

Monopoly pricing with dual-capacity constraints

Robert Somogyi^{1,2} 

¹Department of Finance, Budapest University of Technology and Economics, Budapest, Hungary

²Centre for Economic and Regional Studies, Budapest, Hungary

Correspondence

Robert Somogyi, Department of Finance,
Budapest University of Technology and
Economics, Muegyetem rkp. 3, Budapest
H-1111, Hungary.
Email: somogyi.robert@gtk.bme.hu

Funding information

Nemzeti Kutatási Fejlesztési és Innovációs
Hivatal, Grant/Award Number: OTKA
FK-142492; National Research
Development and Innovation Office
(NKFIH), Grant/Award Number: OTKA
FK-142492

Abstract

This paper studies the price-setting behavior of a monopoly facing two capacity constraints: one on the number of its consumers, and the other on the amount of products it can sell. The characterization of the firm's optimal pricing and optimal customer mix as a function of its two capacities reveals a rich structure. In contrast to the results under one-dimensional capacity constraints with constant marginal cost of production, a firm may optimally respond to an exogenous reduction in one of its capacities by decreasing one of its prices. Moreover, neglecting the existence of the second capacity constraint can reverse some policy interventions' effects on consumer welfare. In particular, easing a regulatory restriction on one of the constraints may harm the average consumer.

1 | INTRODUCTION

The economics literature typically considers one-dimensional capacity constraints (Kreps & Scheinkman, 1983). This is idealized because in real-world production processes, firms typically face several capacity constraints (size of plants, inventories, workforce, etc.). In general, due to the use of supply chains, firms are constrained on even more aspects of their production. The objective of this paper is to develop a theory of monopoly pricing in the presence of multiple capacity constraints.

Examples of industries with profit-maximizing firms facing dual-capacity constraints range from hospitals, through restaurants, to the freight transport industry. *Hospitals* are constrained by the number of beds available in their recovery rooms on the one hand and operating room time on the other. *Restaurants* have to take into account in their pricing decisions both the number of tables they have at their disposal and the size of their kitchen, which limits the amount of food they can prepare. Dual capacities are key characteristics of the *freight transport* industry as well. Both in ocean container shipping and air cargo transport, an important concern of the transporting company is optimizing the mix of items according to both their size and their weight. As physical and regulatory limits are present in both dimensions, profit-maximizing firms cannot avoid taking into account both constraints.

Several questions arise if one wants to understand the consequences of the coexistence of the two types of capacity constraints. How much will the predictions of the model change compared with a model with only one capacity constraint? What are the optimal prices a firm must charge to different consumer groups? How will these optimal prices change as a function of the capacity levels? What is the firm's optimal customer mix given the dual constraints? How do aggregate consumer surplus, profit, and total welfare vary with the capacity levels?

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2023 The Authors. *Journal of Economics & Management Strategy* published by Wiley Periodicals LLC.

To answer these questions, I consider a price-setting monopolist facing two types of capacity constraints. The firm is unable to serve more than K consumers. In addition, it cannot sell more than a quantity Q of its products. Consumers differ in their price-inelastic individual demands and willingness to pay (WTP). There are two consumer groups: high-types intend to buy a larger amount of the product than the low-types. The monopoly can observe the individual demand of each consumer but not their WTP. The WTP of high-types and low-types are distributed along two different intervals. Some high-types have a larger WTP than all the low-types whereas some low-types have a larger *per-unit* WTP than all the high-types. The monopoly is allowed to offer price-quantity bundles to discriminate between consumers.

The existence of a second capacity constraint fundamentally changes the monopoly's optimal behavior when the cost of production is relatively small and linear up to the capacity constraints. The first main result (Propositions 1 and 4) is that, surprisingly, optimal prices can increase in the size of capacities. In particular, the price charged for high-types strictly increases in K , and the price for low-types strictly increases in Q for capacity levels when both constraints bind. This means that following an exogenous reduction of one of its capacity levels (e.g., caused by a natural disaster or a regulatory change), the firm's best response is to decrease the price it charges to one of its consumer groups. This is in sharp contrast to the standard case with only one capacity constraint and constant marginal cost of production, where the monopoly should raise both of its prices after such a loss in capacity.

The intuition for this result is the following. I show that there are always market settings where the firm charges prices in a way that both constraints bind. For small values of K , the capacity on the mass of people served, the monopoly serves mostly high-types. This is because the small K makes the other constraint, Q relatively lax, and high-types use up a lot of that constraint but have a high WTP. As K increases, low-types become relatively more valuable for the firm as they use up less of the Q constraint that becomes relatively stricter. To make space for the low-types, the firm is interested in serving fewer high-types, so it raises the price charged to them.

The second main result of the paper (Propositions 2 and 5) is related to consumer welfare effects. I show that there always exists a range of parameters in which relaxing one of the capacity constraints decreases aggregate consumer surplus. This has important consequences for policy interventions that affect capacity levels. An easing of a restriction on capacity levels (e.g., safety standards for airlines or sanitary regulations for restaurants during a pandemic) can in fact harm the average consumer. This phenomenon cannot occur in models that consider a single-capacity constraint and constant marginal cost of production. The decrease in consumer surplus is a consequence of the monopoly adjusting its optimal mix of consumers in the following sense. When K is very small, only high-types get served. As it increases above a threshold value, the firm starts serving some low-type consumers as well. However, as the transition is smooth, close to the turning point, the firm still serves many high-types and only a few low-types. The price increase suffered by the numerous high-types dominates the gain of the few low-types that start being served, thus the aggregate consumer surplus decreases.

Finally, in Proposition 3 I show that under endogenous capacity choice, that is, in the long run, the monopoly chooses its capacity levels such that they both bind. Notably, this is exactly the parameter region where the novel pricing and welfare effects arise when there is an exogenous change in one of the capacity levels.

For an illustration of the main results of the paper, take the example of a restaurant being constrained both by the number of meals it can produce (K) and the number of chairs it has at its disposal (Q). For simplicity, assume everyone wants to eat the same quantity, one meal. There are two types of consumers: low-type consumers buy the meal as takeout, whereas high-type consumers prefer to eat it in the restaurant thus occupying a chair. During the COVID-19 pandemic, several governments made sanitary rules restricting the occupancy rates of restaurants, resulting in a sudden reduction in chairs, one of the capacity constraints (Q). At the same time, the other capacity constraint remained unchanged: the restaurants could still prepare the same number of meals. Given such a reduction in seating capacity, conventional wisdom suggests that restaurants should increase prices. However, my model suggests that the optimal pricing of the restaurants is more subtle. They should indeed increase their prices charged for high-types who occupy a chair, however, they should decrease their prices for takeout consumers to attract more of them and sell out all of their meals. In addition, this price decrease might increase aggregate consumer welfare: if many takeout consumers benefit from the lower prices, their gain may outweigh the loss of the few high-type consumers. Therefore, a partial easing of the sanitary restrictions (an increase in Q) may even harm the average consumer.

1.1 | Related literature

The monopoly's problem of third-degree price discrimination with a single-capacity constraint is a textbook exercise (see Besanko & Braeutigam, 2010, p. 507, Exercise 12.6). Therefore, the *literature of capacity-constrained pricing* has



mainly focused on the case of competition. A rich body of literature followed the seminal papers of Kreps and Scheinkman (1983) and Davidson and Deneckere (1986).¹ The key contribution of the present paper is that it explicitly models the presence of two capacity constraints, and shows that such coexistence leads to results that are novel qualitatively.

The welfare analysis of the model with general distribution function is related to another stream of recent literature investigating *welfare effects of monopoly's third-degree price discrimination* (Aguirre et al., 2010; Bergemann et al., 2015; Cowan, 2007, 2012). The current paper contributes to this literature by explicitly taking into account the presence of capacity constraints.

My model is also related to the *literature of multiproduct pricing* by a monopolist (Armstrong & Vickers, 2018, 2023; Baumol & Bradford, 1970). Indeed, the monopoly serving two consumer groups in my model can be seen as serving two markets interconnected by dual-capacity constraints. In a recent article, Armstrong and Vickers (2023) show that there exist cost structures such that an increase in the cost of one of the products leads to (i) a decrease of all prices if at least two products are substitutes (“Edgeworth paradox”) and (ii) an increase in consumer surplus (“surplus paradox”). A tightening of one of the capacity constraints can be interpreted as an increase in the cost of production.² My analysis shows, however, that under constant marginal costs, neither of the paradoxes can occur in settings with a single-capacity constraint. Therefore, my main results about the nonmonotonicities of optimal prices and consumer surplus under dual capacities are not simply manifestations of the above paradoxes. I discuss the relevance of the small constant marginal cost assumption in Section 7.

Models where several capacity constraints coexist have so far been relegated to the realms of *operations research and revenue management*. Patient admission planning for scheduled surgeries and patient mix optimization are both important problems hospitals have to face. The multidimensional nature of capacities in hospitals is crucial for such planning. Several papers in the operations research literature focus on solving various problems that arise in a context where the treatment of different categories of patients requires different levels of capacities (Adan & Vissers, 2002; Banditori et al., 2013). My model focuses on the optimal pricing structure of the firm, which is a crucial distinction from this strand of literature.

Multidimensional capacities are crucial in freight transport, such as the air cargo industry or the container shipping industry. Since dynamic pricing is widespread in these industries, a few papers model freight transport pricing by extending standard revenue management models to accommodate multiple capacities (Kasilingam, 1997; Xiao & Yang, 2010). Similarly, the hospitality industry's capacity management literature has also recognized the importance of dual capacities (Bertsimas & Shioda, 2003; Kimes & Thompson, 2004). Being simpler than these dynamic revenue management models, my model reveals nonmonotonicity properties that have not been discovered in more complicated environments.

The rest of the paper is organized as follows. Section 2 discusses a simple benchmark, then outlines the main model. Section 3 describes the monopoly's optimal pricing behavior. Section 4 discusses the consequences of optimal pricing for consumer surplus and total welfare. Section 5 generalizes the model by allowing capacity levels to be chosen endogenously. Section 6 presents the model with more general distribution functions of consumers' WTP. Section 7 discusses the main results and concludes. All omitted proofs are relegated to the appendix.

