Antonella Nocco, Gianmarco I.P. Ottaviano, Matteo Salto

Monopolistic competition and optimum product selection: why and how heterogeneity matters

Article (Accepted version) (Refereed)

Monopolistic Competition and Optimum Product Selection: Why and how heterogeneity matters*

Antonella Nocco, Università del Salento
Gianmarco I.P. Ottaviano, LSE and University of Bologna
Matteo Salto, European Commission

This version: July 2017

Abstract

After some decades of relative oblivion, the interest in the optimality properties of monopolistic competition has recently re-emerged due to the availability of an appropriate and parsimonious framework to deal with firm heterogeneity. Within this framework we show that non-separable utility, variable demand elasticity and endogenous firm heterogeneity cause the market equilibrium to err in many ways, concerning the number of products, the size and the choice of producers, the overall size of the monopolistically competitive sector. More crucially with respect to the existing literature, we also show that the extent of the errors depends on the degree of firm heterogeneity. In particular, the inefficiency of the market equilibrium is largest when selection among heterogenous firms is needed most, that is, when there are relatively many firms with low productivity and relatively few firms with high productivity.

Keywords: monopolistic competition, product diversity, firm heterogeneity, selection, welfare.

J.E.L. Classification: D4, D6, F1, L0, L1.

*For useful comments and suggestions we are grateful to the editor Federico Etro as well as to Paolo Bertoletti, Olivier Biau, Matthieu Lequien, Marc Melitz, John Morrow, Peter Neary, Mathieu Parenti, Jacques Thisse, Evgeny Zhelobodko and seminar participants at various institutions, conferences and workshops. The views expressed here are those of the authors and do not represent in any manner the EU Commission.

1 Introduction

Do monopolistically competitive industries yield an optimal level of product diversity? As discussed by Neary (2004), this 'classic issue' in industrial organization motivated the canonical formalization of the Chamberlinian model (Chamberlin, 1933) as put forth by Spence (1976) and Dixit and Stiglitz (1977). These scholars propose 'reduced form' models that "regard aggregate demands as if they result from the maximisation of a utility function defined directly over the quantities of goods, and the form of the utility function is intended to capture the desire for variety" (Dixit, 2004, p.125).[1] The classic issue can be itself split into four questions concerning the optimality of the market outcome (Stiglitz, 1975): Are there too few or too many products? Are the quantities of the products too small or too large? Are the products supplied by the right set of firms, or are there 'errors' in the choice of technique? Are monopolistically competitive industries too large or too small with respect to the rest of the economy?

The Chamberlinian model makes four basic assumptions (Bishop, 1967; Brakman and Heijdra, 2004): the number of sellers in a group of firms is sufficiently large so that each firm takes the behavior of other firms in the group as given; the group is well defined and small relative to the economy; products are physically similar but economically differentiated so that buyers have preferences for all types of products ('love for variety'); there is free entry. In this setup, optimality rests on how the market mechanism deals with the crucial tradeoff of 'efficiency versus diversity' (Kaldor, 1934).

As forcefully highlighted by Dixit and Stiglitz (1975), there are good reasons to doubt that the market will generally strike the right balance due to the public nature of diversity in the reduced form approach. As in these models the range of products enters utility as a direct argument in addition to the quantities consumed, the range itself becomes a public good whose social benefit is not fully reflected in private incentives. In the words of Spence (1976, pp. 230-231):

> "[T]here are conflicting forces at work with respect to the number or variety of products. Because of setup costs, revenues may fail to cover the costs of a socially desirable product. As a result, some products may be produced at a loss at an optimum. This is a force tending towards too few products. On the other hand, there are forces tending towards too many products. First, because firms hold back output and keep price above marginal cost, they leave more room for entry than would marginal cost pricing. Second, when a firm enters with a new product, it adds its own consumer and producer surplus to the total surplus, but it also cuts into the profits of the existing firms. If the cross elasticities of demand are high, the dominant effect may be the second one. In this case entry does not increase the size of the pie much; it just divides it into more pieces. Thus, in the presence of high cross elasticities of demand, there is a tendency toward too many products".

