Clifford Winston (eds).

Williamson, Oliver. "Selling Expenses as a Barrier to Entry." Quarterly Journal of Economics, 1963.

Figure 1: Hysteresis Band for Finite Horizon Case

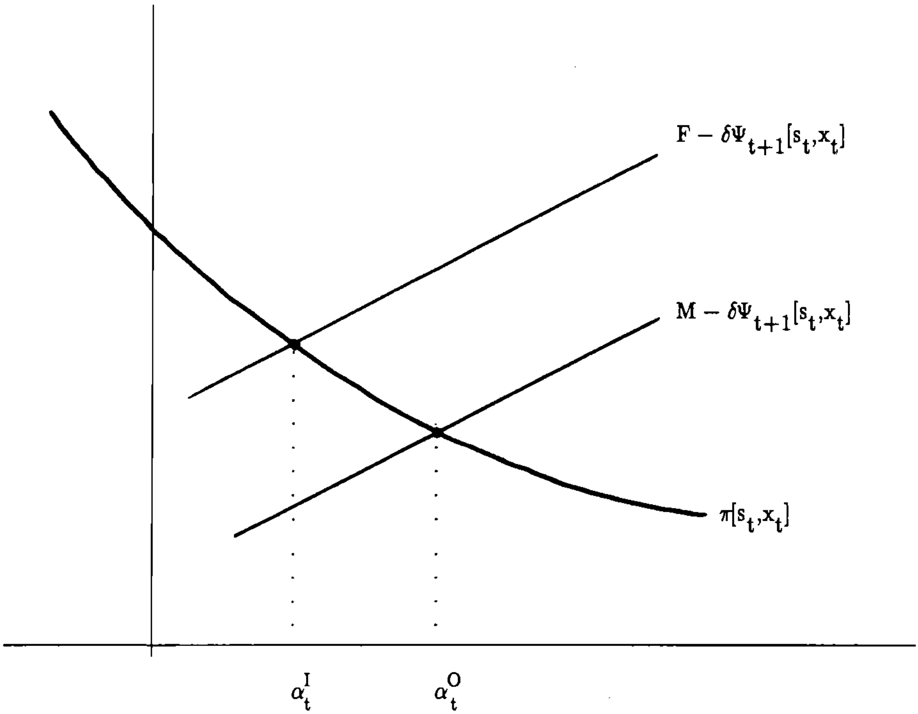

Figure 2: Hysteresis Band with Reversed Slope Condition