2 | THE MODEL

2.1 | A simple benchmark

Throughout the paper, I will compare the results of the dual-capacity model with standard results assuming a single-capacity constraint. For this reason, I shortly present the simplest benchmark model. Consider a price-setting monopolist that can produce for constant marginal cost up to a capacity constraint. All consumers have a unit demand but are heterogeneous in their WTP.³ For capacity levels that are binding, optimal prices are a strictly decreasing function of the capacity constraint, whereas consumer surplus is strictly increasing in the capacity level. Intuitively, as its capacity constraint is relaxed, the monopoly increases its profit by decreasing its price to increase its demand, which clearly benefits consumers. In this paper, I will show that in a market with dual-capacity constraints, the monotonicity of optimal prices, demand, and consumer surplus can all break down.

2.2 | The dual-capacity model

Consider a market served by a monopoly consisting of two consumer groups. Each consumer is characterized by their individual demand and their total WTP for the product. Low-types want to consume a fixed amount of $q_L > 0$ products whereas high-types want to consume a fixed amount of $q_H > q_L$, that is, individual demand is price-inelastic. Hospital patients are a good example of price-inelastic individual demands: even if a long surgery (e.g., a kidney transplant) is very cheap, someone in need of a short surgery (e.g., fixing a broken arm) will never prefer having the longer one.

A consumer of type $i \in \{L, H\}$ with total WTP w has a net consumer surplus of $w - p_i$ if he buys a quantity q_i of the product for price p_i , and 0 otherwise. The total WTP of consumers of type i is uniformly distributed on the interval $[0, v_i]$. Consumers maximize their net surplus, and they demand the good if and only if their net surplus is positive. Assume that the total mass of high-type and low-type consumers is αv_H and $(1 - \alpha)v_L$, respectively, where $0 \leq \alpha \leq 1$ scales the relative weight of the two consumer groups. The monopoly can observe α , the consumers' individual demand, and the distribution of their WTP but not their individual values.

Assumption 1. Let the WTP of consumers satisfy the following conditions:

$$0 < v_L < v_H \quad \text{and} \quad v_L/q_L > v_H/q_H.$$

Assumption 1 guarantees that some high-type consumers' valuation always exceeds all the low-types' valuation, whereas the per-unit WTP of some low-type consumers is greater than the per-unit WTP of all high-type consumers. This assumption corresponds to diminishing marginal utility of consumption in the present setting. Moreover, it restricts the analysis to the most interesting cases, because otherwise, the monopoly would always prefer to serve consumers of one group first, irrespective of the size of capacity constraints. Finally, notice that high-types are not necessarily more valuable for the firm, the terminology refers to the high level of their individual demand.

The heterogeneity of individual demands is a natural assumption in many industries. In hospitals, each consumer (i.e., patient) needs one bed, but consumers differ in their need of operating room time (e.g., patients in need of a short surgery are the low-types, whereas patients in need of a long surgery are the high-types). For restaurants, some consumers prefer dining in (high-types), thus occupying a table, whereas other consumers prefer takeout (low-types).

I analyze a monopoly facing two types of capacity constraints:

- K denotes the maximal mass of consumers the firm can serve,
- Q denotes the maximal total production of the firm.

Both constraints are exogenously given in this baseline model. For simplicity, production is costless up to capacity then it becomes impossible.⁴

The monopoly has an optimal pricing structure that consists of offering at most two price-quantity bundles. Given the price-inelastic individual demands, no consumer would buy any bundle that offers them a quantity different from their desired demand. Moreover, if the firm were to offer several bundles with the same quantity for a different price, consumers would only buy the cheapest one. Let p_H and p_L denote the price of the bundle with high and low quantities, respectively.

3 | OPTIMAL MONOPOLY PRICING AND CUSTOMER MIX

In this section, I describe and solve the monopoly's profit-maximization problem. For any price $p_H \in [0, v_H]$, the high-type consumers willing to buy are the ones who have a higher WTP than p_H . They represent a fraction $\frac{v_H - p_H}{v_H}$ of the high-types, which means that the total mass of high-type consumers who demand the good is given by $\alpha(v_H - p_H)$. Similarly, for any price $p_L \in [0, v_L]$, the total mass of low-types willing to buy is $(1 - \alpha)(v_L - p_L)$. Hence the total demand the monopoly faces is given by

$$\alpha(v_H - p_H)q_H + (1 - \alpha)(v_L - p_L)q_L.$$

Importantly, the monopoly also has the option of not serving one of the consumer groups. Hence, it must choose between serving both consumer groups, excluding low-type consumers, or high-type consumers. Notice that in some cases the latter possibility can be profitable since some low-type consumers have a higher per-unit WTP than all the high-types. The monopoly's maximization problem takes the following form:

$$(P\text{-BL}) \quad \max_{p_L, p_H} \pi = \alpha(v_H - p_H)p_H + (1 - \alpha)(v_L - p_L)p_L \quad \text{s.t.}$$

$$\alpha(v_H - p_H) + (1 - \alpha)(v_L - p_L) \leq K, \quad (1)$$

$$\alpha(v_H - p_H)q_H + (1 - \alpha)(v_L - p_L)q_L \leq Q, \quad (2)$$

$$p_L \leq v_L, \quad (3)$$

$$p_H \leq v_H, \quad (4)$$

$$p_L \geq 0, \quad p_H \geq 0. \quad (5)$$

Constraint (1) provides the upper bound on the maximal mass of people that the monopoly can serve, K . Constraint (2) is the capacity constraint on total production: The mass of people buying times their individual demand cannot exceed Q . Constraints (3) and (4) restrict the monopoly not to choose exceedingly high prices. Note that this is without loss of generality as the monopoly can achieve the same outcome, namely, zero demand from consumer group $i \in \{L, H\}$, by charging $p_i = v_i$ instead of $p_i > v_i$. These constraints will help with the exposition of results as low-types, or high-types will be excluded if and only if (3) or (4) are binding, respectively. Finally, the inequalities in (5) are nonnegativity constraints for the prices.

Solving the optimization problem, there are six regions of the K – Q parameter space where the firm's optimal pricing behavior is qualitatively different. Figure 1 depicts the monopoly's optimal partitioning of the K – Q space.⁵ The firm chooses optimal prices in such a way that it serves some consumers of both types and (i) only capacity K binds in region K , (ii) only capacity Q binds in region Q , (iii) both K and Q bind in region KQ , and (iv) none of the constraints bind in region U . The monopoly excludes one of the consumer groups in the remaining two regions. Namely, only capacity K binds and the monopoly excludes low-types in region EL . Conversely, only capacity Q binds and the monopoly excludes high-types in region EH .⁶

Lemma 1 provides a complete characterization of the firm's optimal choice for any combination of capacity levels.⁷

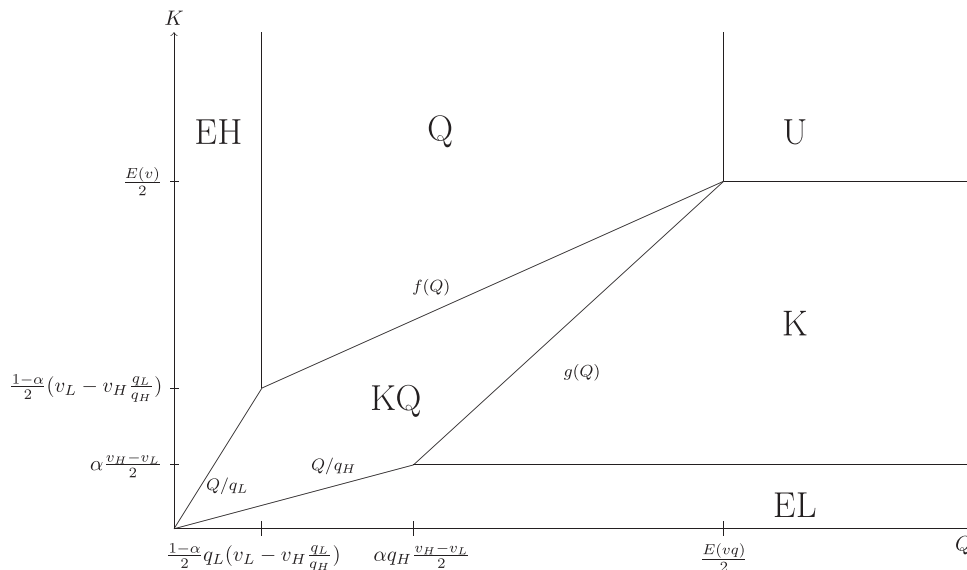


FIGURE 1 Optimal division of the parameter space.

Lemma 1. *The optimal prices of the monopoly are*

- $p_L = v_L + \alpha \frac{v_H - v_L}{2} - K$ and $p_H = v_H - (1 - \alpha) \frac{v_H - v_L}{2} - K$ in region K , where K binds, Q is slack, some consumers of both groups are served;
- $p_L = \frac{q_L}{E(q^2)} \left(\frac{\alpha}{2} (v_L \frac{q_H}{q_L} + v_H) q_H + (1 - \alpha) v_L q_L - Q \right)$ and $p_H = p_L \frac{q_H}{q_L}$ in region Q , where Q binds, K is slack, some consumers of both groups are served;
- $p_L = v_L - \frac{K q_H - Q}{(1 - \alpha)(q_H - q_L)}$ and $p_H = v_H - \frac{Q - K q_L}{\alpha(q_H - q_L)}$ in region KQ , where both K and Q bind, some consumers of both groups are served;
- $p_L = v_L/2$ and $p_H = v_H/2$ in region U , where both K and Q are slack, some consumers of both groups are served;
- $p_H = v_H - \frac{K}{\alpha}$ in region EL , where K binds, Q is slack, low-types are excluded;
- $p_L = v_L - \frac{Q}{q_L(1 - \alpha)}$ in region EH , where Q binds, K is slack, high-types are excluded.

The proof, relegated to the appendix, shows that Assumption 1 guarantees the existence of all six regions enumerated in Lemma 1 and depicted in Figure 1. The main driver of the results is the two consumer groups' varying relative attractiveness for the firm. For any given Q , an increase in K makes the constraint on total production, Q , tighter. This in turn makes low-types, who consume less of the tighter capacity, relatively more valuable for the firm, so in optimum, the firm adjusts its prices to attract more of them.

