

Vertical Distribution, Parallel Trade, and Price Divergence in Integrated Markets*

Mattias Ganslandt[†] The Research Institute of Industrial Economics
Keith E. Maskus[‡] University of Colorado at Boulder

May 2006

Abstract

We develop a model of vertical pricing in which an original manufacturer sets wholesale prices in two markets that are integrated at the distributor level by parallel imports (PI). The manufacturing firm needs to set these two prices to balance three competing interests: restricting competition in the PI-recipient market, avoiding resource wastes due to actual trade, and reducing the double-markup problem in the PI-source nation. These tradeoffs imply the counterintuitive result that both retail prices could diverge as a result of declining trading costs, even as the volume of PI increases. Thus, in some circumstances it may be misleading to think of PI as an unambiguous force for price integration.

JEL Codes: F1, L1

Keywords: Parallel imports, vertical control, price divergence

*We thank the Swedish Competition Authority and the Jan Wallander and Tom Hedlius Foundation for financial support. We are also grateful for comments from Sofronis Clerides, Wilfred Ethier, Jim Markusen, Lars Persson and two referees. We also thank seminar participants at Lund University, University of Kiel, IUI, NOITS 2004 the CEPR International Research and Policy Symposium, and the 2005 spring Midwest International Economics meeting.

[†]The Research Institute of Industrial Economics, P.O. Box 55665, SE-102 Stockholm, SWEDEN e-mail mattias.ganslandt@iui.se, fax +46-8-6654599.

[‡]Corresponding Author: Department of Economics, Campus Box 256, University of Colorado, Boulder, CO 80309-0256, Phone (303) 492-7588, e-mail: keith.maskus@colorado.edu.

1. INTRODUCTION

The European Union has adopted a rigorous regime of regional exhaustion, which states that parallel trade - i.e. cross-border arbitrage - is impermissible from outside the community but the first sale within its territory exhausts the rights of the original manufacturing firm to restrict further trade of its product in the common market.¹ Both the European Commission and the European Court of Justice have consistently defended the presumed importance of parallel trade in support of the single market (Ganslandt and Maskus, 2003). The essential justification for this policy is a belief that parallel imports generate competition at the retail level, inducing a tendency toward retail price convergence and pro-competitive gains from price integration.²

Despite these long-standing efforts to integrate markets, there remains considerable retail price divergence within the European Union, which continues to puzzle observers. For example, the European Central Bank (2002, p. 39), recently stated that, "The data available suggest that price level dispersion for many tradeable goods and services remains higher between euro area countries than within individual countries, implying that further improvements in the functioning of the Internal Market are possible."

One explanation for this failure of prices to move together could be that imperfectly competitive firms may be capable of sustaining divergent retail prices in reaction to reductions in trade barriers. In this paper we set out such a model and explore fully the price impacts of PI in a setting where an original manufacturer - i.e. the holder of the intellectual property rights - sells its product through independent distributors in two national markets.³ Our point of departure is the observation that the bulk of PI actually takes place at the wholesale level, rather than the retail level (Maskus and Chen, 2002; National Economic Research Associates [NERA], 1999).⁴ Because manufacturing firms place goods

¹The policy is essentially equivalent to the U.S. "first-sale doctrine" under which initial sale of a good inside the territorial United States exhausts further rights to control distribution. Thus, our analysis of PI and exclusive territories applies also to "gray-market" unauthorized trade within the United States.

²There exists evidence that this is indeed relevant in some markets. For instance, Ganslandt and Maskus (2004) present empirical evidence that competition from European PI firms in the pharmaceutical sector has a significantly negative effect on original manufacturers' prices in Sweden.

³Friberg and Martensen (2001) offer a model with similar conclusions but in a very different analytical context.

⁴The study by National Economic Research Associates (1999) reported survey evidence of significant flows of parallel trade within the European Union in the early 1990s. While the report tended to focus on retail price differences, it pointed out that the bulk of parallel trade happens at the wholesale or distributor level. Maskus and Chen (2002) performed econometric analysis with detailed U.S. export-price data and discovered that the international distribution of wholesale export prices did follow the predicted U-shaped relationship in U.S. tariff rates and that the level of such prices across destination regions depended on the legal treatment of parallel trade in those regions. Thus, empirical evidence points to the vertical-control

on the market initially through vertical contracts with local distributors, it is important to study the implications of such contracts for PI volumes and the consequent scope for retail price integration. In this context, EU competition policy also rigorously supports parallel trade at the distributor level as a means of reducing market power conveyed by exclusive territories.

In fact, we find that the claim that parallel imports unambiguously bring down retail prices in expensive locations is misleading. We show that in some market circumstances, for an important and empirically relevant range of trade costs, the existence of parallel imports can cause retail prices both to diverge between markets and to increase in high-cost locations, precisely opposite to the conventional intuition.⁵ The essential reason is that original manufacturing firms that own intellectual property rights undertake vertical pricing decisions to manage the threat and costs of PI, including consolidation of the number of distributors as trade costs fall. Thus, a reorientation of competition policy within the EU toward understanding vertical restraints seems in order.

In addition, we provide a welfare analysis of permitting parallel trade and compare it with market segmentation based on, for instance, exclusive territories or legal restraints.⁶ The analysis advanced here builds on the insights in a paper by Chen and Maskus (2005) in several important ways. They noted three essential efficiency tradeoffs in a model where a licensed distributor in one market can trade goods back to the home market of the original manufacturer. The manufacturer has to balance the losses from a pro-competitive price effect of PI into its own market, on the one hand, and the resource costs wasted in the activity of parallel trade and the double-markup problem in inducing a profit-maximizing retail price in the foreign market, on the other. The authors discovered a U-shaped wel-

problem as being central to PI. This line of inquiry was pursued extensively within the EU context by Ganslandt and Maskus (2003). Their empirical evidence supported the view that there are multiple causes for PI. Econometric analysis of European prices suggested that both horizontal arbitrage and vertical-control problems are important practical explanations for such trade. Pricing behavior by exporters from high-price markets - such as Denmark and the United Kingdom - indicated that such firms increase export prices in countries that are in close proximity (and therefore have low trade costs) in an attempt to deter PI.

⁵Arbitrage against differences in retail prices was the focus of early literature (Tarr 1985; Hilke 1988), which found that rapid and large dollar appreciation generated rising volumes of parallel imports into the United States.

⁶The welfare effects of price-integrating horizontal arbitrage have been studied in several papers. In an important theoretical paper, Malueg and Schwartz (1994) argued that a regime of uniform retail pricing would be globally inferior to one in which firms could price-discriminate on the basis of countries grouped by demand elasticity. Also arguing from a standard pricing model, Richardson (2002) argued that neither a global policy of uniform pricing (international exhaustion) nor of full segmentation (national exhaustion) could be supported as a Nash equilibrium, suggesting that negotiations to achieve a consistent global exhaustion regime at the World Trading Organization would be frustrated.

fare curve in the cost of trading PI goods. If trade costs are low, the optimal regulatory policy is to permit free parallel trade, for the pro-competitive gains dominate the other efficiency losses. However, at intermediate-to-high trade costs it is efficient to ban PI to avoid inefficient vertical pricing.

Our model is different and more general than that in Maskus and Chen (2002) and Chen and Maskus (2005) as we permit vertical distributors in two markets, an approach that generates a richer menu of potential price effects even within the framework of Cournot competition. With two independent distributors, permission of parallel trade may harm consumers due to the possibility that such trade could result in retail price divergence as trade costs fall. Consequently, consumers may gain if policy permitted exclusive territories. Maskus and Chen (2002) found a similar result but only for high PI costs.⁷

We believe that there is substantial scope for understanding better the implications of parallel trade by focusing on models of vertical pricing behavior in markets integrated at the wholesale level. It is conceivable that wholesale prices may be set in a way that offsets or even counteracts the anticipated impacts of an open PI regime. This is the point of departure for the analysis that follows.

The paper is organized as follows. In the next section we set out a basic model of parallel imports in two vertically controlled markets that are linked by the possibility of PI. The following section provides an analysis of three different cases that emerge as equilibria for different ranges of trade costs. In the fourth section we take a closer look at the equilibrium retail prices, profits and consumer welfare. In the fifth section we separately analyze the implications of introducing two important modifications of the model: retail arbitrage and product differentiation. We offer concluding remarks in the final section.

2. A THEORETICAL MODEL OF PARALLEL IMPORTS

We develop a model of vertical pricing with parallel imports. The model demonstrates the three deleterious impacts of PI on the manufacturing firm's profits. First, competition between distributors in the PI-recipient market reduces profits through a "pro-competitive effect" even as it raises consumer welfare there. Second, the act of engaging in PI wastes resources in transport, which reduces profits through a "trade-cost effect". Finally, the need to manage PI could generate an inability to set the optimal wholesale price of zero (equal

⁷In two related papers, Raff and Schmidt (2005a, 2005b) analyzed the implications of exclusive territories used by firms in international trade. They concluded that trade liberalization may lead manufacturers to offset the loss of tariff barriers with contracts imposing exclusive territories, which may decrease competition and welfare.

to marginal production costs) in the PI-source market. Specifically, this "double-markup effect" reduces profitability the greater the gap between wholesale price and marginal cost. The original manufacturer must strike a balance among these impacts in choosing its prices.