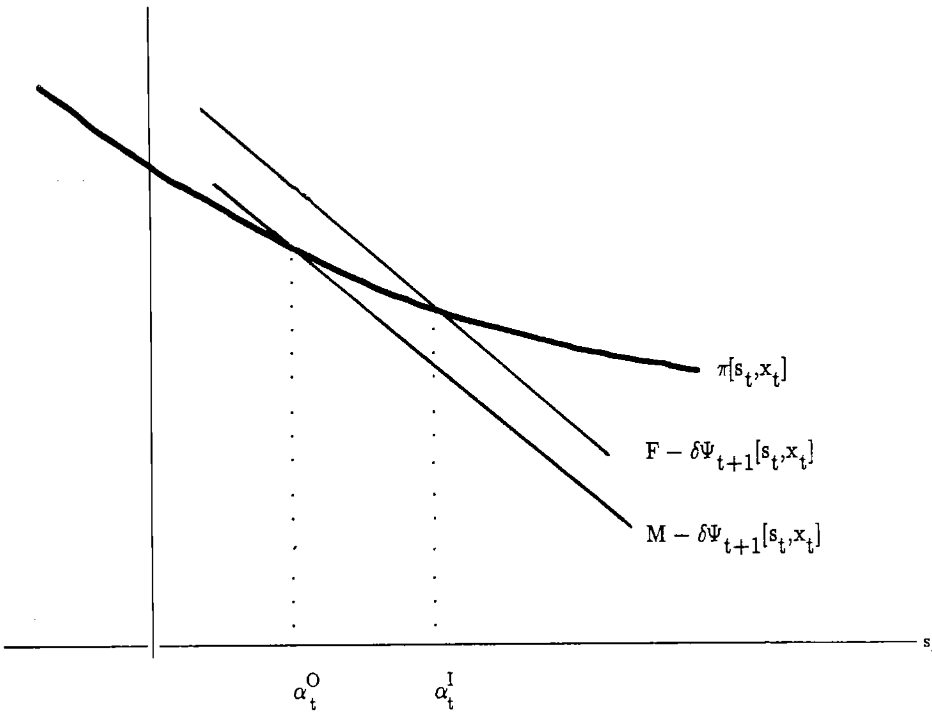

__Figure 3__: The Effect on Ψ of Changes in Rho

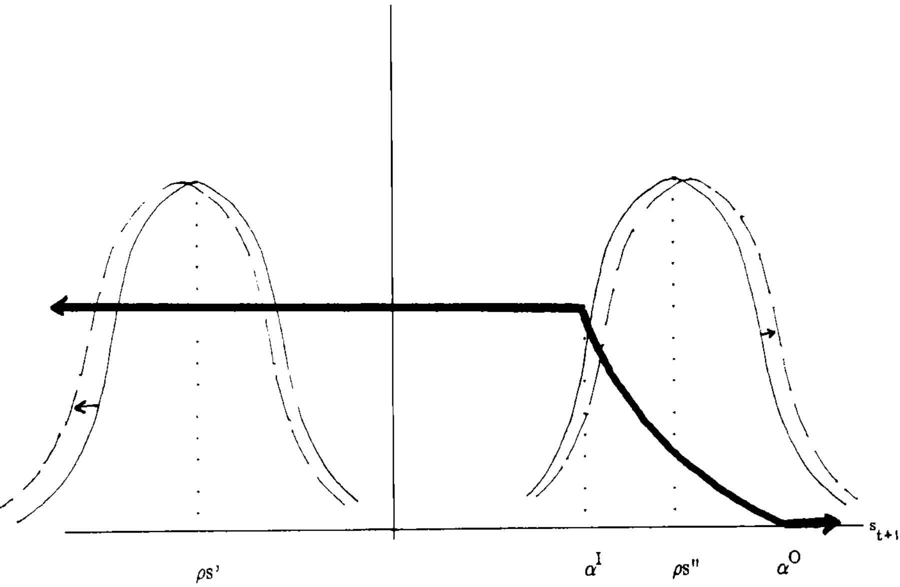

Figure 4: Counter Example of Band Width and Volalitity

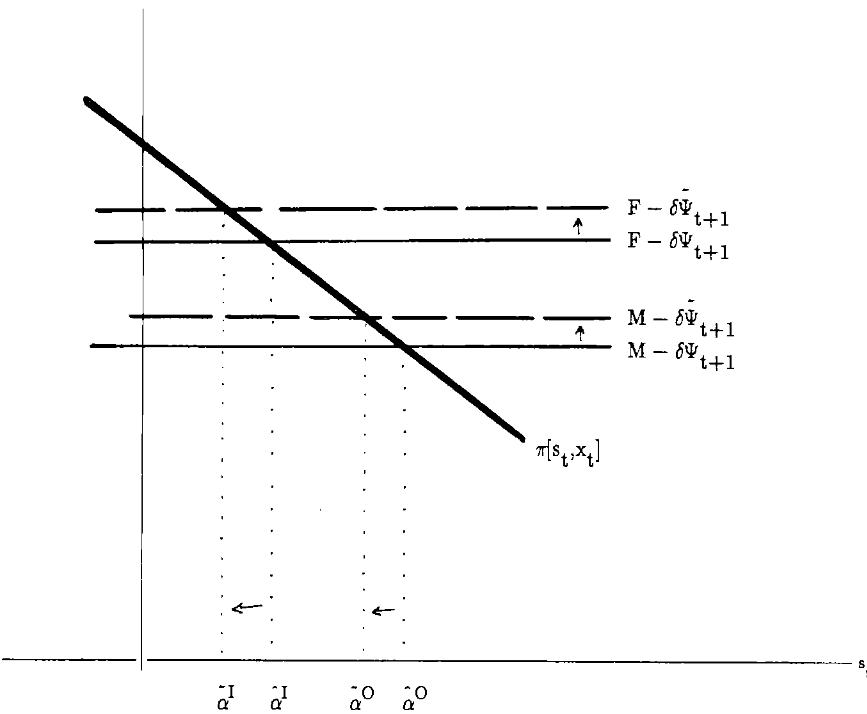

Figure A4.1:

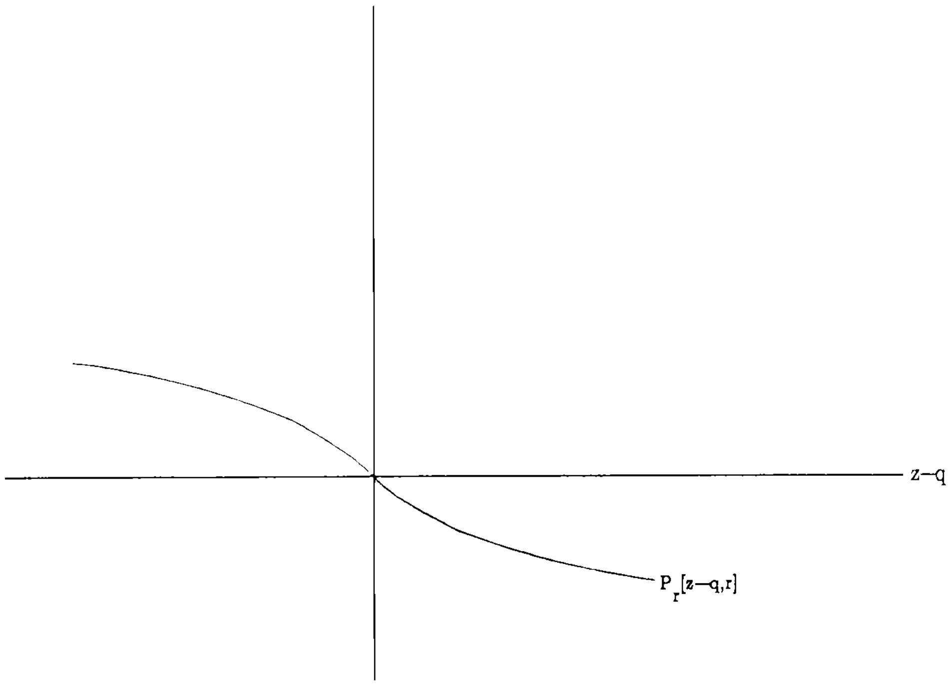

Figure 5: Band Widening for Random Walk

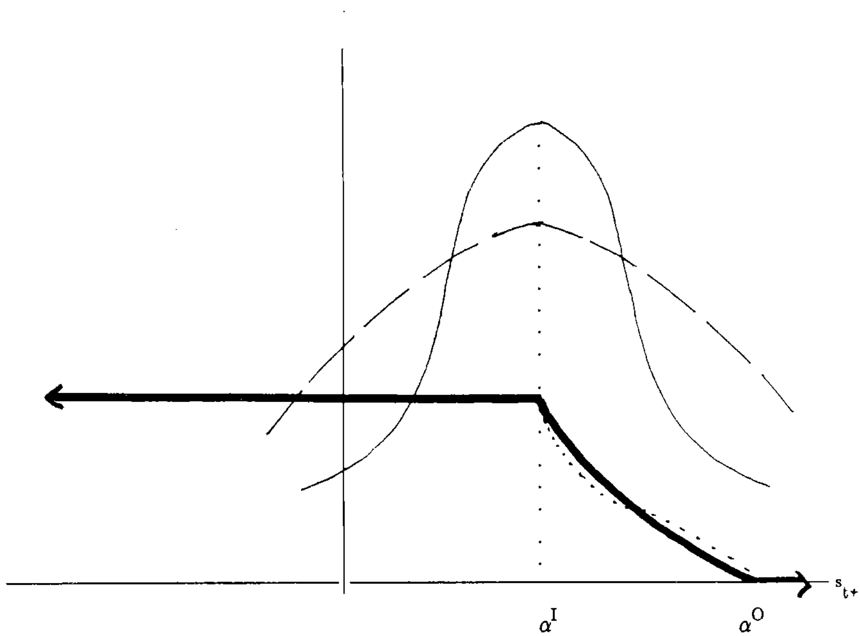

Figure 6: Hysteresis Band for Maximization in Foreign Currency