Next, to provide some intuition, I briefly describe the six regions. First, the monopoly serves only high-types in region EL . This occurs when the constraint on the mass of people (or the number of meals in the restaurant example, K) is very tight, and at the same time, the constraint on total production (the number of chairs in the restaurant example, Q) is abundant. Clearly, the monopoly prefers filling its tight capacity with high-type consumers with a higher WTP. Second, the opposite is true in the EH region where the constraint on total production is very tight. Indeed, the monopoly prefers selling all its production capacity to low-types thanks to their larger per-unit WTP.

Third, arguably the most interesting region is the diamond-shaped region KQ . In Section 5 I will show that when a firm can choose its capacities endogenously, it will choose them to be in this region. Both capacities are of intermediate size in this region and large enough that some consumers of both groups are served. Moreover, both capacity constraints simultaneously bind. One of the main results of the paper, captured by Proposition 1, is that optimal prices behave in an unexpected way in this region.

The remaining three regions follow quite intuitively if one increases capacity levels starting from region KQ . Indeed, in region K capacity constraint K binds and Q is slack. Vice versa, capacity constraint Q binds and K is slack in region Q . Finally, both constraints are slack when they are very large in region U . In fact, this region corresponds to the monopoly's unconstrained optimum.

Proposition 1 highlights the first main result of the paper.

Proposition 1. *There always exist market settings, namely, capacity-pairs in the KQ region, where the optimal response of the monopoly to an exogenous reduction in capacity K is to reduce its prices for high-type consumers and increase its prices for low-type consumers. Conversely, the optimal response to a reduction of Q is an increase in prices charged to high-types and a decrease in prices for low-types.*

Proposition 1 shows that the monopoly's pricing strategy with zero marginal costs under dual-capacity constraints is qualitatively different than its strategy under a single-capacity constraint. The proof is straightforward by taking the derivatives of optimal prices in region KQ given in Lemma 1.

For intuition, take the example of a restaurant serving dine-in customers (high-types) and takeout customers (low-types). For simplicity, everyone wants to eat one meal, and dine-in consumers have a higher total WTP for it than take-away consumers. Let K denote the number of meals its kitchen can prepare, and Q denote the number of tables. Imagine that a pandemic results in a sudden reduction in the number of available tables, Q . Then, to reoptimize its profit, the firm will want to serve fewer dine-in consumers who take up its reduced table capacity, and more takeout

consumers as it can still prepare the same number of meals. To attract this new consumer mix, it must increase the prices it charges for high-types, but *reduce* the prices it charges to low-types.⁸ This latter action is in direct contrast to the standard recommendation obtained from single-capacity models with constant marginal costs, namely that firms should increase their prices if their capacity is reduced. Indeed, following that strategy and ignoring the presence of a second capacity constraint would lead to suboptimal profits.

This result is also in line with some pre-COVID trends in the healthcare sector. In particular, hospitals are bound by the capacity of their operating rooms (K) and their recovery rooms (Q). Some surgery types (e.g., some types of orthopedic surgery) can be performed both in an ambulatory and a nonambulatory way. In the former case, patients are called outpatients and do not use beds in the recovery room at all, whereas in the latter case, the so-called inpatients use some of the recovery room capacity. Thus, the two types of patients' need for operating room capacity is identical whereas they differ in their use of recovery room capacity. Clearly, inpatients correspond to high-types and outpatients to low-types in my model. It is a well-documented fact that the number of available beds in North American hospitals' recovery rooms has substantially decreased between 2000 and 2010 (see, e.g., Halpern et al., 2016). Furthermore, over the same period, the number of ambulatory surgeries has increased (Crawford et al., 2015). This coincides exactly with the predictions of my model: capacity in the recovery room being scarcer (i.e., after a reduction in Q), by means of pricing the hospitals alter their patient mix in a way that attracts more outpatients (i.e., low-types) and fewer inpatients (i.e., high-types).

Finally, Proposition 1 has an important corollary regarding the optimal prices as a function of capacity level.

Corollary 1. *For every $Q < \bar{Q}$, the optimal price charged by the monopoly to high-type consumers is nonmonotonic in capacity level K . Moreover, for every $K < \bar{K}$, the optimal price charged by the monopoly to low-type consumers is nonmonotonic in capacity level Q .*

The proof of the corollary is straightforward by comparing the derivatives of optimal prices in Lemma 1. Intuitively, looking at optimal prices at all possible capacity levels, the nonmonotonicity is a direct consequence of the prices increasing in the KQ region. Indeed, prices decrease, that is, behave as predicted by models with a single-capacity constraint, in all other regions, hence the nonmonotonicity.⁹

4 | WELFARE

In this section, I investigate the welfare properties of the monopoly's optimal pricing behavior in the presence of dual-capacity constraints. I first analyze aggregate consumer surplus as a function of the capacity constraints given that the monopoly chooses its profit-maximizing prices described in Lemma 1. Next, I calculate total welfare as the sum of consumer surplus and the monopoly's profit, although most of the interest comes from the study of consumer surplus.

Aggregate consumer surplus is¹⁰

$$CS = \frac{\alpha}{2}(v_H - p_H)^2 + \frac{1 - \alpha}{2}(v_L - p_L)^2.$$

Clearly, consumer surplus is weakly decreasing in both prices p_L and p_H . As shown in Section 3, with the exception of the KQ region where both constraints bind, prices are decreasing in the size of capacities. Hence consumer surplus is increasing in capacity levels for any capacity-pairs outside of the KQ region.

However, the problem is more complicated in the KQ region where both capacities bind. In this region, p_H is increasing whereas p_L is decreasing in K . The following proposition sheds light on the effect of this trade-off.

Proposition 2. *Aggregate consumer surplus is nonmonotonic in the size of the capacity constraints in the parameter region where both capacities are binding. In particular, there always exists a region of capacity-pairs inside KQ where consumer surplus is decreasing in K and increasing in Q . Moreover, there exists a second, disjoint region inside KQ where consumer surplus is decreasing in Q and increasing in K .*

The proof of Proposition 2 shows that consumer surplus is decreasing in K in the KQ region iff $K \leq Q \frac{E(q)}{E(q^2)}$ (light gray area in Figure 2). The intuition for this result is the clearest when Q is relatively low and one can consider the

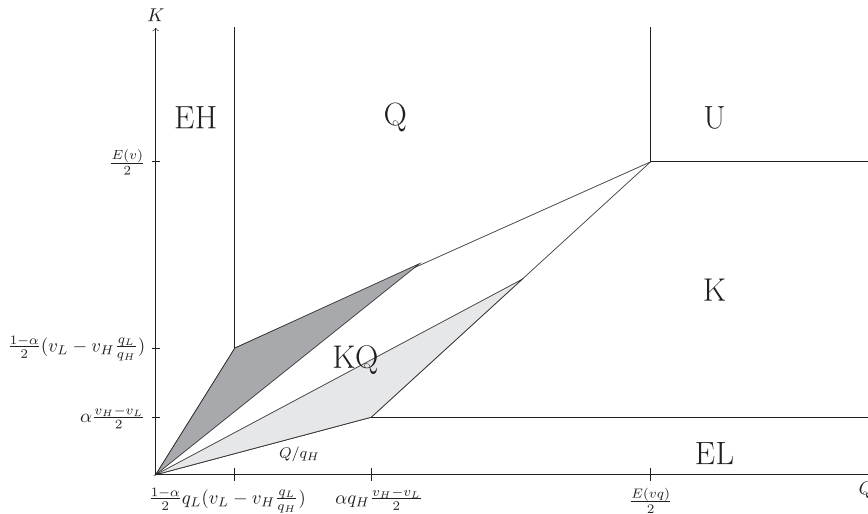


FIGURE 2 Nonmonotonicity of consumer surplus.

transition between the regions EL and KQ. When K is increasing from a value lower than Q/q_H the monopoly first excludes all low-types and serves only high-types. As K reaches Q/q_H it starts to be profitable to serve some low-type consumers as well such that both constraints bind. To accommodate low-types, the firm starts serving fewer high-types thus it increases p_H whereas it lowers p_L .

Consumer surplus is affected by three factors. First, as K is binding in both regions, the total mass of consumers served goes up which ceteris paribus increases consumer surplus. Second, the decrease in p_L has the same effect as well. Third, the increase of p_H goes in the opposite direction. To see that this third effect dominates the first two, one should consider the mass of consumers affected. Indeed, when K is relatively small, the firm serves a lot more high-types than low-types, so the loss suffered by the high-types dominates the gain of the few low-types. Similarly, consumer surplus is decreasing in Q in the KQ region iff $K \geq Q/E(q)$ (dark gray area in Figure 2), and the arguments are analogous to the ones described above.

Consumer surplus decreasing in a capacity level is a novel result in the literature of capacity-constrained pricing. Indeed, this result can never arise in a model with a single-capacity constraint under the standard assumption of constant marginal cost of production. When a firm's (binding) single capacity is shrunk, it serves fewer consumers at higher prices, therefore consumers always suffer from the loss of capacity. In contrast, when a firm faces dual-capacity constraints and one of them is shrunk, Proposition 2 demonstrates that the firm may want to reoptimize its customer mix in a way that ultimately benefits consumers.

This result highlights that one has to be careful when evaluating the welfare effects of regulatory interventions impacting capacity levels in markets with dual-capacity constraints. As a first example, take the reduction of recovery bed capacity in the United States I describe after Proposition 1. Crawford et al. (2015) suggest that this reduction happened due to changed incentives provided by the Medicare program.¹¹ As discussed above, the result of this reduction in capacity (Q) was an increase in the number of outpatients (low-types consumers). Crawford et al. (2015) demonstrate that this was a rather positive change from a medical perspective. On the basis of the results of models with a single-capacity constraint and constant marginal costs, one would think this came at the price of a decreased financial well-being of the average consumer (lower capacity resulting in fewer patients treated at higher prices). However, the result in Proposition 2 implies that if a sufficiently large share of the patients is outpatients, the regulatory change may have instead made the average patient financially better off as well.

In a second example, I show that in a market with dual-capacity constraints, a well-intentioned regulatory intervention relaxing one of the constraints can backfire by harming the average consumer. Airplanes are bound by the number of crew needed (K) on the one hand and by the number of seats (Q) on the other hand. In many short-haul flights in Europe, the part of the plane reserved for business-class consumers can be flexibly changed by simply moving a curtain between rows. The middle seat in rows serving business-class passengers is left empty for their comfort. Thus these passengers correspond to high-types in my model by effectively using 1.5 seats (compared with low-type, economy class passengers using exactly 1 seat). Now imagine that the regulator eases safety requirements relating to emergency evacuation times, allowing the airlines to fit in more seats. This is an exogenous increase in Q , which by Proposition 1 decreases the number of low-types served by the airline and increases the price charged to them.