[1] 'Structural' models, instead, "give an explicit model of a consumer's choice where diversity plays a role; discrete choice from a collection of products differentiated by location in a characteristic space in the most common framework" (Dixit, 2004, p.125). See Anderson, de Palma and Thisse (1992) for microfoundations of the representative-consumer reduced form approach based on random-utility models of discrete choice.

As the issue of optimal product diversity does not admit a general settlement, explicit models with a detailed formulation of demand are used to isolate and analyze the four questions described above. The canonical choice is to model an economy consisting of two sectors. The first sector is monopolistically competitive and is the focus of the analysis. The second sector is perfectly competitive and represents the rest of the economy. Its purpose is to hold factor prices in check and to create the slack needed to answer the question whether the monopolistically competitive sector is too small or too big. This way the market is allowed to eventually misallocate resources not only within the monopolistically competitive sector but also between this sector and the rest of the economy.

The best known insights of the canonical model concern the special case in which the 'group utility' defined over differentiated products is separable across them, the demand of each product is CES and firms are homogeneous. In this case, the model shows that the first-best ('unconstrained') optimum calls for more product variety than the market provides. From a normative perspective, however, this result is traditionally regarded of little practical relevance for policy intervention because implementing the unconstrained optimum requires the use of lump-sum instruments that are hardly available in reality. These are needed to subsidize the entry of firms that otherwise would not cover their setup ('entry') costs due to marginal cost pricing at the optimum. A lot of attention has, therefore, been devoted to the 'constrained' optimum in which the monopolistically competitive sector is financially self-sufficient. Under this constraint, the market is shown to provide the optimal number of products, the optimal size of firms and hence the optimal size of the sector.

The robustness of these results has been investigated along several dimensions, with particular attention devoted to the impact of variable demand elasticity and firm heterogeneity. These extensions are already discussed by Stiglitz (1975), Spence (1976) and Dixit and Stiglitz (1977), who show that, when the elasticity of demand is allowed to vary, the market equilibrium ceases to be constrained optimal. In particular, products are too many (too few) and are supplied in too small (too large) quantities when the elasticity of 'product utility' is increasing (decreasing) in the quantity consumed. As for firm heterogeneity, Dixit and Stiglitz (1977) consider a variant of their model in which there are two groups of differentiated products that are perfect substitutes for each other with each group having CES sub-utility. Both fixed and marginal costs are allowed to differ between the two groups but not within them. Dixit and Stiglitz (1977) use this variant to show that the determination of the set of products to be supplied depends on a richer list of factors: fixed and marginal costs, the elasticity of the demand schedule, the level of the demand schedule and the cross-elasticities of demand. As a result, constrained optimality eventually applies only to a zero-measure set of parametrizations. A more exhaustive treatment of this issue can be found in Spence (1976), while Stiglitz (1975) reaches similar conclusions in a model of the capital market in which firms with heterogeneous costs issue securities whose returns are imperfectly correlated with each other.

Some decades of relative oblivion followed, with relatively rare contributions in the theory of monopolistic competition. An exception is Pascoa (1997), which first discussed classes of models with heterogeneous goods and firms to show the relative role of demand elasticity and increasing returns to scale in generating monopolistically competitive equilibria. But a diffuse interest in the optimality properties of monopolistic competition has re-emerged only recently due to the 'heterogeneous firms revolution' in international trade theory (Melitz and Redding, 2012). This has been initiated by Melitz (2003), who shows that a Dixit-Stiglitz model with CES demand, endogenous firm heterogeneity and fixed export costs (but without the homogeneous good sector) predicts 'new' gains from trade liberalization through the selection of the most efficient firms.[2] Subsequent papers show that a similar result holds when demand exhibits variable elasticity, though fixed export costs are not necessarily needed for the result to materialize in this case (Melitz and Ottaviano, 2008; Behrens and Murata, 2012). Bertoletti, Etro and Simonovska (2017) show that with firm heterogeneity, indirect additivity generates lower gains from trade than under CES, with the difference linked to the average pass-through.