2A. Model Assumptions

Here we set out the basic model assumptions and offer a justification for them. Consider the situation in which an original manufacturer, M , sells its product in two countries, A and B . Firm M sells its product through an independent exclusive distributor in country A and another independent exclusive distributor in country B . The demand in A is $Q_A = 1 - p$, and that in B is $Q_B = S(1 - bp)$, where S is the (relative) population of market B , letting the population in A be normalized to 1. Parameter $b > 1$ governs (relative) demand elasticity. Manufacturer M has a constant marginal cost of production normalized to zero, and the marginal cost of retailing in both countries is normalized to zero as well.

In our primary model, goods sold in both markets are homogeneous, which is a reasonable assumption in the PI context because the ultimate origin of all products is the single manufacturing firm. However, the NERA (1999) survey suggests that PI goods may be considered slightly inferior to goods sourced domestically and we extend the analysis to a situation of quality differentiation in the final analytical section.

Initially, we shall assume that consumer markets are segmented. Neither consumers nor firms can buy the product at the retail level in one country and ship it to the other in quantities sufficient to affect prices. Segmentation of consumer markets at the retail level can be motivated on several grounds. In some industries arbitrage is illegal, while parallel trade at the wholesale level is permitted. This is the case with prescription drugs in the European Union and alcoholic beverages in the United States. In other industries, physical products and non-tradeable services are tied, causing effective market segmentation at the retail level. Local warranties bundled with capital goods and local calling plans bundled with cell phones are examples in this category. Finally, retail arbitrage can be prohibitively costly, while available margins could support parallel trade at the wholesale level. Near the end of the paper we shall relax this assumption and illustrate the effects of arbitrage based on retail price differentials.

Suppose that the manufacturing firm M can offer the distributor in market i ($i = A, B$) any contract in the form (w_i, T_i) , where w_i is the wholesale price at which the distributor purchases from M and T_i is a transfer payment (franchise fee) from the distributor to M . Thus, we assume that the manufacturing firm can only control supply with regular

two-part tariffs of the kind envisioned here. There is complete information, permitting the manufacturer to extract all profits from distributors with these contracts.⁸ We rule out contracts that incorporate an agreement to limit parallel trade directly or indirectly. The manufacturing firm can neither control quantities with supply curbs nor target parallel trade with a "dual-pricing" scheme (i.e., charging higher prices for products that are exported than for those sold domestically). We consequently focus on the effects of parallel imports for distributor-level competition when the manufacturing firm uses a standard strategy to solve the vertical control problem (differentiating wholesale prices while avoiding double marginalization) rather than trying to target parallel trade as such. More sophisticated strategies may also be illegal under competition law in jurisdictions such as the European Union.

Regarding competition, we assume that M cannot prevent the distributor in B from selling the product in market A , either directly or through such intermediaries as firms specializing in parallel trade. That is, either the manufacturer cannot legally limit the distributor's territory of sales or it is too costly to enforce any such constraint. If the distributor in B sells in market A it incurs an additional marginal cost of $t \geq 0$ and competes with the distributor in A in a Cournot fashion. We also assume that only the licensed distributor in market B can sell in that market. Thus, the situation we have in mind describes one market protected from PI competition but a vertical relationship that is vulnerable to wholesale-level competition in the other market.

There are several practical justifications for this assumption of one-way trade. First, if there are fixed costs (not explicitly modeled in this paper) and market B is small while market A is large, an asymmetry is possible, with products flowing from the small to the large country but not the other way around. Portugal and the United Kingdom within the European Union would exemplify this case. Second, there may be asymmetric product standards between the two countries. Third, the countries may vary in their legal treatment of PI, with international exhaustion the rule in A and national exhaustion in B . Examples of the former are Australia and Hong Kong, which are open to PI in copyright goods, and examples of the latter are Japan and the United States, which are not (Maskus, 2000).

Another important assumption is that the manufacturing firm sells the products to its distributors prior to the opening of the retail markets. The wholesale price can only be charged based on the quantity sold by the manufacturing firm to the distributor, but not

⁸Under incomplete information there would be a principal-agent problem, with two agents competing. This problem would be more difficult and would offer a greater incentive for vertical integration. However, our framework permits a tractable investigation of the tradeoffs arising in the common situations where vertical integration is incomplete.

on the quantities sold by the distributors to consumers. This is consistent with the manufacturing firm's inability to directly control or monitor quantities sold by the distributors. Consequently, the distributor contracts cannot be based on subsequent retail supplies and the manufacturing firm must therefore offer a contract with a non-negative wholesale price, i.e. $w_i \geq 0$. If the wholesale price could be negative, it would pay the distributor to order an infinite quantity from the manufacturing firm and sell only a limited quantity to consumers, a case we rule out on the grounds of realism.⁹

The situation we describe rules out other strategies or market circumstances that might mitigate the PI problem. For example, the manufacturer might acquire the B distributor and, through this vertical integration, prevent products from escaping to the A market. For another, if the manufacturer used competitive wholesale dealers in market B , it could preclude PI and avoid double marginalization by setting the monopoly price there, knowing that the dealers could not charge a markup themselves.¹⁰ Our justification for the specific case we consider is based on two general observations. First, in many products PI exist at the wholesale level, which would not be possible under vertical integration or competitive distribution, suggesting that some market power arises at that level. Second, it is common for original manufacturers to establish single, but independent, distributors in each national market. For example, the Coca-Cola company has separate general distributors for several countries in the EU. More specifically, Coca-Cola Enterprise SA is the sole licensed bottler in France, though it has five bottling plants and two distribution warehouses.¹¹ Automobiles also are commonly marketed through single wholesale dealers in the EU (NERA, 1999). To illustrate, Ferrari Great Britain Ltd. is the exclusive dealer for the company in the United Kingdom.¹² The recorded music industry may best exemplify our assumptions. The Sony-BMG company sustains single national distributors in most European countries, including such relatively low-income locations as Hungary, the Czech Republic, and Spain.¹³ At the same time, the volume of PI in recorded music within the EU overall was between five

⁹It is in this context that our assumption of zero marginal production costs is significant. As we discuss later, under some circumstances involving positive PI the manufacturing firm would like to subsidize sales in market A , charging a wholesale price below the marginal cost of production. However, the only way to avoid the distributor placing excessive order amounts is to have a strictly positive wholesale price. It is consequently impossible to subsidize the distributor unless the distributor can be forced to sell everything it has ordered from the manufacturing firm. This inefficiency in vertical pricing generates the outcome in our model that our intermediate pro-competitive range (where the non-negative wholesale price in A is a binding condition) would gradually diminish for higher marginal costs and at some point would vanish. Ganslandt and Maskus (2006) consider the case of positive marginal production costs.

¹⁰We are grateful to a referee for making this point.

¹¹See http://www2.coca-cola.com/ourcompany/cfs/cfs_france.htm, last visited 15 January 2006.

¹²See <http://www.ferrari.co.uk/home.php>, last visited 15 January 2006.

¹³See <http://www.sonybmg.com>, last visited 29 January 2006.

and ten percent of the aggregate market in 1998, with some releases experiencing up to 25 percent of sales arriving through parallel trade, in part because of low transport costs (NERA, 1999). The NERA report (p. 81) further noted about this industry that "vertical restraints...are not so tight that they would prevent parallel trade."

The strategy of having exclusive national dealers arises largely because of the need to support market-specific services, advertising and guarantees, which would be difficult to sustain under the free-riding that would exist with competitive distribution. Further, it limits reductions in quality or dilution of trademark value that could occur under open competition. Finally, there may be scale economies in distribution. Indeed, these are major pro-competitive justifications for permitting exclusive territories.

2B. Basic Analysis

Under the Cournot assumption, let the quantities sold in A by the two distributors be q_A^* and q_B^* , respectively, and the quantity sold in B by the sole distributor be q_B . A subgame-perfect Nash equilibrium is a pair (q_A^*, q_B^*) that constitute a Nash equilibrium for any (w_i, T_i) for $i = A, B$, together with an optimal choice of q_B by the distributor in market B for any (w_B, T_B) and an optimal choice of (w_i, T_i) for $i = A, B$ by the manufacturing firm M . Let w denote the vector (w_A, w_B) and T denote the vector (T_A, T_B) .

Our main objective is to analyze how the manufacturing firm sets the wholesale prices and the transfer payments to maximize its profit. The manufacturing firm's profit is equal to the total revenues in equilibrium minus real costs incurred. More precisely, the objective of the manufacturing firm is to maximize:

$$\Pi(w) = p_A(q_A^* + q_B^*) - tq_B^* + p_Bq_B \tag{1}$$

where the first term on the right hand side is the total revenue in market A , the second term is the real cost of parallel trade between the markets and the third term is the total revenue in market B . Note that with two-part tariffs and complete information the manufacturer is able to extract all economic profits from both distributors.

We assume that at least one distributor can sell in market A , for any w and T accepted by the distributors. The equilibrium quantity for distributor A is¹⁴

¹⁴Details are provided in the appendix.

$$q_A^*(w) = \begin{cases} \frac{1-2w_A^*+w_B^*+t}{3} & \text{if } 2(w_B^*+t)-1 < w_A^* < \frac{1+w_B^*+t}{2} \\ \frac{1-w_A^*}{2} & \text{if } w_A^* \leq 2(w_B^*+t)-1 \\ 0 & \text{if } \frac{1+w_B^*+t}{2} \leq w_A^* \end{cases} \quad (2)$$

and the corresponding quantity for distributor B in market A , i.e. the volume of PI, is

$$q_B^*(w) = \begin{cases} \frac{1+w_A^*-2w_B^*-2t}{3} & \text{if } 2w_A^*-1 < w_B^*+t < \frac{1+w_A^*}{2} \\ \frac{1-w_B^*-t}{2} & \text{if } w_B^*+t \leq 2w_A^*-1 \\ 0 & \text{if } \frac{1+w_A^*}{2} \leq w_B^*+t. \end{cases} \quad (3)$$

and in market B the unique subgame equilibrium is the monopoly quantity

$$q_B(w) = S\left(\frac{1-bw_B}{2}\right) \text{ if } w_B < \frac{1}{b}, \quad (4)$$

where the condition for the wholesale price ensures that the distributor sells a positive quantity in that market. The quantities given in equations(2), (3) and (4) characterize all subgame equilibria and we can proceed to analyze the decision of the manufacturing firm. More precisely, the firm chooses optimal wholesale prices for the distributors.