Proposition 2 shows that such a regulatory easing may end up harming the average consumer due to the reoptimization by the airline. In particular, it decreases aggregate consumer surplus if the share of economy-class consumers is high.

For completeness, I also consider total welfare. In this context total welfare simply equals the sum of the monopoly's profit and consumer surplus, that is,

$$TW = CS + \pi = \frac{\alpha}{2}(v_H^2 - p_H^2) + \frac{1-\alpha}{2}(v_L^2 - p_L^2),$$

whenever some consumers of both types are served. The next proposition provides comparative statics results for total welfare.

Lemma 2. *Total welfare is increasing in both capacity levels for every parameter value.*

I show in the appendix that in both areas where consumer surplus is decreasing, profit increases faster than consumer surplus decreases. This property clearly holds for other capacity-pairs where both consumer surplus and profit are increasing in capacities.

5 | ENDOGENOUS CAPACITY CHOICE

In this section, I investigate the monopoly's optimal choice of capacity levels in the long run. Although in the short run, it is reasonable to assume that the monopoly chooses its prices facing fixed capacity levels, in the long run, firms can extend or shrink both of their capacities.

Let $c_K(K)$ denote the cost of building capacity K and let $c_Q(Q)$ denote the cost of building Q . I assume that the costs are separable. In the hospital example, although there is a fixed cost of constructing the hospital building, the additional costs of adding beds and equipping the operating rooms are separable.

Assume that these costs are strictly positive whenever the capacity levels are strictly positive. The monopoly maximizes its profit which is now a function of its two capacities as well as its prices. The optimal choice of prices for any capacity-pair is the one described in Lemma 1. Proposition 3 holds under these very general conditions.

Proposition 3. *If the capacity choice is endogenous and the cost of building capacities is strictly positive then the monopoly chooses its prices and capacity levels in such a way that both constraints bind, that is, the optimal capacities are always chosen from the KQ region.*

The proof is straightforward. To get a contradiction, assume that the monopoly chooses its optimal capacities in such a way that only one of the capacities binds. Then it could increase its profit by reducing the capacity building cost and still be able to serve the same demand at the same prices, so the original capacity choice could not have been optimal. Intuitively, in a world of deterministic demand, it is never profitable for a monopoly to build unused capacity.

Proposition 3 underpins the importance of the main results of the paper, Propositions 1 and 2 in the following way. In the long run, when the monopoly is free to choose its capacities, it chooses its optimal capacity-pair from the KQ region. However, this is exactly the region where the unexpected comparative statics results arise in the short run. Therefore, it shows that the starting point of the most relevant short-run analyses is indeed the KQ region, so the main results identified above are likely to be relevant in real-world settings.

6 | GENERAL DISTRIBUTION OF CONSUMERS

In the baseline model presented in Section 2, consumers' WTP is distributed uniformly within each group, which implies linear total demand that in turn leads to the clear-cut results presented in Propositions 1 and 2. In this section, I generalize the baseline model by assuming a general distribution function of consumers' WTP. I identify sufficient conditions on the distribution functions for the main results of the paper to carry over. In particular, I show that Myerson's regularity condition and the analog of Assumption 1 guarantee the existence of the six regions discovered in

the baseline model (illustrated in Figure 1), and therefore optimal prices are nonmonotonic and consumer surplus can decrease in capacity levels.

Let the WTP of consumers of type $i \in \{L, H\}$ be distributed according to the twice continuously differentiable cumulative distribution function F_i with support $[0, \theta_i]$. Let f_i denote the density function, that is, the first derivative of F_i . For any $p \geq 0$ let $D_i(p) = 1 - F_i(p)$ be the total demand function that measures the proportion of i -type consumers willing to buy at price p . Let $P_i(N_i)$ denote the inverse demand function¹² of group i , that is, $N_i = D_i(P_i(N_i))$. Individual demand of consumers is the same as in the baseline model: $q_H > q_L > 0$. Let α and $1 - \alpha$ denote the total mass of high-types and low-types, respectively. As before, the monopoly is constrained by K , the total mass of consumers it can serve on the one hand, and by Q , its maximal production on the other hand. The maximization problem of the monopoly in terms of quantities written as

$$\begin{aligned}
 \text{(P-GEN)} \quad \max_{N_L, N_H} \pi &= \alpha N_H P_H(N_H) + (1 - \alpha) N_L P_L(N_L) \quad \text{s.t.} \\
 \alpha N_H + (1 - \alpha) N_L &\leq K, & (\lambda_K) \\
 \alpha q_H N_H + (1 - \alpha) q_L N_L &\leq Q, & (\lambda_Q) \\
 N_L &\geq 0, & (\lambda_L) \\
 N_H &\geq 0. & (\lambda_H)
 \end{aligned}$$

It will prove useful to define the virtual valuation functions

$$\phi_i(p) \equiv p - \frac{1 - F_i(p)}{f_i(p)}, \quad i \in \{L, H\}.$$

Assumption 2. Let the distributions of the WTP of consumers satisfy the following conditions:

- (i) $\phi'_i(p) > 0$ for $i \in \{L, H\}$,
- (ii) $0 < \theta_L < \theta_H$ and $\theta_L/q_L > \theta_H/q_H$.

It is a standard result in the literature¹³ that $\phi_i(p)$ increasing, a.k.a. Myerson's regularity condition, is a sufficient condition for the concavity of $N_i P_i(N_i)$ in N_i . Part (ii) is analogous to Assumption 1 in the baseline model, guaranteeing that at least some high-type consumers have a higher WTP than all low-type consumers, whereas some low-type consumers have higher per-unit WTP than all high-type consumers.¹⁴

Proposition 4 states that under these conditions, the optimal pricing strategy of the monopolist is remarkably similar to the one used in the baseline model.

Proposition 4. Under Assumption 2, the monopoly's optimal partitioning of the K - Q parameter space consists of the same six regions as identified in Lemma 1. In particular, there exists a nonempty core region where both capacity constraints bind and the optimal price charged to high-type (resp., low-type) consumers is increasing in K (resp., Q).

The proof in the appendix provides the optimal partitioning and the optimal mass of consumers served from each group in all of the six regions. I show that the monopoly's optimal pricing strategy is qualitatively the same as in the baseline model. One of the few differences is that the equivalents of the $f(Q)$ and $g(Q)$ lines are nonlinear under the more general distribution of WTP values. However, I show that these curves always exist and they are unique, increasing, and their slope is always larger than Q/q_H and lower than Q/q_L , respectively. Therefore, the picture of the optimal division of the parameter space looks almost exactly like the one depicted in Figure 1. The exact coordinates are of course only implicit functions of the cumulative distribution functions.

Consumer surplus in the context of general distributions can be written as

$$CS = \alpha \int_{p_H}^{\theta_H} (w - p_H) f_H(w) dw + (1 - \alpha) \int_{p_L}^{\theta_L} (w - p_L) f_L(w) dw.$$

Clearly, the consumer surplus is a decreasing function of both prices. In five parameter regions, both prices are decreasing in both capacity levels, thus aggregate consumer surplus increases in both K and Q , just like in standard models with a single-capacity constraint. However, in the core KQ' region where both capacities bind, Proposition 4

shows that p_H is increasing in K and decreasing in Q and conversely, p_L is increasing in Q and decreasing in K . Therefore, the consumer surplus of the high-types is decreasing whereas the consumer surplus of low-types is increasing in K . Proposition 5 generalizes Proposition 2.

Proposition 5. *Under Assumption 2, there exist two disjoint regions inside of the core region KQ' such that the consumer surplus decreases in K in one of the regions, and it decreases in Q in the other region.*

The proof in the appendix reveals that consumer surplus is decreasing in Q whenever at the optimal prices the demand elasticity to price ratio is larger for high-types than for low-types. Furthermore, consumer surplus is decreasing in K when at the optimal prices the demand elasticity to unit price ratio is higher for the low-types than for the high-types. This concludes the generalization of the results of the baseline model.

7 | DISCUSSION AND CONCLUSION

Several capacity constraints coexist in various real-world industries. This paper provides a formal economic analysis of the effects of dual-capacity constraints on optimal firm behavior. It reveals a rich structure of optimal monopoly pricing which in the short run is qualitatively different from the predictions of models of firms bound by a single capacity with constant marginal cost of production. First, the firm should decrease its prices for some of its consumers as an optimal response to an exogenous reduction in one of its capacities. Second, policy interventions affecting firms' capacity constraints may have unintended consequences in markets with dual-capacity constraints. Assuming constant marginal cost of production, these two main results can never arise in models with one single-capacity constraint. Moreover, the main results hold for a fairly general class of distribution of consumers' WTP.

7.1 | “Short run” versus “long run”

The novel predictions of considering dual capacities arise, both in terms of managerial implications and in terms of welfare, in the “short run,”¹⁵ that is, for exogenous capacity levels, in market settings when both capacities bind. Proposition 3 reveals that in the long run, the monopoly chooses its capacities exactly in a way that both constraints bind. These are exactly where the novel phenomena are expected to happen. There are at least three factors that may suddenly and exogenously reduce one of the capacity constraints. First, as in the restaurant example above, a global pandemic or other natural disasters instantaneously destroy capacities. Second, temporary supply chain problems (due to accidents like a ship blocking the Suez Canal, or, for example, due to mismanagement of the employee leave system¹⁶) can decrease production capacities. Third, regulation affects effective capacity levels in many markets with dual-capacity constraints (e.g., safety regulations for airlines, sanitary regulations during pandemics for restaurants, and social security reimbursements for hospitals). Therefore, the results of the model with exogenous capacity are often relevant in real-life situations.

7.2 | Low and constant marginal cost

In this paper, I found that with zero marginal cost of production up to capacity, nonmonotonicity of optimal prices and consumer surplus can occur only under dual-capacity constraints and never in setups with a single capacity. Zero marginal cost is a very common assumption in the literature of capacity-constrained pricing.¹⁷ However, as there are two consumer groups in my model, one may worry that the results are driven by their production cost being equal. To show this is not the case, I demonstrate in the appendix that the results of my model hold in case serving a low-type and high-type consumer costs cq_L and cq_H , respectively, if c is sufficiently small. I also show that nonmonotonicities do not arise in the analogous setting with a single capacity. Therefore, my model provides novel insights for capacity-constrained pricing when production costs are relatively low and linear up to the capacities.