The validity of these (among other) insights on international trade issues when alternative specifications of demands are allowed for is discussed by Zhelobodko, Kokovin, Parenti, and Thisse (2012). Using a framework with variable elasticity of substitution (VES), they show that CES is just a knife-edge case. While this finding is reminiscent of the conclusions by Stiglitz (1975), Spence (1976) and Dixit and Stiglitz (1977), Zhelobodko, Kokovin, Parenti, and Thisse (2012) do not discuss its implications for optimum product variety as those early contributors do. This is done, instead, by Dhingra and Morrow (2016) who fully characterize the optimality properties of a general demand system derived from separable 'group utility'. Their normative analysis thus complements the positive analysis of Zhelobodko, Kokovin, Parenti, and Thisse (2012), showing that, in the absence of the homogeneous sector, the market outcome achieves the (unconstrained) optimum under CES but not under VES. When a homogeneous sector is instead introduced, Melitz and Redding (2012) show that CES leads to constrained rather than unconstrained optimality due to the misallocation of resources between sectors. In other words, with CES firm heterogeneity does not change the welfare insights of the original Dixit-Stiglitz framework while things change in the case of VES.

The present paper goes back to the full set of classic questions laid down at the beginning of this introduction, with renewed emphasis on the question whether in the market equilibrium the products are supplied by the right set of firms. As Nocco, Ottaviano and Salto (2014), the paper does so in a Melitzian framework of endogenous firm heterogeneity with variable demand elasticity *à la* Melitz and Ottaviano (2008). However, it differs from Nocco, Ottaviano and Salto (2014) in terms of focus. The main of purpose of Nocco, Ottaviano and Salto (2014) is to highlight that, with variable demand elasticity and endogenous firm heterogeneity, the market outcome errs in several ways, with respect to the number of products, the size and the choice of producers, and the overall size of the monopolistically competitive sector. Differently and crucially with respect to the existing literature, the main purpose of the present paper is to also show that the extent of those errors depends on the degree of firm heterogeneity.

Apart from Nocco, Ottaviano and Salto (2014), none of the papers previously cited simultaneously addresses the four classic questions on the optimality of monopolistic competition in a framework with variable demand elasticity and endogenous firm heterogeneity. Moreover, none of them, including Nocco, Ottaviano and Salto (2014), provides a systematic quantitative analysis of the

[2] See Arkolakis, Costinot and Rodriguez-Clare (2010) as well as Melitz and Redding (2012) for a discussion of the actual novelty of these findings.

impact of different degrees of firm heterogeneity on the extent of market inefficiencies. The discussion in Spence (1976) is systematic but qualitative, while Dixit and Stiglitz (1977) confine themselves to the special scenario discussed above. Ottaviano and Thisse (1999) characterize the social planner outcome with homogeneous firms within the same framework with variable demand elasticity adopted here. Bertoletti and Etro (2016) provide the first characterization of the social planner problem for any symmetric preferences comparing it with the corresponding market equilibrium in the case of homogeneous firms. Neither contribution considers firm heterogeneity. Dhingra and Morrow (2016) are closer to what the present paper tries to achieve but the focus of their comparative statics is on the parametrization of demand rather than on the parametrization of firm heterogeneity. In addition, not having the homogeneous good sector prevents them from discussing between-sector misallocation. Differently, Stiglitz (1975) presents comparative statics results on the heterogeneity parameters but his heterogeneity is not endogenous and his approach, based on a utility defined over alternative portfolios of assets, is quite distinct from the canonical model of monopolistic competition.