The optimal choices depend on trade costs. Four possible outcomes emerge as equilibria for different ranges. First, in the range of low trade costs the manufacturing firm is able to set optimal wholesale prices in both markets and PI exist. Second, for intermediate trade costs the firm is constrained to set a non-negative price in A but can choose an optimal price in B and PI occur in equilibrium. Third, for high trade costs the manufactureur sets wholesale prices to deter PI. Finally, sufficiently high trade costs are prohibitive and markets will be segmented.

For both low and intermediate trade costs the manufacturing firm sets wholesale prices to accommodate the effects of arbitrage. Accordingly, parallel imports occur in equilibrium. The quantities chosen by distributors are all positive. The equilibrium profit is denoted $\Pi^{PI}(w)$, where superscript PI refers to "parallel imports".

For high trade costs, on the other hand, parallel trade is deterred and no arbitrage occurs. The margin between the retail price in market A and the wholesale price in market B must be smaller than the unit trade cost, i.e.

$$p_A(w) - w_B \leq t. \quad (5)$$

In this case, the quantity sold in market A by distributor B is zero and we refer to this outcome as the arbitrage-free equilibrium. The profit is denoted $\Pi^{NA}(w)$, where NA refers

to "no arbitrage".

Finally, for prohibitive trade costs each distributor would only sell in its designated market.¹⁵ In the segmented equilibrium the manufacturing firm can maximize its profit without a constraining arbitrage condition. It follows immediately that in the segmented equilibrium the manufacturing firm set wholesale prices equal to zero (marginal production costs) in both markets (i.e. $w_A = 0$ and $w_B = 0$) to eliminate the double mark-up problem and the corresponding segmented retail prices are

$$p_A = \frac{1}{2} \text{ and } p_B = \frac{1}{2b}. \quad (6)$$

We now proceed to analyze the wholesale prices set by the manufacturing firm in the PI equilibrium and the arbitrage-free equilibrium and make a comparison with the segmented equilibrium. We then proceed to analyze consumer welfare

3. OPTIMAL PRICING AND PARALLEL TRADE

3A. Retail Price Divergence

It is natural to start with the analysis in the range of low trade costs. In this situation the distributor in market B finds it profitable to sell a positive quantity in market A , implying that parallel imports occur in equilibrium. The analysis in this section shows that parallel trade does not unambiguously result in retail price convergence.

To find the optimal wholesale prices we differentiate the accommodation profit $\Pi^{PI}(w)$ with respect to wholesale prices in the two markets. The first-order condition with respect to the wholesale price in market A , w_A , is

$$\frac{d\Pi^{PI}(w)}{dw_A} = \left[\frac{1}{9} - \frac{2}{9}w_A - \frac{2}{9}w_B - \frac{2}{9}t \right] - \left[\frac{1}{3}t \right] = 0. \quad (7)$$

and the first-order condition with respect to the wholesale price in market B , w_B , is

$$\frac{d\Pi^{PI}(w)}{dw_B} = \left[\frac{1}{9} - \frac{2}{9}w_A - \frac{2}{9}w_B - \frac{2}{9}t \right] + \left[\frac{2}{3}t \right] - \left[\frac{Sbw_B}{2} \right] = 0. \quad (8)$$

The first-order condition in market A illustrates two of the three problems facing the manufacturer. The first term in square brackets is the pro-competitive effect of parallel imports in that market and it characterizes the manufacturer's incentive to control the

¹⁵More generally, segmentation may also be caused by contractual restrictions and legal barriers.

total supply there. Competition between distributors lowers the manufacturer's profits. Note that the pro-competitive effect is positive when w_B and the unit trade cost are low. In this situation the firm would moderate the pro-competitive effect of arbitrage by raising its wholesale price in market A . In contrast, the effect is negative when w_B and the trade cost are high and the firm would choose to reduce w_A . The incentive to reduce the volume of arbitrage and charge a low wholesale price in market A is therefore stronger when the trade cost is high. The second term is the trade-cost effect, which captures the incentive to save resources wasted in parallel trade. Clearly, the higher is t the greater the incentive to limit PI by reducing the wholesale price in A .

In the first-order condition for market B , the first term in square brackets is the pro-competitive effect of parallel imports in market A . High trade costs permit the firm to reduce w_B and still limit this competition. Note that the pro-competitive effect is identical in both (7) and (8). In other words, the manufacturing firm has two instruments with which to control the total supply in market A : the wholesale price in A and the wholesale price in B . The second term in square brackets also reflects the problem with resources used in trade costs. Here, the trade-cost effect is positive and, to moderate the volume of arbitrage, the manufacturing firm would charge a high wholesale price in market B , while simultaneously charging a low wholesale price in market A .

Note that a third term appears in the latter first-order condition, reflecting the double-markup effect. This is the profit-reducing impact of double marginalization in market B , capturing the inability of the manufacturer to set the efficient wholesale price of zero while simultaneously limiting the volume of parallel trade. The higher is w_B , other things equal, the greater the incentive to reduce that price toward vertical pricing efficiency. If the manufacturing firm is forced by PI to charge a positive wholesale price in B its revenues there would be reduced because the local monopoly distributor there would set a retail price above the optimal (revenue-maximizing) level for the IPR holder.¹⁶ The manufacturing firm thus has an incentive to keep the wholesale price in B low in order to minimize the double-markup problem.

Having analyzed the incentives of the manufacturing firm we can solve the system of equations consisting of the two first-order conditions (7) and (8) to find the optimal wholesale

¹⁶This insight explains why there is no double-markup problem in A , where the distributor must compete with imports and is not a monopolist. Increases in w_A accordingly reflect conditions of competition and generate profit-maximizing retail prices there.

prices. The system has a unique solution:

$$w_A = \frac{1}{2} - \frac{5}{2}t - \frac{2}{Sb}t, \quad (9)$$

$$w_B = \frac{2}{Sb}t, \quad (10)$$

where both prices are continuous and linear functions in the unit trade cost t . The wholesale price in market A (B) is a decreasing (increasing) function of the trade cost. Note that the wholesale price in market A is non-negative for $t \in [0, \underline{t}]$, which we term the range of low trade costs. The upper bound for this interval is determined by $w_A(t) = 0$ and the relevant threshold level is

$$\underline{t} = \frac{Sb}{4 + 5Sb}. \quad (11)$$

Our results support two observations. First, the lower are trade costs, the stronger is the incentive to reduce the pro-competitive effect by raising the wholesale price in A (note that w_A rises as t falls). Moreover, few resources are wasted by PI in this range of costs. Second, declining trade costs permit the firm to decrease the wholesale price in market B and reduce the double-markup distortion. Thus, parallel trade has an effect on vertical price control that is opposite to the typical effect of arbitrage in price-discrimination models. The wholesale price in the importing country increases as the unit trade cost falls, despite an increasing volume of parallel imports. Correspondingly, the wholesale price in the export market decreases when the unit trade cost falls.

Readers may prefer to think in terms of rising trade costs. Thus, put another way, as trade charges increase in this low range the manufacturer would face diminished PI and lower competitive pressure, permitting it profitably to decrease the wholesale price in A . In B the distributor price would rise, which would exacerbate the double-markup problem. Rising trade costs would also increase resources wasted in PI activities.

We insert the optimal wholesale prices into (2) and (3) to obtain the subgame-perfect equilibrium quantities. We then insert the equilibrium outputs into the inverse demand functions to obtain the retail prices in markets A and B :

$$p_A = \frac{1}{2} - \frac{1}{2}t, \quad (12)$$

$$p_B = \frac{1}{2b} + \frac{t}{Sb}, \quad (13)$$

where both functions are continuous and linear in trade costs. Thus, in the range of low

trade costs a decline (increase) in t generates a divergence (convergence) in retail prices.¹⁷ We can summarize our result for the range of low trade costs as follows:

Proposition 1 *Assume that retail markets are segmented and the unit trade cost of parallel trade is $t \in [0, \underline{t}]$. If the trade cost increases in this range, the retail price in market A decreases, the retail price in market B increases and the volume of PI decreases.*

Proof. The proof of the first part of the proposition follows immediately from the optimal retail prices in (12) and (13). The volume of arbitrage is equivalent to the quantity chosen by distributor B in market A . We insert the equilibrium wholesale prices to obtain the equilibrium volume of parallel imports:

$$q_B^* = \frac{1}{2} - \frac{2}{Sb}t - \frac{3}{2}t \text{ if } t \in [0, \underline{t}] \quad (14)$$

which is a continuous and declining function in the trade cost. ■

It is also worth noting that the volume of arbitrage is an increasing function in the size parameter S as well as the price-elasticity (captured by parameter b) in market B . The manufacturing firm has an incentive to keep the wholesale price in that market low when it is large and when the consumers are price-sensitive. In this case the manufacturing firm is willing to accept a larger volume of arbitrage to avoid a serious double-markup problem in B .