I believe that the small positive marginal cost assumption is realistic in all of the motivating examples of my model. Namely, the marginal cost of serving an additional passenger (economy or business-class) is arguably very small both compared with the fixed cost of flying the airplane and to the opportunity cost of leaving a seat empty. Customers of

restaurants who dine in are somewhat more costly to serve than take-away consumers, but both costs are low compared with the cost of expanding the restaurant or even to consumers' WTP. In the hospital example, inpatients are somewhat more costly to treat than outpatients, but both costs are low compared with the patients' WTP or to the cost of expanding the hospital wing. Finally, it is also realistic in all three markets that the cost of serving a given type of consumer increases approximately linearly up to capacity.

7.3 | Competition

There is more than one firm in many of the motivating industries. Somogyi (2016, Chap. 3) shows that the main qualitative results under monopoly carry over for the symmetric equilibria of a corresponding symmetric duopoly, where both firms face dual-capacity constraints. In particular, equilibrium prices and consumer surplus may be nonmonotonic in capacities.

Future research could extend the results in several important aspects. One could verify the model's robustness by approximating the capacity constraints with convex and continuous cost functions. Moreover, one could test whether the model can accommodate more than two consumer groups.

ACKNOWLEDGMENTS

This paper is an improved version of a chapter of my Ph.D. dissertation prepared at Ecole Polytechnique and CREST. I am indebted to Francis Bloch, my advisor, for his invaluable guidance and support throughout this project. I would like to thank Øyvind Aas, Mark Armstrong, Fatma Aslan, Eric Avenel, Daniel Garrett, Piero Gottardi, Daniel Herrera, Johannes Johnen, Antonin Macé, Marc Möller, Andrew Rhodes, Francisco Ruiz-Aliseda, Gyula Seres, Christopher Stapenhurst, Pál Valentiny, Rafael Treibich, Thibaud Vergé, Xavier Vives, and Xavier Wauthy for fruitful discussions and comments; and participants at various seminars and conferences for suggestions and comments. All remaining errors are mine. I thank the support of the National Research Development and Innovation Office (NKFIH) under grant number OTKA FK-142492.

ORCID

Robert Somogyi  <http://orcid.org/0000-0003-1033-1754>

ENDNOTES

- ¹ Recent contributions include Cabon-Dhersin and Drouhin (2020), Fabra and Llobet (2023), Hunold and Muthers (2019a), Lemus and Moreno (2017), Montez and Schutz (2021), Somogyi (2020), and Somogyi et al. (2023).
- ² I would like to thank an anonymous referee for pointing out this connection.
- ³ The qualitative predictions are the same for a more complicated benchmark model with two consumer groups and third-degree price discrimination. Solving both models is straightforward. For details, see an earlier version of this paper (Somogyi, 2016, Chap. 2).
- ⁴ I show in the appendix that the results hold more generally under small but strictly positive constant marginal cost of production.
- ⁵ Figure 1 depicts the case of $(1 - \alpha)q_L(v_L - v_H \frac{q_L}{q_H}) < \alpha q_H(v_H - v_L)$. When this ordering is reversed, the figure changes accordingly. However, the coordinates of all lines and critical points remain the same. Functions $f(Q)$ and $g(Q)$ are defined in the appendix.
- ⁶ Throughout the paper, *EL* stands for “excluding low-types” and *EH* stands for “excluding high-types.”
- ⁷ I use the following notation throughout the paper to simplify the exposition of results: $E(x)$ denotes the weighted average of any two variables x_L and x_H , that is, $E(x) = \alpha x_H + (1 - \alpha)x_L$.
- ⁸ Some high-end restaurants indeed started selling takeout meals during the COVID pandemic at prices much below their usual price range (Madeira et al., 2021, p. 4).
- ⁹ For a more detailed analysis of the firm's pricing behavior, one should divide the Q - K plane into four regions by the three vertical lines going through the 3 values that appear on the Q axis in Figure 1. For details, see an earlier version of this paper (Somogyi, 2016, Chap. 2).
- ¹⁰ Consumer surplus in the *EL* (*EH*) region can be obtained by setting $p_L = v_L$ ($p_H = v_H$).
- ¹¹ In addition, Parkinson et al. (2019) show direct evidence that hospitals respond strongly to changes in the profitability of a consumer group. A policy reform in England in 2011 led to 76% and 152% increases in the volumes of two newly incentivized treatments.
- ¹² In the model with the generalized distribution of consumers, it proves useful to cast the monopoly's maximization problem in terms of quantities instead of prices, as this allows one to use standard results in the literature to show the uniqueness of the optimum.

- ¹³ The result comes from a direct comparison of the second derivative of the revenue function with respect to quantity and the first derivative of the virtual valuation function with respect to price. See, for example, Caplin and Nalebuff (1991) and Cowan (2007).
- ¹⁴ Note that $0 < \theta_L < \theta_H$ is a weaker assumption than first-order stochastic dominance or the monotone likelihood ratio property that are standard in mechanism design. Similarly, $\theta_L/q_L > \theta_H/q_H$ is weaker than assuming first-order stochastic dominance relative to per-unit WTP values.
- ¹⁵ The “short run” can in fact be very long: the airport capacity expansion project of South-East England comes to mind that has already taken decades: <https://www.ft.com/content/440560ba-1f27-11e5-aa5a-398b2169cf79>. Similarly, planning, designing, and building new hospital wings can take several years. Moreover, the recent example of Hungary buying thousands of ventilators during the peak of the COVID-19 pandemic only to find out that there is a lack of trained doctors and nurses to operate them constitutes a rather sad example of dual constraints: <https://shorturl.at/aHIT2>. Both links accessed on January 16, 2023.
- ¹⁶ <https://www.nytimes.com/2021/07/17/world/middleeast/suez-canal-stuck-ship-ever-given.html> and <https://www.economist.com/blogs/gulliver/2017/09/pilot-light> Both links accessed on January 16, 2023.
- ¹⁷ See, for example, Acemoglu et al. (2009), de Frutos and Fabra (2011), Garcia et al. (2023), Hunold and Muthers (2019b), Lemus and Moreno (2017), Möller and Watanabe (2010), Moreno and Ubeda (2006), Nocke and Peitz (2007), and Somogyi et al. (2023).
- ¹⁸ For all the six parameter regions that appear in Figure 1, I will denote the corresponding region in the generalized model with a superscript’.

REFERENCES

- Acemoglu, D., Bimpikis, K., & Ozdaglar, A. (2009). Price and capacity competition. *Games and Economic Behavior*, 66(1), 1–26.
- Adan, I. J. B. F., & Vissers, J. M. H. (2002). Patient mix optimisation in hospital admission planning: A case study. *International Journal of Operations & Production Management*, 22(4), 445–461.
- Aguirre, I., Cowan, S., & Vickers, J. (2010). Monopoly price discrimination and demand curvature. *The American Economic Review*, 100(4), 1601–1615.
- Armstrong, M., & Vickers, J. (2018). Multiproduct pricing made simple. *Journal of Political Economy*, 126(4), 1444–1471.
- Armstrong, M., & Vickers, J. (2023). Multiproduct cost passthrough: Edgeworth's paradox revisited. *Journal of Political Economy*. Forthcoming. <https://www.journals.uchicago.edu/doi/abs/10.1086/724573>
- Banditori, C., Cappanera, P., & Visintin, F. (2013). A combined optimization–simulation approach to the master surgical scheduling problem. *IMA Journal of Management Mathematics*, 24(2), 155–178.
- Baumol, W. J., & Bradford, D. F. (1970). Optimal departures from marginal cost pricing. *The American Economic Review*, 60(3), 265–283.
- Bergemann, D., Brooks, B., & Morris, S. (2015). The limits of price discrimination. *The American Economic Review*, 105(3), 921–957.
- Bertsimas, D., & Shioda, R. (2003). Restaurant revenue management. *Operations Research*, 51(3), 472–486.
- Besanko, D., & Braeutigam, R. (2010). *Microeconomics* (4th ed.). John Wiley & Sons.
- Cabon-Dhersin, M.-L., & Drouhin, N. (2020). A general model of price competition with soft capacity constraints. *Economic Theory*, 70(1), 95–120.
- Caplin, A., & Nalebuff, B. (1991). Aggregation and imperfect competition: On the existence of equilibrium. *Econometrica*, 59(1), 25–59.
- Cowan, S. (2007). The welfare effects of third-degree price discrimination with nonlinear demand functions. *The RAND Journal of Economics*, 38(2), 419–428.
- Cowan, S. (2012). Third-degree price discrimination and consumer surplus. *The Journal of Industrial Economics*, 60(2), 333–345.
- Crawford, D. C., Li, C. S., Sprague, S., & Bhandari, M. (2015). Clinical and cost implications of inpatient versus outpatient orthopedic surgeries: A systematic review of the published literature. *Orthopedic Reviews*, 7(4). <https://doi.org/10.4081/or.2015.6177>
- Davidson, C., & Deneckere, R. (1986). Long-run competition in capacity, short-run competition in price, and the Cournot model. *The RAND Journal of Economics*, 17(3), 404–415.
- de Frutos, M.-Á., & Fabra, N. (2011). Endogenous capacities and price competition: The role of demand uncertainty. *International Journal of Industrial Organization*, 29(4), 399–411.
- Fabra, N., & Llobet, G. (2023). Auctions with privately known capacities: Understanding competition among renewables. *The Economic Journal*, 133(651), 1106–1146.
- Garcia, D., Janssen, M. C. W., & Shopova, R. (2023). Dynamic pricing with uncertain capacities. *Management Science*. Forthcoming. <https://pubsonline.informs.org/doi/epdf/10.1287/mnsc.2022.4613>
- Halpern, N. A., Goldman, D. A., Tan, K. S., & Pastores, S. M. (2016). Trends in critical care beds and use among population groups and medicare and medicaid beneficiaries in the United States: 2000–2010. *Critical Care Medicine*, 44(8), 1490.
- Hunold, M., & Muthers, J. (2019a). Spatial competition and price discrimination with capacity constraints. *International Journal of Industrial Organization*, 67, 102–524.
- Hunold, M., & Muthers, J. (2019b). Spatial competition and price discrimination with capacity constraints. *International Journal of Industrial Organization*, 67, 102524.
- Kasilingam, R. G. (1997). Air cargo revenue management: Characteristics and complexities. *European Journal of Operational Research*, 96(1), 36–44.