Clearly, as pointed out by Stiglitz (1975) and others, without some appropriate parametrization of the problem, it would be hard to cut any new ground on the issues of interest. As Nocco, Ottaviano and Salto (2014), the present paper relies on the specific parametrization of linear demand introduced by Ottaviano, Tabuchi and Thisse (2002) as applied to endogenous firm heterogeneity by Melitz and Ottaviano (2008). This parametrization is less general than the VES systems studied by Dhingra and Morrow (2016) and Zhelobodko, Kokovin, Parenti, and Thisse (2012) in terms of product utility but allows for cross-product effects that are absent in the former paper and only touched upon in the latter. Moreover, differently from Dhingra and Morrow (2016) and Zhelobodko, Kokovin, Parenti, and Thisse (2012), we follow Melitz and Ottaviano (2008) also in parametrizing heterogeneity in terms of a Pareto distribution of firm productivity. Within this framework we are able to revisit all the aforementioned four questions on the optimality of the market outcome emphasizing how the extent of firm heterogeneity affects the answers. We do so in the wake of Maignan, Ottaviano, Pinelli and Rullani (2003) and Ottaviano (2012) associating the scale and shape parameters of the Pareto productivity distribution with the dimensions of 'cost-increasing richness' and 'cost-decreasing evenness' of firm heterogeneity.

Our exposition is organized around the comparison of results obtained with and without firm heterogeneity in the equilibrium and in the first best ('unconstrained') optimum as this allows us to encompass the findings of Ottaviano and Thisse (1999) in a similar setup with homogeneous firms. First, we show that the total output of the monopolistically competitive sector in the market equilibrium is smaller than optimal with and without firm heterogeneity. With heterogeneous firms, more cost-decreasing evenness decreases (increases) the gap in the total output of the differentiated varieties between the market equilibrium and the optimum when evenness is initially limited (pronounced). More cost-increasing richness always increases the gap.

Second, we find that, while with homogeneous firms the output of each firm in the market equilibrium is inefficiently small, with firm heterogeneity this holds true only for the average output per firm: the equilibrium quantities produced by less productive firms are inefficiently large whereas those supplied by more productive firms are inefficienty small. More cost-decreasing evenness makes the overprovision of varieties relatively more likely than its underprovision in the market equilibrium. Cost-increasing richness has no impact on this.

Third, we show that the market provides an inefficient number of varieties, both when firms are homogeneous and when they are heterogeneous. In both cases the number of varieties supplied in the market equilibrium can either be richer or poorer than optimal depending on the parameters of the model. In particular, product variety will be inefficiently rich for large market size, high substitutability between varieties, low entry costs and low unit costs of production. Less cost-decreasing evenness and more cost-increasing richness makes the under-provision of variety relatively more likely than its over-provision in the market equilibrium.

Fourth and last, obviously the issue of selection among firms producing with different techniques can only be analyzed when firms are heterogeneous, as with homogeneity all firms (or none) produce in equilibrium. We find that products present in the market are not supplied by the right set of firms as selection in the market equilibrium is weaker than optimal and there are varieties supplied by low productivity firms that would not be supplied in the optimum. More cost-decreasing evenness increases the gap in selection between the market equilibrium and the optimum. Cost-increasing richness has no impact on this gap.

Through all these channels, the degree of heterogeneity has a multifaceted impact on the degree of inefficiency of the market equilibrium. Analyzing overall welfare, the key insight is that the inefficiency of the market equilibrium is largest when selection is needed most, that is, when there are a lot of low productivity firms and few high productivity ones. We show that this conclusion holds also when policy makers lack the tools needed to implement the first best ('unconstrained') optimum and have to settle for the second best ('constrained') optimum in which they cannot affect firms' pricing and selection but can only target product variety as in the traditional literature with homogeneous firms.

The rest of the paper is organized in six sections. Section 2 briefly presents the model by Melitz and Ottaviano (2008). Sections 3 and 4, respectively, derive and compare the market equilibrium and the first best ('unconstrained') optimum, focusing first on the case in which firms are homogeneous as in Ottaviano and Thisse (1999), and then on the case of heterogeneous firms as in Nocco, Ottaviano and Salto (2014). Section 5 investigates how the extent of firm heterogeneity affects the gap between the equilibrium and first best optimum outcomes. Section 6 discusses the second best ('constrained') optimum where policy makers may not use lump sum trasfers for firms. Section 7 concludes.