One interesting aspect of the parallel trade equilibrium is that at a unit trade cost of zero the retail prices are identical to those that would be set by a vertically integrated monopolist. In other words, the manufacturing firm can solve the vertical control problem perfectly if no real resources are wasted in arbitrage and consumer markets are segmented.¹⁸ The wholesale price in market B would be set to zero and the wholesale price in market A would be set at a prohibitive level for the distributor there, pushing it out of business. Distributor B would sell the revenue-maximizing quantity in both markets and A would be supplied from B but no real resources would be used in transportation. In addition, there would be no double-markup problem in market B . Consequently, the retail prices in an integrated equilibrium would be identical to the prices set in a completely segmented equilibrium.

¹⁷Strictly speaking, because these prices depend on the values of S and b , it is possible to have $p_b > p_a$ at \underline{t} , in which case as t falls these prices would come together before diverging as in the text. This situation is consistent with PI flowing from the high-retail price nation to the low-retail price nation, a result found in Chen and Maskus (2005).

¹⁸In principle, the manufacturer could solve the problem even with positive trade costs by vertically integrating with the distributor in B and not shipping any goods to A . As noted earlier, we rule this case out in order to focus on the situation of independent distributors and the vertical control problem. Vertical integration is not feasible in a variety of circumstances and independent wholesalers are common.

This result stems from the fact that as trade costs disappear it is profit-maximizing to shift the source of supply from the distributor in A to the distributor in B . Once distributor B becomes the sole supplier at zero trade cost, the manufacturer would choose to authorize exports from B to A , implying that PI would be replaced by standard trade. In effect, the result would be a market integrated at the wholesale level but segmented at the retail level by a single distributor with two territories.

3B. A Pro-competitive Effect of Parallel Imports

For sufficiently high trade costs, i.e. a unit cost above the critical threshold \underline{t} , the optimal wholesale price in market A would be negative, meaning that the manufacturing firm would like to sponsor the supply of distributor A in that market through a variable subsidy and recoup this subsidy by charging a higher fixed fee. As discussed earlier we have ruled out this possibility on grounds that effective monitoring may be impossible. Therefore, the firm must determine the profit-maximizing wholesale price in market B subject to the condition that the wholesale price in market A is zero.

This constraint adds an additional distortion and further complicates the vertical control problem by limiting the manufacturer to a single price instrument. The maximization problem is now reduced to finding the solution to the first-order condition for the manufacturing firm with respect to the wholesale price in market B :

$$\frac{d\Pi^{PI}(w)}{dw_B} = \left[\frac{1}{9} - \frac{2}{9}w_B - \frac{2}{9}t \right] + \left[\frac{2}{3}t \right] - \left[\frac{1}{2}Sbw_B \right] = 0. \quad (15)$$

As before, the first term reflects the pro-competitive effect in A , the second term is the trade-cost effect and the third term reflects the double-markup problem in B .

The important difference compared to the first-order condition when the manufacturing firm could charge an optimal wholesale price in market A resides in the first term. The manufacturing firm has an incentive to charge a lower wholesale price in market B to avoid generating a double-markup problem in market A .¹⁹ Hence, the wholesale price in B is lower than it would be if w_A could be negative. We solve the first order condition (15) to obtain the equilibrium wholesale price

$$w_B = 2 \left(\frac{1 + 4t}{4 + 9Sb} \right). \quad (16)$$

¹⁹In contrast to the unconstrained case, the possibility of a double-markup problem in A arises because of the minimum pricing constraint, which might prevent establishment of the optimal retail price.

In this case, as trade cost goes up the manufacturer would reduce the volume of PI by raising the wholesale price in market B . However, it is not profitable for the manufacturing firm to block PI completely by this strategy because it would create an excessively large double-markup effect in B . The higher the trade cost, the stronger is the incentive to reduce the volume of PI. This incentive is a combination of two effects: a higher trade cost wastes more resources in PI and also protects market A from competition. These effects reduce the net cost of limiting PI with an inefficient wholesale price in market B . Accordingly, the wholesale price in B increases with t while the wholesale price in A would be kept at the minimum level of zero. As previously we insert the subgame-perfect equilibrium quantities from (2) and (3) in the inverse demand functions to obtain the retail prices in markets A and B . The equilibrium prices are given by

$$p_A = \frac{1}{3} + \frac{2}{3} \left(\frac{1+4t}{4+9Sb} \right) + \frac{t}{3}, \quad (17)$$

$$p_B = \frac{1}{2b} + \left(\frac{1+4t}{4+9Sb} \right), \quad (18)$$

where both functions are continuous and linear in t . It is interesting to note that both retail prices would increase with higher trade costs.

For sufficiently high trade costs, parallel trade is unprofitable and no arbitrage occurs, whereas for low and intermediate trade costs, i.e. $t \in [0, \bar{t})$, the volume of parallel trade is positive. The critical threshold for eliminating arbitrage is given by

$$p_A - w_B = t, \quad (19)$$

where p_A is given by (17) and w_B is given by (16). Accordingly, the upper threshold at which PI vanish is

$$\bar{t} = \frac{3}{2} \left(\frac{Sb}{4+3Sb} \right) \quad (20)$$

which takes values on the open interval $(0, 1/2)$. The threshold level is consequently strictly positive and less than the prohibitive unit trade cost ($t = 1/2$) for relevant values of S and b . We have the following result.

Proposition 2 *Assume that retail markets are segmented and the unit trade cost of parallel trade is $t \in (\underline{t}, \bar{t})$. If the trade cost increases in this range, the retail prices in market A and B both increase and the volume of PI decreases.*

Proof. For the purpose of proving the proposition, note that the optimal retail prices (17)

and (18) are linear and increasing functions in t . Second, insert the equilibrium wholesale prices to obtain the equilibrium volumes of parallel imports:

$$q_B^* = \frac{1}{3} - \frac{4}{3} \left(\frac{1 + 4t}{4 + 9Sb} \right) - \frac{2t}{3} \text{ if } t \in (\underline{t}, \bar{t}) \quad (21)$$

which is a continuous and declining function of trade costs. ■

The evolution of the retail price differential between markets A and B is also interesting in the intermediate range of trade costs. Both retail prices rise with costs in this interval, but they actually diverge. Using the equilibrium prices (17) and (18) we compute the difference between retail prices, i.e. $p_A - p_B$. The derivative of the difference with respect to the unit trade cost is

$$\frac{d(p_A - p_B)}{dt} = \frac{3Sb}{4 + 9Sb}, \quad (22)$$

which is strictly positive. In other words, retail prices differ more for high trade costs than for low trade costs in this range. This result is the outcome of the manufacturing firm's incentive to set a wholesale price in market B to reduce the volume of PI while keeping the wholesale price in market A at zero. Put differently, suppose unit trade cost declines in this range. The manufacturer would accommodate a larger volume of PI while also choosing a lower wholesale price in market B . As a result, the pro-competitive effect of PI in market A gets stronger and the double-markup effect in market B gets weaker. The retail prices in both markets consequently decrease and tend to converge as trade costs fall.

The essence of the analysis in this subsection is that, because the manufacturer no longer has the ability to set the wholesale price in A optimally, it is limited to using only the wholesale price in B to manage these various tradeoffs. Accordingly, the firm must permit a higher PI volume, magnifying the pro-competitive impact in market A .

3C. A Double-Markup Problem with Arbitrage-Free Prices

For high trade costs, i.e. $t \in [\bar{t}, 1/2]$, the manufacturing firm sets wholesale prices to block PI and the equilibrium is, consequently, arbitrage-free. In this equilibrium the distributor in B finds it profitable to sell a positive quantity in her own market but does not ship goods to market A .

Noting that the wholesale price in A must be non-negative, i.e. $w_A \geq 0$, we differentiate the arbitrage-free profit function $\Pi^{NA}(w)$ with respect to wholesale prices in the two markets, subject to the condition in inequality (5) to find the optimal wholesale prices. These

prices in the arbitrage-free equilibrium are

$$w_A = 0, \tag{23}$$

$$w_B = \frac{1}{2} - t. \tag{24}$$

This result has a straightforward interpretation. The higher the trade cost, the more protected is market A by the natural barrier and, further, the wholesale price in market B can be lowered to reduce the double-markup effect in B . This price would be set both to deter PI and limit the price distortion in B . When t reaches its maximum level (i.e. $t = 1/2$), the optimal wholesale price in B becomes zero and the price distortion disappears. The corresponding equilibrium retail prices in A and B are given by

$$p_A = \frac{1}{2}, \tag{25}$$

$$p_B = \frac{1}{2b} + \frac{1}{4} - \frac{t}{2}. \tag{26}$$

We have the following result:

Proposition 3 *Assume that retail markets are segmented and the unit trade cost of parallel trade is $t \in (\bar{t}, 1/2)$. In this range of trade costs, the retail price in market A is equivalent to its value in the segmented equilibrium. Moreover, as trade cost increases, the retail price in market B declines.*

Proof. The proposition follows from (25) and (26). ■

The retail price in B declines with trade costs because the manufacturing firm sets a wholesale price in this market to deter PI. This wholesale price reflects a double-markup distortion that diminishes with higher t . Thus, for higher trade costs the manufacturing firm would eliminate PI with a lower wholesale price and experience a correspondingly weaker double-marginalization problem. In this case equilibrium retail prices diverge as trade cost increases, which is more in line with the usual intuition. The potential competition from arbitrage is strong when the trade cost is low and weak when the trade cost is high.

Note that $p_B = \frac{1}{2b}$ at $t = 1/2$. For prohibitive trade costs, i.e. $t > 1/2$, markets are fully segmented.