- Kimes, S. E., & Thompson, G. M. (2004). Restaurant revenue management at chevys: Determining the best table mix. *Decision Sciences*, 35(3), 371–392.
- Kreps, D. M., & Scheinkman, J. A. (1983). Quantity precommitment and bertrand competition yield Cournot outcomes. *The Bell Journal of Economics*, 14(2), 326–337.
- Lemus, A. B., & Moreno, D. (2017). Price caps with capacity precommitment. *International Journal of Industrial Organization*, 50, 131–158.
- Madeira, A., Palrao, T., & Mendes, A. S. (2021). The impact of pandemic crisis on the restaurant business. *Sustainability*, 13(1), 40. <https://doi.org/10.3390/su13010040>
- Möller, M., & Watanabe, M. (2010). Advance purchase discounts versus clearance sales. *The Economic Journal*, 120(547), 1125–1148.
- Montez, J., & Schutz, N. (2021). All-pay oligopolies: Price competition with unobservable inventory choices. *The Review of Economic Studies*, 88(5), 2407–2438.
- Moreno, D., & Ubeda, L. (2006). Capacity precommitment and price competition yield the Cournot outcome. *Games and Economic Behavior*, 56(2), 323–332.
- Nocke, V., & Peitz, M. (2007). A theory of clearance sales. *The Economic Journal*, 117(522), 964–990.
- Parkinson, B., Meacock, R., & Sutton, M. (2019). How do hospitals respond to price changes in emergency departments? *Health Economics*, 28(7), 830–842.
- Somogyi, R. (2016). *Essays on capacity-constrained pricing* (Ph.D. Dissertation). Ecole Polytechnique, Université Paris-Saclay, 2016SACLX024.
- Somogyi, R. (2020). Bertrand-edgeworth competition with substantial horizontal product differentiation. *Mathematical Social Sciences*, 108, 27–37.
- Somogyi, R., Vergote, W., & Virag, G. (2023). Price competition with capacity uncertainty—feasting on leftovers. *Games and Economic Behavior*, 140, 253–271.
- Xiao, B., & Yang, W. (2010). A revenue management model for products with two capacity dimensions. *European Journal of Operational Research*, 205(2), 412–421.

How to cite this article: Somogyi, R. (2024). Monopoly pricing with dual-capacity constraints. *Journal of Economics & Management Strategy*, 33, 155–174. <https://doi.org/10.1111/jems.12556>

APPENDIX A: PROOFS

Proof of Lemma 1. Define

$$f(Q) = \frac{E(q)}{E(q^2)}Q + \frac{1}{2E(q^2)}\alpha(1-\alpha)(v_L q_H - v_H q_L)(q_H - q_L)$$

and

$$g(Q) = \frac{1}{E(q)}\left(Q - \alpha(1-\alpha)(q_H - q_L)\frac{v_H - v_L}{2}\right).$$

Consider the maximization problem from the main text with the corresponding Karush–Kuhn–Tucker multipliers in parentheses:

$$\begin{aligned}
 (\text{P-BL}) \quad \max_{p_L, p_H} \pi &= \alpha(v_H - p_H)p_H + (1-\alpha)(v_L - p_L)p_L \quad \text{s.t.} \\
 \alpha(v_H - p_H) + (1-\alpha)(v_L - p_L) &\leq K, & (\lambda_1) \\
 \alpha(v_H - p_H)q_H + (1-\alpha)(v_L - p_L)q_L &\leq Q, & (\lambda_2) \\
 p_L &\leq v_L, & (\lambda_3) \\
 p_H &\leq v_H, & (\lambda_4) \\
 p_L &\geq 0, p_H &\geq 0.
 \end{aligned}$$

The nonnegativity constraints are omitted and verified ex post. Multiplying the third and fourth constraint by $1 - \alpha$ and α , respectively, the objective function for deriving the Karush–Kuhn–Tucker conditions writes as

$$\begin{aligned}
L(p_L, p_H, \lambda_1, \lambda_2, \lambda_3, \lambda_4) = & \alpha(v_H - p_H)p_H + (1 - \alpha)(v_L - p_L)p_L \\
& - \lambda_1[\alpha(v_H - p_H) + (1 - \alpha)(v_L - p_L) - K] \\
& - \lambda_2[\alpha(v_H - p_H)q_H + (1 - \alpha)(v_L - p_L)q_L - Q] \\
& - \lambda_3(1 - \alpha)(v_L - p_L) - \lambda_4\alpha(v_H - p_H).
\end{aligned}$$

Hence the two first-order conditions are

$$2p_L = v_L + \lambda_1 + \lambda_2 q_L + \lambda_3 \quad \text{and} \quad 2p_H = v_H + \lambda_1 + \lambda_2 q_H + \lambda_4.$$

Case of $\lambda_3 = \lambda_4 = 0$: These conditions correspond to the case of potentially serving both consumer groups. One must distinguish four subcases based on the values of the remaining two KKT multipliers.

Subcase 1 (Region K): $\lambda_1 > 0$ and $\lambda_2 = 0$.

From the FOCs: $\lambda_1 = 2p_L - v_L = 2p_H - v_H$ and $\lambda_1 > 0$ implies that K binds: $\alpha(v_H - p_H) + (1 - \alpha)(v_L - p_L) = K$. The optimal prices can be calculated from these two equations:

$$p_L^K = v_L + \alpha \frac{v_H - v_L}{2} - K \quad \text{and} \quad p_H^K = v_H - (1 - \alpha) \frac{v_H - v_L}{2} - K.$$

The capacity constraint on production rewrites as

$$\alpha \left(K + (1 - \alpha) \frac{v_H - v_L}{2} \right) q_H + (1 - \alpha) \left(K - \alpha \frac{v_H - v_L}{2} \right) q_L \leq Q,$$

which is equivalent to $K \leq g(Q)$ by definition. $\lambda_1 > 0$ implies $K < E(v)/2$ and finally, $p_L \leq v_L$ implies $\frac{\alpha}{2}(v_H - v_L) \leq K$. The nonnegativity constraints and $p_H \leq v_H$ are always satisfied in this optimum.

Subcase 2 (Region Q): $\lambda_1 = 0$ and $\lambda_2 > 0$.

From the FOCs: $\lambda_2 = \frac{2p_L - v_L}{q_L} = \frac{2p_H - v_H}{q_H}$ and $\lambda_2 > 0$ implies that Q binds: $\alpha(v_H - p_H)q_H + (1 - \alpha)(v_L - p_L)q_L = Q$. It is straightforward to calculate the optimal prices from these two equations:

$$p_L^Q = \frac{q_L}{E(q^2)} \left(\frac{\alpha}{2} \left(v_L \frac{q_H}{q_L} + v_H \right) q_H + (1 - \alpha) v_L q_L - Q \right) \quad \text{and} \quad p_H^Q = p_L^Q \frac{q_H}{q_L}.$$

The capacity constraint K must be slack, replacing the optimal prices into the constraint leads to $f(Q) \leq K$. Moreover, $p_H \leq v_H$ implies $\frac{1-\alpha}{2} q_L \left(v_L - v_H \frac{q_L}{q_H} \right) \leq Q$ and finally, $\lambda_2 > 0$ implies $Q < E(vq)/2$. The nonnegativity constraints and $p_L \leq v_L$ are always satisfied.

Subcase 3 (Region KQ): $\lambda_1 > 0$ and $\lambda_2 > 0$.

Both K and Q bind, hence the optimal values of the prices follow directly from these two equations. The four borders of the KQ regions can be calculated as follows. The values of the two multipliers can be calculated from the FOCs by substituting the optimal prices:

$$\begin{aligned}
\lambda_2 = & \frac{v_H - v_L}{q_H - q_L} + 2 \frac{Kq_H - Q}{(1 - \alpha)(q_H - q_L)^2} - 2 \frac{Q - Kq_L}{\alpha(q_H - q_L)^2} \quad \text{and} \\
\lambda_1 = & v_H - 2 \frac{Q - Kq_L}{\alpha(q_H - q_L)} - q_H \left(\frac{v_H - v_L}{q_H - q_L} + 2 \frac{Kq_H - Q}{(1 - \alpha)(q_H - q_L)^2} - 2 \frac{Q - Kq_L}{\alpha(q_H - q_L)^2} \right).
\end{aligned}$$

It follows that $\lambda_1 > 0$ is equivalent to $K < f(Q)$ and $\lambda_2 > 0$ is equivalent to $K > g(Q)$. Furthermore, $p_L \leq v_L$ implies $K \leq Q/q_L$ and $p_H \leq v_H$ implies $K \geq Q/q_H$.

Subcase 4 (Region U): $\lambda_1 = 0$ and $\lambda_2 = 0$.

Neither capacity constraint binds under these conditions. This means that the monopoly can implement its unconstrained optimal prices, that is, $p_L = v_L/2$ and $p_H = v_H/2$. Substituting these values back to the constraints provides the two borders of this region: $K \geq E(v)/2$ and $Q \geq E(vq)/2$.

Excluding one consumer group (regions EL and EH):

The remaining three cases when either λ_3 or λ_4 or both are positive correspond to situations where it is optimal for the firm to exclude one consumer group. Clearly, $\lambda_3 > 0$ and $\lambda_4 > 0$ jointly lead to zero profit and hence can never be never optimal.

As $\lambda_3 > 0$ implies $p_L = v_L$, this case corresponds to excluding the low-types. Then the two capacity constraints can be rewritten as $\alpha(v_H - p_H) \leq \min(K, Q/q_H)$. The firm serves high-types up to the tighter capacity constraint for a price $p_H = v_H - \min(K, Q/q_H)/\alpha$, consequently its profit equals $v_H \min(K, Q/q_H) - \frac{(\min(K, Q/q_H))^2}{\alpha}$. Similar arguments show that for $\lambda_4 > 0$, when the firm excludes high-types, the optimal price is $p_L = v_L - \frac{\min(K, Q/q_L)}{1-\alpha}$ and the firm's profit is $v_L \min(K, Q/q_L) - \frac{(\min(K, Q/q_L))^2}{1-\alpha}$.