2 The Model

Following Melitz and Ottaviano (2008) and Nocco, Ottaviano and Salto (2014), consider an economy populated by L consumers, each endowed with one unit of labor. Preferences are defined over a continuum of differentiated varieties indexed $i \in \Omega$, and a homogeneous good indexed 0. All consumers own the same initial endowment $\overline{q}_0$ of this good and share the same utility function given by

$$U = q_0^c + \alpha \int_{i\in\Omega} q_i^c di - \frac{1}{2}\gamma \int_{i\in\Omega} (q_i^c)^2\, di - \frac{1}{2}\eta \left(\int_{i\in\Omega} q_i^c di \right)^2 \tag{1}$$

with positive demand parameters α, η and γ, the latter measuring 'love for variety' and the others measuring the preference for the differentiated varieties with respect to the homogeneous good.[3] The initial endowment $\overline{q}_0$ of the homogeneous good is assumed to be large enough for its consumption to be strictly positive at the equilibrium and optimal outcomes.

Labor is the only input. It can be employed for the production of the homogeneous good under constant returns to scale with unit labor requirement equal to one. It can also be employed for the production of the differentiated varieties. In this case the technology requires a preliminary R&D effort of $f > 0$ units of labor to design a new variety and its production process, which is also characterized by constant returns to scale. The R&D effort leads to the design of the new variety with certainty whereas the unit labor requirement c of the corresponding production process is uncertain, being randomly drawn from a continuous distribution with cumulative density $G(c)$ over support $[0, c_M]$ The R&D effort cannot be recovered and is therefore a sunk labor requirement.

3 Equilibrium and Unconstrained Optimum

3.1 Market Outcome

In the market equilibrium consumers maximize utility under their budget constraints, firms maximize profits given their technological constraints, and markets clear. It is assumed that the labor market as well as the market of the homogeneous good are perfectly competitive. This good is chosen as numeraire, which under perfect competition implies that the wage equals one given its unit labor requirement. The market of differentiated varieties is, instead, monopolistically competitive with a one-to-one relation between firms and varieties.

The first order conditions for utility maximization under the budget constraint gives individual inverse demand for variety i as

$$p_i = \alpha - \gamma q_i^c - \eta Q^c \tag{2}$$

for $q_i^c > 0$, where $Q^c = \int_{i\in\Omega} q_i^c di$ is total individual quantity consumed of the differentiated good. Aggregate demand for variety i can be derived from (2) as

$$q_i \equiv L q_i^c = \frac{\alpha L}{\eta N + \gamma} - \frac{L}{\gamma} p_i + \frac{\eta N}{\eta N + \gamma} \frac{L}{\gamma} \bar{p},\ \forall i \in \Omega^* \tag{3}$$

where the set Ω^* is the largest subset of Ω such that demand is positive, N is the measure ('number') of varieties in Ω^* and $\bar{p} = (1/N) \int_{i\in\Omega^*} p_i di$ is their

[3] The demand system exhibits 'love of variety' because, holding the distribution of prices constant (namely holding the mean $\bar{p}$ and variance σ_p^2 of prices constant), utility rises with product variety N. Melitz and Ottaviano (2008) show that rewriting the indirect utility function in terms of average price and price variance reveals that it decreases with average prices $\bar{p}$, but rises with the variance of prices σ_p^2 (holding $\bar{p}$ constant), as consumers then re-optimize their purchases by shifting expenditures towards lower priced varieties as well as the numeraire good.

average price. Variety i belongs to this set if its price satisfies

$$p_i \leq \frac{1}{\eta N + \gamma} (\gamma\alpha + \eta N\bar{p}) \equiv p_{\max} \tag{4}$$

where $p_{\max} \leq \alpha$ is the price threshold at which demand for a variety is driven to zero.