4. PRICES, PROFITS AND CONSUMER WELFARE

The analysis of prices in the previous section characterizes the equilibrium for all trade costs and other parameter values. In this section we take a closer look at retail prices,

profits and consumer welfare. We are particularly interested in comparing the equilibrium in a market subject to actual or potential PI competition with the segmented equilibrium.

4A. Retail Prices

We first consider retail prices for different trade costs. Equilibrium retail prices in the two markets are illustrated in Figure 1 for parameter values $S = 1$, $b = 2$. The upper line corresponds to p_A and the lower line corresponds to p_B . The horizontal lines indicate fully segmented price levels. The first vertical line demarcates the lower threshold, below which the non-negativity constraint does not bind and retail prices diverge with a decreasing trade cost. This range corresponds to accommodation equilibrium conditions and PI are positive. Proceeding to intermediate trade costs, prices increase but grow apart as t increases in the second range, which corresponds also to positive volumes of PI. The second vertical line demarcates the upper threshold of t for which PI exist. In the third range, corresponding to the arbitrage-free case, retail price is at its monopoly level in market A but falls in market B . Above that range of trade cost the markets become fully segmented.

It is straightforward to see that the arbitrage-free equilibrium is Pareto-dominated by a completely segmented equilibrium. This follows directly from the fact that the retail price in market B is higher than the segmented retail price (because of the double-markup problem), the retail price in market A equals the segmented retail price and the profit of the manufacturing firm is lower than it would be under full segmentation.

4B. Profits

Next, we compute the manufacturing firm's profit in equilibrium (see the appendix for details). It is illustrated as a function of the variable trade cost in Figure 2 for the same parameter values. For low trade costs ($t < \underline{t}$), the profit decreases as trade cost rises, both because of more resources lost in parallel trade activities and an increasing inability to induce optimal retail prices in both markets. For intermediate trade costs ($\underline{t} \leq t < \bar{t}$), the profit goes up as trade costs increase. The importance of the trade cost as a natural barrier increases; parallel imports vanish, fewer resources are wasted and the double-markup problem in market B is moderated. For high trade costs ($t \geq \bar{t}$), the profit also increases with t . The equilibrium is arbitrage-free and no resources are wasted in trade. The double-marginalization problem in the export market is gradually reduced as the markets move towards segmentation.

4C. Consumer Welfare

Finally, we analyze consumer welfare with PI and exclusive territories. Both competition policy and market integration policy typically have the objective of enhancing consumer welfare and maximizing consumer surplus. It is therefore of interest to evaluate the effects of legalizing PI on consumer welfare when the manufacturing firm employs two-part tariffs for vertical control purposes. We compare these effects to the welfare outcome when the manufacturing firm is permitted to segment the wholesale market with exclusive territories and discriminate with different prices in the two countries.

Aggregate (two-country) consumer surplus with an integrated wholesale market is

$$CS = \left[\left(Q_A - \frac{Q_A^2}{2} \right) - p_A Q_A \right] + \left[\frac{1}{b} \left(Q_B - \frac{Q_B^2}{2S} \right) - p_B Q_B \right] \quad (27)$$

where the first term is consumer surplus in A and the second term is consumer surplus in B . We insert the demand functions in markets A and B to get consumer surplus as a function of retail prices:

$$CS(p) = \frac{1}{2} (1 - p_A)^2 + \frac{S}{2b} (1 - bp_B)^2 \quad (28)$$

which is continuous, concave and decreasing in both p_A and p_B .

Differentiating consumer surplus with respect to t in the accommodation range of low trade costs ($t < \underline{t}$) gives the following result

$$\frac{dCS}{dt} = -\frac{1}{2b} + \frac{1}{4} + \frac{t}{4Sb} + \frac{t}{4}, \quad (29)$$

and in the accommodation range of intermediate trade costs ($\underline{t} \leq t < \bar{t}$),

$$\frac{dCS}{dt} = -\frac{2S(b+1)+2}{4+9Sb} + \frac{(4+Sb)t}{4+9Sb} \quad (30)$$

and, finally, in the arbitrage-free range of high trade costs ($t \geq \bar{t}$):

$$\frac{dCS}{dt} = \frac{2S - Sb + 2Sbt}{8}. \quad (31)$$

It is straightforward to show that combined consumer surplus in markets A and B has its unique global minimum at $t = \bar{t}$. This has an intuitive explanation. In any arbitrage-free equilibrium, the retail price in market A is at its maximum and identical to the segmented price. At the same time, the retail price in B has its maximum at \bar{t} and is strictly higher

than the segmented price. Here, the manufacturing firm charges the highest wholesale price in that market to ensure that prices are arbitrage-free, since the trade cost is only a partial barrier for parallel imports. In other words, for high trade costs both combined consumer surplus and producer profits are lower when parallel trade is permitted than when the manufacturing firm can segment the wholesale market with exclusive territories.

In addition, for two sufficiently similar markets, i.e. $b < \underline{b}(S)$, the policy of permitting parallel trade has a negative effect on aggregated consumer surplus at low and intermediate trade costs. The unique potential local maximum for consumer welfare at interior trade costs, $t \in (0, 1/2)$, is at the lower critical threshold \underline{t} . Accordingly, market segmentation dominates PI as long as consumer welfare at this trade cost is lower than the consumer welfare in a segmented equilibrium. We insert equilibrium prices from the PI equilibrium at \underline{t} and the segmented equilibrium into the consumer surplus function. We then set the two levels equal and solve for b . The critical level is

$$\underline{b}(S) = \frac{10}{11} - \frac{6}{11S} + \frac{2}{11S} \sqrt{9 + 14S + 25S^2} \quad (32)$$

and we have the following result.

Proposition 4 *For sufficiently similar markets, i.e. $b < \underline{b}$, joint consumer surplus has its global maxima at $t = 0$ and $t = \frac{1}{2}$ (see Figure 3).*

Proof. First, note that retail prices are equal to the segmented prices at $t = 0$ and $t = \frac{1}{2}$. Consumer surplus is consequently the same at these two points. Second, consumer surplus in the arbitrage-free range of high trade costs ($t \geq \bar{t}$) is strictly less than consumer surplus in the segmented equilibrium as p_A is identical to the segmented price and p_B is strictly higher. Third, in the accommodation range of intermediate trade costs ($\underline{t} \leq t < \bar{t}$), the derivative of the consumer surplus function is strictly negative and consumer surplus reaches its maximum at \underline{t} in this interval (both retail prices increase in t in this range). Finally, consumer surplus in a segmented equilibrium is constant and consumer surplus in the range of low trade costs is quadratic and strictly convex in t (the derivative is linear and increasing in t). The two consumer surplus levels, consequently, can only be equal at two trade cost levels and consumer surplus must be lower in any interior PI equilibrium than in the segmented equilibrium. The first trade cost is $t = 0$. The second is

$$t = \frac{S(4 - 2b)}{Sb + 4} \quad (33)$$

and this threshold is larger than \underline{t} if

$$b < \frac{10}{11} - \frac{6}{11S} + \frac{2}{11S} \sqrt{9 + 14S + 25S^2} \quad (34)$$

which is sufficient for the consumer surplus to be lower in the PI equilibrium than in the segmented equilibrium. Consumer surplus is consequently lower in an equilibrium open to PI than in a segmented equilibrium for any $t \in (0, 1/2)$, if $b < \underline{b}$. ■

While open parallel trade is beneficial for consumers in the import market for low and intermediate trade costs, it is not in the interest of the consumers of both countries taken as a group and it also damages producers. On the contrary, if the manufacturing firm uses two-part tariffs to solve vertical control problems, exclusive distribution territories can benefit consumers.

On the other hand, for sufficiently different countries parallel trade may increase consumer welfare (see Figure 4). In the range of low trade costs consumer surplus is an increasing function in t for $b \geq 2$. In this case, combined consumer surplus has its maximum at \underline{t} . Consumers can jointly benefit when the manufacturing firm sets prices to minimize the volume of parallel trade. The difference in retail prices reaches its minimum at \underline{t} , where the aggregated consumer surplus is highest. This is akin to the classical welfare effect of uniform pricing: price convergence increases aggregate consumer surplus as long as total output is unchanged or higher compared to a situation with price discrimination.

This result shows that consumers in both markets may jointly benefit from parallel imports for low and intermediate trade costs as long as the difference in the price elasticity of demand is sufficiently large between the two consumer groups. The intuition for this result is that the pro-competitive effect of PI in A would dominate the double-markup effect in B since the optimal price in the latter market would be sufficiently low to induce a significant price reduction in the former market. This outcome would pertain to situations where consumers in the PI-source country were substantially more price-sensitive than in the PI-recipient country. Indeed, licensee prices of original manufacturer's goods generally seem to follow this pattern across countries, at least to the degree that per-capita income differences are negatively correlated with demand elasticities.²⁰ In that regard, where trade costs are low permitting PI among countries at significantly different income levels could raise overall consumer welfare, even as it would harm consumers in the lower-income countries.

As a final observation in this section we note that the two most important parameters for reaching welfare conclusions are b and t , which are the price sensitivity in market B

²⁰Ganslandt and Maskus (2004) found this for pharmaceuticals. See also NERA (1999).

and the trade cost, respectively. Parameter S , the size of market B , is of minor importance. Our main conclusion, that permitting parallel trade can reduce combined consumer surplus for positive trade costs, is true independent of the size of market B as long as the price-sensitivity in that market is below the limit of the critical threshold as S goes to zero, which requires only that $(b < 4/3)$.²¹

5. MODEL EXTENSIONS

In this section we consider two important relaxations of our model assumptions: the possibility of arbitrage at the retail level and the existence of quality-differentiated products.