In region EL, that is, if $K \leq \min\left(Q/q_H, \frac{\alpha}{2}(v_H - v_L)\right)$ then K must bind, so one must compare the profit of $v_H K - \frac{K^2}{\alpha}$ when excluding low-types with the profit of $v_L K - \frac{K^2}{1-\alpha}$ obtained by excluding the high-types. $K \leq \frac{\alpha}{2}(v_H - v_L)$ implies that the first profit, that is, excluding the low-types is always more profitable in region EL. An analogous argument shows why excluding high-types is more profitable than excluding low-types when $Q \leq \min\left(Kq_L, \frac{1-\alpha}{2}q_L(v_L - v_H \frac{q_L}{q_H})\right)$, that is, in region EH. This concludes the proof of the lemma. \square

Positive, constant marginal cost of production:

Assume an extension of the baseline model where instead of zero marginal costs, producing quantity q costs $c \cdot q$, where $c > 0$ denotes the constant marginal cost of production up to capacity. In other words, producing the goods for low-type and high-type consumers cost cq_L and cq_H , respectively. The (P-BL) maximization problem is modified as follows:

$$\begin{aligned}
 \text{(P-c)} \quad \max_{p_L, p_H} \pi &= \alpha(v_H - p_H)(p_H - cq_H) + (1 - \alpha)(v_L - p_L)(p_L - cq_L) \quad \text{s.t.} \\
 \alpha(v_H - p_H) + (1 - \alpha)(v_L - p_L) &\leq K, & (\lambda_1) \\
 \alpha(v_H - p_H)q_H + (1 - \alpha)(v_L - p_L)q_L &\leq Q, & (\lambda_2) \\
 p_L &\leq v_L, & (\lambda_3) \\
 p_H &\leq v_H, & (\lambda_4) \\
 p_L &\geq cq_L, p_H &\geq cq_H.
 \end{aligned}$$

Focusing on the case where some consumers of both types are served ($\lambda_3 = \lambda_4 = 0$) the new first-order conditions are

$$2p_L = v_L + \lambda_1 + \lambda_2 q_L + cq_L \quad \text{and} \quad 2p_H = v_H + \lambda_1 + \lambda_2 q_H + cq_H.$$

I first show that under a single-capacity constraint with constant marginal cost, optimal prices are monotone decreasing in the level of capacity. Notice that subcase 1 above corresponds exactly to a situation when Q is so large that only one capacity, namely, K binds. Following the same steps as above, the new optimal prices are

$$p_L^K = v_L + \alpha \frac{v_H - v_L}{2} - K - \frac{\alpha}{2}(q_H - q_L)c \quad \text{and} \quad p_H^K = v_H - (1 - \alpha) \frac{v_H - v_L}{2} - K + \frac{1 - \alpha}{2}(q_H - q_L)c.$$

Thus, if c is sufficiently low ($c \leq \min\{p_L^K/q_L; p_H^K/q_H\}$), the optimal prices under positive c are simply shifted by a term proportional to c , independent of K . Therefore, optimal prices are still monotone decreasing in K .

Similarly, subcase 2 corresponds to a situation when only capacity Q binds. It is easy to see that in this case, c cancels out in the first-order conditions, therefore the optimal prices charged by the monopolist remain the same as in

the baseline model as long as $c \leq \min\{p_L^Q/q_L; p_H^Q/q_H\}$. I conclude that optimal prices are monotone decreasing under positive c if there is a single-capacity constraint, no matter whether it is a constraint on the mass of people served or the amount of products sold.

Next I show that the optimal prices are nonmonotonous in the capacities in the KQ region where both capacities bind. It is straightforward to see that the optimal prices are unaffected by the marginal cost whenever it is sufficiently low, that is, $c \leq \min\{p_L^{KQ}/q_L; p_H^{KQ}/q_H\}$. Finally, the existence of the KQ region for sufficiently low c is guaranteed by the continuity of all four functions that delimit the region.

I conclude that the main results of the baseline model hold under small, positive constant marginal cost of production.

Proof of Proposition 2. Substituting the optimal prices p_L and p_H of region KQ into the general formula, the consumer surplus equals

$$\begin{aligned} CS &= \frac{\alpha}{2} \left(\frac{Q - Kq_L}{\alpha(q_H - q_L)} \right)^2 + \frac{1-\alpha}{2} \left(\frac{Kq_H - Q}{(1-\alpha)(q_H - q_L)} \right)^2 \\ &= \frac{1}{2\alpha(1-\alpha)(q_H - q_L)^2} (Q^2 - 2E(q)KQ + E(q^2)K^2). \end{aligned}$$

Therefore the first derivative of the consumer surplus with respect to K and Q are

$$\begin{aligned} \frac{\partial CS}{\partial K} &= \frac{1}{\alpha(1-\alpha)(q_H - q_L)^2} (-E(q)Q + E(q^2)K) \quad \text{and} \\ \frac{\partial CS}{\partial Q} &= \frac{1}{\alpha(1-\alpha)(q_H - q_L)^2} (-E(q)K + Q). \end{aligned}$$

Thus consumer surplus is decreasing in K whenever $K \leq E(q)/E(q^2)Q$ and it is decreasing in Q if $K \geq Q/E(q)$. The two regions delimited by these lines are disjoint since they both cross the origin and the slope of $K = E(q)/E(q^2)Q$ is smaller than the slope of $K = Q/E(q)$ as $(E(q))^2 < E(q^2)$. \square

Proof of Lemma 2. Total welfare is given by the sum of consumer surplus and profit:

$$\begin{aligned} TW = CS + \pi_{KQ} &= \frac{1}{(q_H - q_L)} ((v_H - v_L)Q + (v_L q_H - v_H q_L)K) \\ &\quad - \frac{1}{2\alpha(1-\alpha)(q_H - q_L)^2} (Q^2 - 2E(q)KQ + E(q^2)K^2). \end{aligned}$$

Notice that wherever consumer surplus is decreasing in either K or Q , the third term of the above equation is increasing, so do the first two terms which means that total welfare is always an increasing function of both capacity levels. \square

Proof of Proposition 4. I fully characterize the monopoly's optimal strategy under the general distribution of WTP values in five steps. In Step 1, I show that a core region, KQ' exists,¹⁸ and determine the monopoly's optimal decision in that region. In Step 2, I show that this region is uniquely determined by the primitives of the model and I provide the formula to determine the borders of this region. In Step 3, I describe the U' region and show it is unique. In Step 4, I describe the K' and EL' regions and show they are uniquely determined. Finally, in Step 5 I describe the remaining two regions, that is, Q' and EH' .

The objective function corresponding to (P-GEN) for deriving the Karush–Kuhn–Tucker conditions writes as

$$\begin{aligned} L(N_L, N_H, \lambda_K, \lambda_Q, \lambda_L, \lambda_H) &= \alpha N_H P_H(N_H) + (1-\alpha) N_L P_L(N_L) - \lambda_K [\alpha N_H + (1-\alpha) N_L - K] \\ &\quad - \lambda_Q [\alpha N_H q_H + (1-\alpha) N_L q_L - Q] + (1-\alpha) \lambda_L N_L + \alpha \lambda_H N_H. \end{aligned}$$

Let $MR_i(N_i) = P_i(N_i) + N_i P'_i(N_i)$ denote the marginal revenue of the monopoly from consumer group i . Notice that $MR_i(0) = \theta_i$, moreover, Myerson's regularity condition in Assumption 2 guarantees $MR'_i(N_i) < 0$. Then the two first-order conditions can be written as

$$MR_H(N_H) = \lambda_K + \lambda_Q q_H - \lambda_H \quad \text{and} \quad MR_L(N_L) = \lambda_K + \lambda_Q q_L - \lambda_L.$$

As Step 1, I will now show the existence of the KQ' region, defined as the region where both constraints bind and some consumers of both types are served. This latter property means that $\lambda_L = \lambda_H = 0$ must be satisfied if such a region exists. Moreover, both constraints binding immediately imply that the optimal quantities must satisfy

$$N_L^{KQ} = \frac{Kq_H - Q}{(1 - \alpha)(q_H - q_L)} \quad \text{and} \quad N_H^{KQ} = \frac{Q - Kq_L}{\alpha(q_H - q_L)}.$$

Nonnegativity constraints are then satisfied if and only if $\frac{Q}{q_H} \leq K \leq \frac{Q}{q_L}$. Finally, optimality also requires that the two multipliers on capacity levels be nonnegative:

$$\lambda_Q \geq 0 \Leftrightarrow MR_H(N_H) \geq MR_L(N_L) \quad \text{and} \quad \lambda_K \geq 0 \Leftrightarrow \frac{MR_L(N_L)}{q_L} \geq \frac{MR_H(N_H)}{q_H}.$$

At $K = Q = 0$, even the strict versions of these inequalities are satisfied by Assumption 2. Indeed, $K = Q = 0$ clearly imply $N_i = 0$ and $MR_i(0) = \theta_i$. Therefore, by continuity of the marginal revenue functions, there must be a neighborhood of $(0,0)$ where all four conditions are satisfied, proving the existence of a nonempty KQ' region.

In Step 2, I will now show that all curves bordering this region are uniquely determined by the primitives of the model. This is trivial for the lines $\frac{Q}{q_H}$ and $\frac{Q}{q_L}$. To be in line with the notation of the baseline model, let the curve $K = g(Q)$ denote the border of the KQ' region where constraint Q starts to be slack and thus $\lambda_Q = 0$. From the inequalities above, we know that this curve is defined by the points that satisfy $MR_H(N_H^{KQ}) = MR_L(N_L^{KQ})$.

To show the existence and uniqueness of $g(Q)$, take any point (Q_1, K_1) in the inside of KQ' and start increasing Q . At (Q_1, K_1) by construction we have $MR_H(N_H^{KQ}) > MR_L(N_L^{KQ})$. Next, I will show that $MR_H(N_H^{KQ})$ is monotone decreasing in Q , whereas $MR_L(N_L^{KQ})$ is monotone increasing in Q , therefore they must cross exactly once, proving the uniqueness. Indeed,

$$\frac{\partial MR_H(N_H^{KQ})}{\partial Q} = \frac{MR'_H(N_H^{KQ})}{\alpha(q_H - q_L)} < 0 \quad \text{and} \quad \frac{\partial MR_L(N_L^{KQ})}{\partial Q} = -\frac{MR'_L(N_L^{KQ})}{(1 - \alpha)(q_H - q_L)} > 0,$$

where the inequalities are a result of the decreasing marginal revenue curves, that is, Assumption 2. Similarly, let the curve $K = f(Q)$ denote the border of the KQ' region where constraint K starts to be slack and thus $\lambda_K = 0$. From above, we know that this curve is defined by the points that satisfy $MR_H(N_H^{KQ})/q_H = MR_L(N_L^{KQ})/q_L$. Analogous arguments to the uniqueness of $g(Q)$ also show the uniqueness of $f(Q)$. Indeed, $MR_H(N_H^{KQ})/q_H < MR_L(N_L^{KQ})/q_L$ in the interior of KQ' , moreover $MR_H(N_H^{KQ})/q_H$ is increasing in K whereas $MR_L(N_L^{KQ})/q_L$ is decreasing in K , thus they must cross exactly once.