When a variety is produced by a firm with unit labor requirement c, the corresponding first order conditions for profit maximization are satisfied by an output level equal to

$$q^m(c) = \begin{cases} \frac{L}{2\gamma}(c^m - c) & c \leq c^m = \alpha - \frac{\eta}{L}Q^m \\ 0 & c > c^m \end{cases} \tag{5}$$

where 'm' labels equilibrium variable, $c^m = p_{\max}$ and $Q^m \equiv N_E^m \int_0^{c_M} q^m(c)dG(c) = N^m \int_0^{c^m} q^m(c)dG^m(c)$ is the total supply of differentiated varieties with N_E^m and N^m respectively denoting the number of firms entering the market and the number of them eventually producing. This is also the number of varieties supplied and is related to the number of entrants by $N^m = N_E^m G(c^m)$ where $G^m(c) = G(c)/G(c^m)$ is the conditional distribution of unit input requirements for varieties supplied. Expression (5) defines a cutoff rule for survival: only entrants that are productive enough ($c \leq c^m$) eventually produce. For them the price that corresponds to the profit-maximizing output $q^m(c)$ is $p^m(c) = (c^m + c)/2$, implying markup $\mu^m(c) = p^m(c) - c = (c^m - c)/2$ and maximized profit

$$\pi(c) = \frac{L}{4\gamma}(c^m - c)^2. \tag{6}$$

Due to free entry and exit, in equilibrium expected profit is exactly offset by the sunk entry cost stemming from the sunk labor requirement

$$\int_0^{c^m} \pi(c)dG(c) = f.$$

Given (6), this 'free entry condition' can be rewritten as

$$\frac{1}{4}\int_0^{c^m} (c^m - c)^2 dG(c) = \frac{\gamma f}{L}. \tag{7}$$

Finally, the number of producers can be determined as a function of c^m by observing that marginal firms with unit labor requirement $c = c^m$ make zero profit, i.e. $p(c^m) = c^m = p_{\max}$. Recalling (4), this implies the following 'zero cutoff profit condition'

$$c^m = \frac{1}{\eta N^m + \gamma}(\gamma\alpha + \eta N^m \bar{p}^m) \tag{8}$$

where, due to the law of large numbers, $\bar{p}^m$ is the *ex ante* expected price conditional on producing as well as the *ex post* average price of producers: $\bar{p}^m = \int_0^{c^m} p(c)dG^m(c)$. The zero cutoff profit condition (8) can be solved to obtain the equilibrium number of producers (and varieties produced) as a function of the equilibrium cutoff

$$N^m = \frac{2\gamma}{\eta}\frac{\alpha - c^m}{c^m - \bar{c}^m} \tag{9}$$

with $\bar{c}^m = \left[\int_0^{c^m} c dG^m(c)\right]$. The number of entrants is then given by $N_E^m = N^m/G(c^m)$.

3.2 Optimal Outcome

As the quasi-linearity of (1) implies transferable utility, social welfare may be expressed as the sum of all consumers' utilities. This implies that a benevolent planner chooses the number of varieties and their output levels so as to maximize the social welfare function given by individual utility (1) times the number of consumers L, subject to the resource constraint, the homogenous good's production function, the varieties' production functions and the stochastic 'innovation production function' (i.e. the mechanism that determines each variety's unit labor requirement as a random draw from $G(c)$ after f units of labor have been allocated to R&D).

Specifically, given (1), the planner chooses the number N_E of R&D projects and the output levels of associated varieties so as to maximize social welfare

$$\begin{array}{c} W = q_0^c L + \alpha N_E \int_0^{c_M} \left[q^c(c)L\right] dG(c) - \frac{1}{2}\frac{\gamma}{L} N_E \int_0^{c_M} \left[q^c(c)L\right]^2 dG(c) \\ -\frac{1}{2}\frac{\eta}{L}\left[N_E \int_0^{c_M} \left[q^c(c)L\right] dG(c)\right]^2 \end{array} \tag{10}$$

with respect to q_0^c, $q^c(c)$ and N_E subject to the aggregate resource constraint

$$q_0^c L + f N_E + N_E \int_0^{c_M} c q^c(c) L dG(c) = L + \bar{q}_0 L \tag{11}$$

whereby the supply of the homogeneous good ($q_0^c L$), the supply of differentiated varieties ($N_E \int_0^{c_M} c q^c(c) L dG(c)$) and the R&D investment ($f N_E$) are constrained by the amount of the available endowments of labor (L) and homogenous good ($\bar{q}_0$).