5A. Retail Arbitrage

We initially assumed that retail markets are segmented, justifying the realism of this assumption on the grounds of varying policies, product standards, and differential market sizes. While we think these factors are relevant in a broad class of goods, readers may wonder about the implications of retail arbitrage in the model. Thus, here we relax the assumption of retail segmentation.

First, consider the possibility that consumers could engage in perfectly elastic retail arbitrage. Consumer arbitrage is not profitable if the retail price differential is less than the unit trade cost t . Formally, the no-arbitrage condition is

$$p_A - p_B \leq t \tag{35}$$

and we insert the retail prices and solve for the critical trade cost. A sufficient condition for retail markets to be segmented in this case is $t \geq \tilde{t}$, where

$$\tilde{t} = \frac{S(b-1)}{2+3Sb}. \tag{36}$$

Note that this threshold is lower than \underline{t} meaning that this form of arbitrage limits the range of price divergence, for given market parameters, more than does the first form. Nevertheless, a range of price divergence continues as trade costs fall. Observe that the threshold is close to zero for b close to 1. Markets remain segmented at the retail level by the natural barrier t for similar markets or high trade costs. For dissimilar markets and low trade costs, arbitrage at the retail level could limit the scope for price differentiation and

²¹We demonstrate this assertion in a separate appendix available on request.

prices would converge. Figure 5 illustrates the consumer-level retail arbitrage conditions. The solid lines depict retail prices in markets A and B . The dotted vertical line corresponds to the threshold level \tilde{t} at which consumer arbitrage would kick in. For trade costs below that level prices would tend to converge along the dashed lines making up the retail price cone (p'_A and p'_B), as shown.

Next, consider the possibility that the retail services provided in market A are necessary complements for consumption there. In other words, the distributor in market A (or an agent) can import the product from B but consumers in A cannot buy the product at the retail level in B . Arbitrage by distributor A (in which he buys at the retail level abroad in order to find a cheaper source of supply than the available wholesale price) is not profitable if the margin between the wholesale price in market A and the retail price in market B is lower than the unit trade cost. Formally, the no-arbitrage condition in this case is

$$w_A - p_B \leq t \tag{37}$$

and in order to determine whether this condition is slack in equilibrium we insert the accommodation wholesale price in market A and the retail price in market B . A sufficient condition for retail markets to be segmented in this case is

$$t > \frac{S(b-1)}{6+7Sb} \tag{38}$$

where the right hand side is close to zero for b close to 1. Markets are consequently segmented at the retail level by the natural barrier t for similar markets or high trade costs. For dissimilar markets and low trade costs, however, arbitrage at the retail level could limit the scope for price differentiation and prices would converge.

We reiterate that perfect retail arbitrage is a strong assumption and inconsistent both with the fact that the bulk of PI occurs at the distributor level and with persistent differences in retail prices within the EU. There are good reasons to expect limited arbitrage of this kind even without restrictions on parallel trade. First, as noted above there are likely to be complementarities in retail services that cannot be provided by arbitrageurs. Second, there may be significant fixed costs in organizing cross-border retail trade. Third, there may be large information costs for consumers in determining product prices and characteristics for purposes of organizing arbitrage. Thus, we think our analysis of distributor-level PI is valid in many realistic circumstances.

5B. Product Differentiation

In the previous sections we modeled parallel imports and domestic products as homogeneous goods. In some situations, however, it is more realistic to assume that the PI product is differentiated from the good sold by the local distributor. More specifically, the PI variety can be of slightly inferior quality, perhaps due to the absence of pre-sale and post-sale services, limited warranties, longer distribution lags or a need for domestic adaptation. In this subsection we explore the effects of quality differentiation within our framework of one-way parallel trade. We restrict our attention to the range of low trade costs, with a strictly positive volume of PI, since differentiation in the absence of PI competition is not interesting. Further, we consider both quantity competition and price competition.

Quantity competition.—

Assume that demand in market B is given by $Q_b = S(1 - p_b)$ as before, where p_b denotes the price of the product sold there. Next, consider market A . Because the PI good is inferior, it offers utility at a fraction $\lambda < 1$, of the product sold by the local distributor.²² The demand for the product sold by the local distributor in A and the demand for the PI product are

$$q_A = 1 - \frac{p_A - p_{PI}}{1 - \lambda} \quad (39)$$

$$q_{PI} = \frac{\lambda p_A - p_{PI}}{(1 - \lambda)\lambda} \quad (40)$$

where p_A is the price of the variety sold by the local distributor and p_{PI} is the price of the PI variety. We solve the equilibrium in the usual backward fashion. For a sufficiently low trade cost, the equilibrium quantities in market A are

$$q_A(w) = \frac{2 - \lambda + w_B + t - 2w_A}{4 - \lambda} \quad (41)$$

$$q_{PI}(w) = \frac{\lambda + w_A\lambda - 2w_B - 2t}{\lambda(4 - \lambda)} \quad (42)$$

while the subgame in market B , for a given w_B , is the same as before. Working backwards, the manufacturing firm chooses wholesale prices to maximize aggregated profit. The

²²We treat the parameter λ as fixed and exogenous. For some products and markets product differentiation may be endogenous. For example, the manufacturing firm could increase product differentiation by local adaptation or limits on the geographical scope of warranties. This is an interesting topic for future research.

optimization problem has a unique solution in which equilibrium wholesale prices are

$$w_A = \frac{S\lambda^3b - t(4S\lambda b + 4 + S\lambda^2b)}{2(4(1 - \lambda) + Sb\lambda(4 - 3\lambda))} \quad (43)$$

$$w_B = \frac{2(1 - \lambda)\lambda + t\lambda}{4(1 - \lambda) + Sb\lambda(4 - 3\lambda)} \quad (44)$$

It follows from these expressions that the wholesale price in A decreases and that in B increases in t . This is qualitatively the same result we found with homogeneous products. We insert the equilibrium wholesale prices in the demand functions to obtain the following retail prices

$$p_A = \frac{4(1 - \lambda) + Sb\lambda(2 - \lambda)^2 - tSb\lambda^2}{2(4(1 - \lambda) + Sb\lambda(4 - 3\lambda))} \quad (45)$$

$$p_{PI} = \frac{4\lambda(1 - \lambda) + Sb\lambda^2(2 - \lambda) + (4(1 - \lambda) + Sb\lambda(4 - 5\lambda))t}{2(4(1 - \lambda) + Sb\lambda(4 - 3\lambda))} \quad (46)$$

$$p_B = \frac{1}{2b} + \frac{\lambda(1 - \lambda) + t\lambda}{4(1 - \lambda) + Sb\lambda(4 - 3\lambda)} \quad (47)$$

where the denominators are strictly positive for $\lambda < 1$.

It follows that the retail price for the locally distributed product in A decreases and the retail price in B increases, if the trade cost increases. This result corresponds qualitatively to our result for homogeneous products.

Interestingly, the result for the PI variety depends on the degree of product differentiation. For sufficiently similar products (high λ), the retail price for the PI product decreases in t . If the PI product is a perfect substitute for the local variety both products must sell for the same price. However, if the PI product is very inferior compared to the local variety, the retail price for the PI product increases in t since the distributor can pass on the higher marginal cost to consumers.

Price competition.—

Assuming that products are differentiated we can analyze the equilibrium when distributors set prices rather than quantities. Let both distributors choose profit-maximizing prices in market A while distributor B also sets the price in her market. The optimal retail prices as functions of wholesale prices are

$$p_A(w) = \frac{2 - 2\lambda + 2w_A + w_b + t}{4 - \lambda} \quad (48)$$

$$p_{PI}(w) = \frac{\lambda(1 - \lambda) + 2w_B + w_A\lambda + 2t}{4 - \lambda} \quad (49)$$

and the volume of PI is

$$q_{PI}(w) = \frac{\lambda(1 - \lambda) + w_A\lambda - (2 - \lambda)(w_B + t)}{(4 - \lambda)(1 - \lambda)\lambda} \quad (50)$$

In other words, if the wholesale price in A is sufficiently high and that in B is sufficiently low, parallel trade is strictly positive for low unit trade costs. Working backwards, we insert the retail prices in the demand functions and solve the manufacturing firm's optimization problem to obtain the unique equilibrium wholesale prices

$$w_A = \frac{2(S\lambda b + 1)\lambda - 2S\lambda^3 b}{4 - 3S\lambda^2 b + 4S\lambda b} - \frac{1}{2}t \quad (51)$$

$$w_B = \frac{2\lambda}{4 - 3S\lambda^2 b + 4S\lambda b} \quad (52)$$

and it follows that the wholesale price for distributor A decreases in t , while the wholesale price for distributor B is constant. The result is qualitatively similar, but not identical, to quantity competition. The manufacturing firm has three major problems to solve in both cases. It has to moderate the pro-competitive effect in market A , reduce the waste of resources and avoid a double-markup problem in market B . The difference between quantity and price competition is that the problem of resource waste in PI activities is less severe in the latter case. The trade cost also has a more direct effect on competition in market A . In addition, product differentiation has a positive value in market A , which adds a fourth effect of legal PI.