As Step 3, I will prove that the $f(Q)$ and $g(Q)$ curves cross exactly once, moreover, this intersection point determines the borders of the U' region. For this, I use the implicit function theorem for both functions to determine their slopes. By definition, at $K = g(Q)$

$$\xi(K, Q) = MR_H\left(\frac{Q - Kq_L}{\alpha(q_H - q_L)}\right) - MR_L\left(\frac{Kq_H - Q}{(1 - \alpha)(q_H - q_L)}\right) = 0.$$

By the implicit function theorem and some straightforward transformations,

$$g'(Q) = \frac{\partial K}{\partial Q} = -\frac{\partial \xi / \partial Q}{\partial \xi / \partial K} = \frac{(1 - \alpha)MR'_H(N_H^{KQ}) + \alpha MR'_L(N_L^{KQ})}{(1 - \alpha)q_L MR'_H(N_H^{KQ}) + \alpha q_H MR'_L(N_L^{KQ})},$$

so

$$\begin{aligned} g'(Q) &= \frac{(1 - \alpha)MR'_H(N_H^{KQ}) + \alpha MR'_L(N_L^{KQ})}{(1 - \alpha)q_L MR'_H(N_H^{KQ}) + \alpha q_H MR'_L(N_L^{KQ})} \\ &> \frac{(1 - \alpha)MR'_H(N_H^{KQ}) + \alpha MR'_L(N_L^{KQ})}{(1 - \alpha)q_H MR'_H(N_H^{KQ}) + \alpha q_H MR'_L(N_L^{KQ})} = \frac{1}{q_H}, \end{aligned}$$

where the inequality comes about because $q_H > q_L$ and the numerator of the fractions and $MR'_H(N_H^{KQ})$ are strictly negative. Analogous steps show that the derivative of $f(Q)$ is given by

$$f'(Q) = \frac{(1 - \alpha)q_L MR'_H(N_H^{KQ}) + \alpha q_H MR'_L(N_L^{KQ})}{(1 - \alpha)q_L^2 MR'_H(N_H^{KQ}) + \alpha q_H^2 MR'_L(N_L^{KQ})},$$

and that this ratio is between 0 and Q/q_L . Therefore, we now know that both of the curves are monotone increasing in Q , moreover, the slope of $g(Q)$ is strictly larger. From Step 1, we know that for sufficiently small Q we must have $f(Q) > g(Q)$ so this must hold for all Q up their unique intersection point. Let (\bar{Q}, \bar{K}) denote this intersection. By definition, at this point $MR'_H(N_H^{KQ}) = MR'_L(N_L^{KQ})$ and $MR'_H(N_H^{KQ})/q_H = MR'_L(N_L^{KQ})/q_L$ which imply $MR'_H(N_H^{KQ}) = MR'_L(N_L^{KQ}) = 0$. In other words, the monopoly can implement its unconstrained maximal prices and quantities at (\bar{Q}, \bar{K}) .

Indeed, this point corresponds to the solution of the optimization problem where none of the constraints bind, that is, $\lambda_K = \lambda_Q = 0$. By optimality, (\bar{Q}, \bar{K}) is the lowest capacity-pair that admits the unconstrained optimal quantities, N_H^U and N_L^U , given implicitly by the zero marginal revenues. Clearly, any pair of capacities satisfying $Q \geq \bar{Q}$ and $K \geq \bar{K}$ also allows the unconstrained optimum, hence (\bar{Q}, \bar{K}) indeed provide the borders of the U' region, as in the baseline model.

At this point, I finish showing that the KQ' region is qualitatively similar to the KQ region by showing that $\frac{\bar{Q}}{q_H} < \bar{K} < \frac{\bar{Q}}{q_L}$. Together with the fact that the slope of $f(Q)$ is smaller than the slope of Q/q_L and the slope of $g(Q)$ is larger than the slope of Q/q_H , this implies that KQ' is indeed delimited by four curves of different slopes as in the baseline model. To get a contradiction for the first inequality, assume $\frac{\bar{Q}}{q_H} \geq \bar{K}$. Then the point $(\bar{Q}, \frac{\bar{Q}}{q_H})$ is in the U' region where none of the constraints bind. In particular, it must be that $\alpha N_H^U + (1 - \alpha)N_L^U \leq \frac{\bar{Q}}{q_H}$, but that implies the following contradiction:

$$\alpha q_H N_H^U + (1 - \alpha)q_H N_L^U \leq \bar{Q} = \alpha q_H N_H^U + (1 - \alpha)q_L N_L^U.$$

A similar argument ensures $\bar{K} \leq \frac{\bar{Q}}{q_L}$.

Next, in Step 4, I describe regions K' and EL' . Region K' corresponds to the solution of the optimization problem when $\lambda_Q = \lambda_H = \lambda_N = 0$ and $\lambda_K \geq 0$. K binding implies that optimal quantities N_H^K, N_L^K in this region satisfy $\alpha N_H^K + (1 - \alpha)N_L^K = K$. First-order conditions imply that $\lambda_K = MR'_L(N_L^K) = MR'_H(N_H^K) > 0$. One can use the same argument as in Step 2 to show that one of the marginal revenues is increasing in K , the other is decreasing in K , proving that N_H^K, N_L^K are unique.

By the definition of $g(Q)$ and \bar{K} , region K' consists of capacity-pairs that satisfy $K \leq \min\{g(Q), \bar{K}\}$. The last border of the region will be given by the constraint $N_L \geq 0$. Using the implicit function theorem on $MR_L(N_L^K) - MR_H\left(\frac{1}{\alpha}(K - (1 - \alpha)N_L^K)\right) = 0$, one gets

$$\frac{\partial N_L^K}{\partial K} = \frac{MR'_H\left(\frac{1}{\alpha}(K - (1 - \alpha)N_L^K)\right)}{\alpha MR'_L(N_L^K) + (1 - \alpha)MR'_H\left(\frac{1}{\alpha}(K - (1 - \alpha)N_L^K)\right)} > 0.$$

Therefore there must be a positive $\hat{K} > 0$ such that $N_L^K(\hat{K}) = 0$, otherwise at $K = 0$ one would get the following contradiction: $\theta_L = MR_L(0) \geq MR_H(0) = \theta_H$. Clearly, \hat{K} is implicitly given by $MR_H\left(\frac{\hat{K}}{\alpha}\right) = \theta_L$.

The monopoly prefers excluding low-types below this \hat{K} threshold, that is, it is region EL' . Indeed, this region corresponds to the optimal solutions of the optimization problem where $\lambda_L > 0$ and $\lambda_K > 0$. Then the optimal quantity is simply $N_H^{EL} = K/\alpha$, and the second border of the region is given by the capacity constraint Q : $K \leq Q/q_H$.

Step 5 is analogous to Step 4. In region Q' , following the same steps as before, the optimal quantities are uniquely determined by

$$\frac{MR_L\left(\frac{Q - \alpha q_H N_H^Q}{(1 - \alpha)q_L}\right)}{q_L} = \frac{MR_H(N_H^Q)}{q_H}.$$

Moreover, the capacity level \hat{Q} satisfying $\theta_H/q_H = MR_L\left(\frac{\hat{Q}}{1 - \alpha}\right)/q_L$ can be shown to be the border between the Q' and the EH' region. Finally, the optimal quantity in the EH' region is $N_L^{EH} = \frac{Q}{1 - \alpha}$ and the second border of this region is given by $K \geq Q/q_L$. This concludes the proof of Proposition 4. \square

Proof of Proposition 5. Consumer surplus being additive in the consumer surplus of the consumer groups,

$$\frac{\partial CS}{\partial K} = \frac{\partial p_H}{\partial K} \frac{\partial}{\partial p_H} \int_{p_H}^{\theta_H} \alpha(w - p_H) f_H(w) dw + \frac{\partial p_L}{\partial K} \frac{\partial}{\partial p_L} \int_{p_L}^{\theta_L} (1 - \alpha)(w - p_L) f_L(w) dw.$$

Using the Leibniz-rule, it follows that

$$\frac{\partial CS}{\partial K} = \frac{q_H}{(1 - \alpha)(q_H - q_L)D'_L} (-(1 - \alpha)D_L) + \frac{-q_L}{\alpha(q_H - q_L)D'_H} (-\alpha D_H)$$

which implies that

$$\frac{\partial CS}{\partial K} < 0 \Leftrightarrow q_L \frac{D'_L(p_L)}{D_L(p_L)} < q_H \frac{D'_H(p_H)}{D_H(p_H)} \Leftrightarrow \frac{\varepsilon_L(p_L)}{\varepsilon_H(p_H)} > \frac{p_L/q_L}{p_H/q_H},$$

where the last equivalence uses the definition of price-elasticities of demand, that is, $\varepsilon_i(p_i) = \frac{D'_i(p_i)}{D_i(p_i)} p_i$ and the fact that they are negative. Analogous steps reveal $\frac{\partial CS}{\partial Q} < 0 \Leftrightarrow \frac{\varepsilon_L(p_L)}{\varepsilon_H(p_H)} < \frac{p_L}{p_H}$.

It follows from Proposition 4 that part of the Q/q_L and Q/q_H lines border the KQ' region. From the demand functions, it is straightforward that $\frac{\partial CS}{\partial Q} < 0$ is satisfied for $K = Q/q_L$ and $\frac{\partial CS}{\partial K} < 0$ holds for $K = Q/q_H$. The existence of the two regions can be proved by continuity arguments. Finally, the two regions must be disjoint otherwise any common point would satisfy $\frac{p_L/q_L}{p_H/q_H} < \frac{p_L}{p_H}$, a contradiction. This concludes the proof of the proposition. \square