We can use the equilibrium wholesale prices to obtain the corresponding equilibrium retail prices.

$$p_A = \frac{2(1 - S\lambda^2 b + S\lambda b)}{4 - 3S\lambda^2 b + 4S\lambda b} \quad (53)$$

$$p_{PI} = \frac{\lambda(2 - S\lambda^2 b + S\lambda b)}{4 - 3S\lambda^2 b + 4S\lambda b} + \frac{1}{2}t \quad (54)$$

$$p_B = \frac{1}{2b} + \frac{\lambda}{4 - 3S\lambda^2b + 4S\lambda b} \quad (55)$$

and the equilibrium PI quantity is

$$q_{PI} = \frac{S\lambda b}{4 - 3S\lambda^2b + 4S\lambda b} - \frac{t}{2(1 - \lambda)\lambda} \quad (56)$$

If distributors compete in prices, retail prices for the locally distributed products remain constant if the unit trade cost is below the critical level establishing a positive volume of PI. Within this range, a higher unit trade cost would generate a lower PI volume and less competition in the recipient market, but retail prices would not diverge as conventional intuition would suggest. Instead, these prices would be unaffected by changes in t . The retail price for the PI variety in market A , however, would be increased by a rise in the trade cost, in line with conventional intuition.

The main difference between quantity and price competition is that retail prices of the locally distributed goods have a tendency to converge with quantity competition if the trade cost increases within the low range, while the price differential is constant in the case of price competition. The main intuition for this difference is that in the former case the manufacturing firm must push retail prices together to avoid wasteful PI activities, while the competitive disadvantage of a higher trade cost would be sufficient to reduce the volume of PI in the latter case.

6. CONCLUDING REMARKS

We developed a model in which a manufacturing firm owns an intellectual property right in two markets but its ability to limit parallel imports from one market to the other is exhausted. In this environment, the firm has the ability to set differential wholesale prices to its independent distributors in the two locations. It will use these instruments to maximize profits within the vertical-control framework. There are three essential tradeoffs for the manufacturer. It wishes to restrict the extent of competition from PI in the A market, limit the amount of PI because it wastes real resources in transport costs, and avoid the double-markup problem in market B arising from the inability to set an efficient (zero) wholesale price.

Our analysis turned up some interesting results. Because of the cross-cutting effects of PI on wholesale prices in the two markets, it is possible to observe a divergence in wholesale prices as trade costs are reduced within a low range. As a result, retail prices may diverge

as well. Arguably, the EU is in a situation of low and declining internal trade restrictions. In this context, paradoxically, the policy of free parallel trade among member states may be a force for retail price divergence.

For intermediate and high trade costs, the firm might wish to set a negative wholesale price in A but could be constrained to a minimum price of zero. As trade costs increase, the volume of PI declines and the double-markup problem in B diminishes. Retail prices rise in both locations as transport costs increase within the intermediate range, then diverge at high trade costs as these prices achieve levels expected with wholesale-market segmentation.

Perhaps the most interesting implication of our analysis is that there is a substantial difference between integrating retail markets and wholesale markets. In the former case, as trade costs fall due to declining trade barriers and transport costs, straightforward arbitrage would push markets toward price convergence, which is the intuitive result and the outcome anticipated by integration policy. However, in the latter case, where PI are prevalent, declining trade costs could integrate wholesale markets even as they push retail markets toward greater segmentation. Again, this possibility suggests that the favorable view of parallel imports in EU competition policy, stemming from a tradition that does not consider vertical distribution arrangements, may be due for reconsideration.

It could be argued that the double-markup problem in the export market is partly due to seller concentration at the retail level. The manufacturing firm perhaps could introduce several competing retailers to reduce the retail margin and moderate the double-marginalization problem. This benefit of having more than one retail distributor could, however, be offset by other disadvantages such as free-riding and duplicated fixed costs in retailing. Hence, extending the analysis to study the effects of parallel trade with an endogenous retail structure is an interesting subject for future research.

REFERENCES

- [1] Chen, Y. and K. E. Maskus (2005), "Vertical Pricing and Parallel Imports," *Journal of International Trade and Economic Development* 14: 1-18.
- [2] European Central Bank (2002), "Price Level Convergence and Competition in the Euro Area," *Monthly Bulletin*, August: 39-50.
- [3] Friberg, R. and K. Martensen (2001), "Endogenous Market Segmentation and the Law of One Price," Stockholm School of Economics, SSE/EFI Working Paper No. 471.
- [4] Ganslandt, M. and K. E. Maskus (2003), "Vertical Restraints, Distribution, and the Price Impact of Parallel Imports: Implications for the European Union and Sweden," in K. Lundvall, ed., *High Prices in Sweden - A Result of Poor Competition?* (Stockholm: Swedish Competition Authority), 160-223.
- [5] Ganslandt, M. and K. E. Maskus (2004), "The Price Impact of Parallel Trade in Pharmaceuticals: Evidence from the European Union," *Journal of Health Economics* 23: 1035-1057.
- [6] Hilke, J. C. (1988), "Free Trading or Free Riding: an Examination of the Theories and Available Evidence on Gray Market Imports," *World Competition* 32: 75-92.
- [7] Malueg, D. A. and M. Schwartz (1994), "Parallel Imports, Demand Dispersion, and International Price Discrimination," *Journal of International Economics* 37: 187-196.
- [8] Maskus, K. E. (2000), "Parallel Imports," *The World Economy* 23: 1269-1284.
- [9] Maskus, K. E. and Y. Chen (2002), "Parallel Imports in a Model of Vertical Distribution: Theory, Evidence, and Policy," *Pacific Economic Review* 7: 319-334.
- [10] National Economic Research Associates (1999), *The Economic Consequences of the Choice of Regime in the Area of Trademarks* (London: National Economic Research Associates).
- [11] Raff, H. and N. Schmitt (2005a), "Endogenous Vertical Restraints in International Trade," *European Economic Review* 49: 1877-1889.
- [12] Raff, H. and N. Schmitt (2005b), "Exclusive Dealing and Common Agency in International Markets", CES-IFO working paper no. 1168.

- [13] Richardson, M. (2002), "An Elementary Proposition Concerning Parallel Imports," *Journal of International Economics* 56: 233-245.
- [14] Tarr, D. G. (1985), "An Economic Analysis of Gray Market Imports," U.S. Federal Trade Commission, Manuscript

APPENDIX

The subgame equilibrium

It is assumed that at least one distributor finds it profitable to supply market A. Accordingly, we have to consider three different cases: both distributors supply that market, only distributor A supplies it or, finally, only distributor B does so. Consider duopoly first, where the distributors compete in quantities in market A. The profit functions of the two distributors are

$$\pi_A^* = (1 - (q_A^* + q_B^*) - w_A^*) q_A^*, \quad (57)$$

$$\pi_B^* = (1 - (q_A^* + q_B^*) - w_B^* - t) q_B^* \quad (58)$$

and the corresponding first-order conditions are

$$1 - 2q_A^* - q_B^* - w_A^* = 0, \quad (59)$$

$$1 - q_A^* - 2q_B^* - w_B^* - t = 0, \quad (60)$$

and the unique solution to this system of equations is

$$q_A^* = \frac{1 - 2w_A^* + w_B^* + t}{3}, \quad (61)$$

$$q_B^* = \frac{1 + w_A^* - 2w_B^* - 2t}{3}. \quad (62)$$

Distributor A's quantity is positive if

$$w_A^* < \frac{1 + w_B^* + t}{2} \quad (63)$$

and distributor B's quantity is positive if

$$w_B^* < \frac{1 + w_A^* - 2t}{2}. \quad (64)$$

These two conditions must simultaneously hold in duopoly. We combine them and rewrite the expression to obtain

$$2w_B^* + 2t - 1 < w_A^* < \frac{1 + w_B^* + t}{2} \quad (65)$$

and

$$2w_A^* - t - 1 < w_B^* < \frac{1 + w_A^* - 2t}{2}. \quad (66)$$

Next, consider a deterrent wholesale cost for distributor B,

$$w_B^* \geq \frac{1 + w_A^* - 2t}{2} \iff w_A^* \leq 2(w_B^* + t) - 1,$$

for which $q_B^* = 0$ in equilibrium. Then distributor A is the only supplier in his market. It maximizes

$$\pi_A^* = (1 - q_A^* - w_A^*) q_A^* \quad (67)$$

and the first-order condition is

$$1 - 2q_A^* - w_A^* = 0 \quad (68)$$

with a unique solution

$$q_A^* = \frac{1 - w_A^*}{2} \quad (69)$$

which is positive for $w_A^* < 1$. Correspondingly, if the wholesale cost for distributor A is a deterrent, i.e.

$$w_A^* \geq \frac{1 + w_B^* + t}{2} \iff w_B^* + t \leq 2w_A^* - 1, \quad (70)$$

then $q_A^* = 0$ in equilibrium and distributor B is the only supplier in A. It maximizes

$$\pi_B^* = (1 - q_B^* - w_B^* - t) q_B^* \quad (71)$$

and the first-order condition is

$$1 - 2q_B^* - w_B^* - t = 0 \quad (72)$$

with a unique solution

$$q_B^* = \frac{1 - w_B^* - t}{2} \quad (73)$$

which is positive for $w_B^* < 1 - t$.

The manufacturing firm's profit

Insert segmented prices in the profit function to obtain the profit in the segmented equilibrium:

$$\Pi^S = \frac{S + b}{4b} \quad (74)$$

and correspondingly retail prices and the PI quantity for low trade costs to obtain the profit in the PI equilibrium for $t < \underline{t}$

$$\Pi_{t < \underline{t}}^{PI} = \frac{Sb + 5Sbt^2 - 2tSb + 4t^2 + S^2}{4Sb} \quad (75)$$

Next insert retail prices and the PI quantity for intermediate trade costs to obtain the profit in the PI equilibrium for $t \geq \underline{t}$

$$\Pi_{t \geq \underline{t}}^{PI} = \frac{8Sb^2 + 20Sb^2t^2 - 8tSb^2 + 9S^2b + 16bt^2 + 4b + 4S}{4b(4 + 9Sb)} \quad (76)$$

and, finally, retail prices for high trade costs to obtain the profit in the no-arbitrage equilibrium

$$\Pi^{NA} = \frac{4b + 4S - Sb^2 + 4tSb^2 - 4Sb^2t^2}{16b}. \quad (77)$$

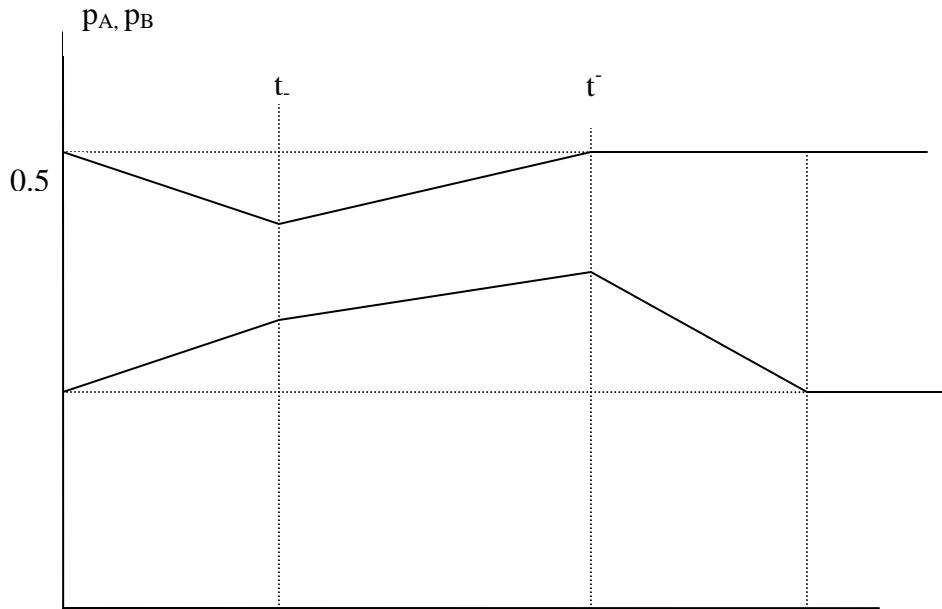


Figure 1. Equilibrium Retail Prices in A, B ($S=1, b=2$)

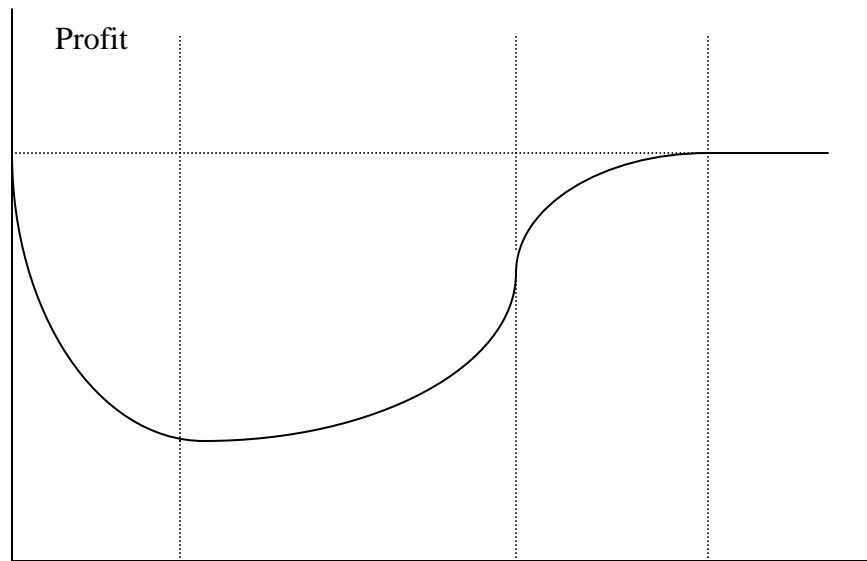


Figure 2. Manufacturing Firm's Profit ($S=1, b=2$)

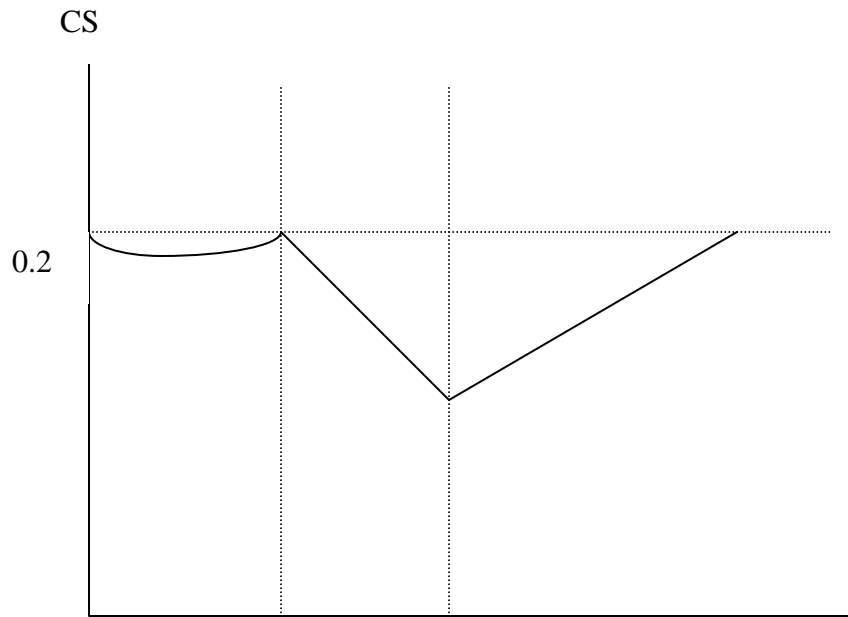


Figure 3. Joint CS ($S = 1, b = b^*$)

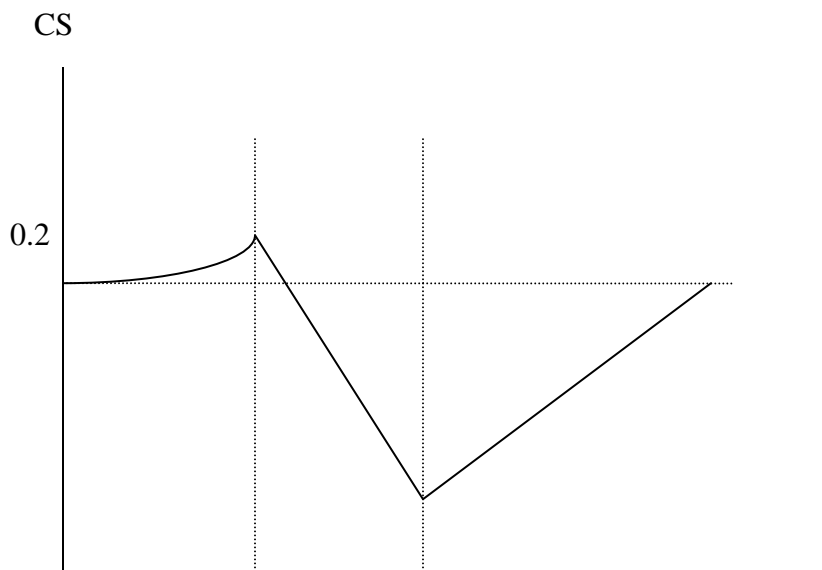


Figure 4. Joint CS ($S = 1, b = 2$)

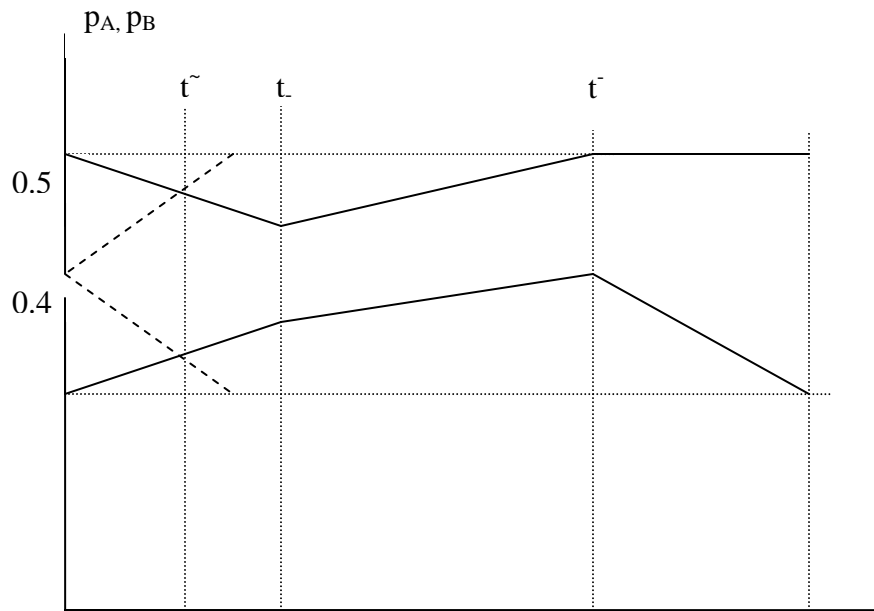


Figure 5. Retail Prices with Retail Arbitrage ($S=1, b=1.5$)