After substituting (11) into (10), the planner's problem can be rewritten as the maximization of

$$\begin{array}{c} W = L + \bar{q}_0 L - f N_E + N_E \int_0^{c_M} (\alpha - c)\, q(c) dG(c) \\ -\frac{1}{2}\frac{\gamma}{L} N_E \int_0^{c_M} \left[q(c)\right]^2 dG(c) - \frac{1}{2}\frac{\eta}{L}\left[N_E \int_0^{c_M} q(c) dG(c)\right]^2 \end{array} \tag{12}$$

with respect to $q(c)$ and N_E. The corresponding first order conditions are then

$$\frac{\partial W}{\partial q(c)} = \left[N_E (\alpha - c) - \frac{\gamma}{L} N_E q(c) - \frac{\eta}{L} (N_E)^2 \int_0^{c_M} q(c) dG(c)\right] dG(c) = 0 \tag{13}$$

for all values of c and

$$\begin{array}{c} \frac{\partial W}{\partial N_E} = -f + \int_0^{c_M} (\alpha - c)\, q(c) dG(c) - \frac{1}{2}\frac{\gamma}{L} \int_0^{c_M} \left[q(c)\right]^2 dG(c) \\ -\frac{\eta}{L} N_E \left[\int_0^{c_M} q(c) dG(c)\right]^2 = 0. \end{array} \tag{14}$$

As utility can take only positive values, it must be $N_E > 0$ at the optimum. Rearranging (13) shows that optimal output $q^o(c)$ has to satisfy

$$q^o(c) = \frac{L}{\gamma}(\alpha - c) - \frac{\eta}{\gamma} N_E^o \int_0^{c_M} q^o(c) dG(c) = \frac{L}{\gamma}(\alpha - c) - \frac{\eta}{\gamma} Q^o \tag{15}$$

where 'o' labels optimum variables, $Q^o \equiv L \int_{i\in\Omega} q_i^c di = N_E^o \int_0^{c_M} q^o(c) dG(c) = N^o \int_0^{c^o} q^o(c) dG^o(c)$ is the total supply of the differentiated good, and $N^o =$

$N_E^o G(c^o)$ is the number of varieties supplied with $G^o(c) = G(c)/G(c^o)$ denoting the conditional distribution of unit input requirements for varieties that the planner produces. Equation (15) and the constraint $q^o(c) \geq 0$ imply that the planner sets

$$q^o(c) = \begin{cases} \frac{L}{\gamma}(c^o - c) & c \leq c^o = \alpha - \frac{\eta}{L}Q^o \\ 0 & c > c^o \end{cases} . \tag{16}$$

Result (16) reveals that as the market also the planner follows a cutoff rule allowing only for the production of varieties with low enough unit labor requirements: $q^o(c) \geq 0$ only for $c \leq c^o$.

Expressions (16) and (2) can be used to show that the optimal output levels would clear the market in a decentralized scenario only if each producer priced at marginal cost. To see this, note that (2) implies $q(c) = [\alpha - p(c)]L/\gamma - \eta Q/\gamma$. Then, imposing $q(c) = q^o(c) = (c^o - c)\,L/\gamma$ and $Q = Q^o = (\alpha - c^o)\,L/\eta$ from (16) respectively on the left and on the right hand sides of $q(c) = [\alpha - p(c)]L/\gamma - \eta Q/\gamma$ gives $p(c) = c$.

Integrating (13) across c gives

$$Q^o = \frac{\gamma N^o}{\gamma + \eta N^o}\frac{L}{\gamma}\left(\alpha - \bar{c}^o\right),$$

with $\bar{c}^o = \int_0^{c^o} c dG^o(c)$. Substituting this result into $c^o = \alpha - \eta Q^o/L$ from (16) and solving for N^o gives

$$N^o = \frac{\gamma}{\eta}\frac{\alpha - c^o}{c^o - \bar{c}^o}, \tag{17}$$

which is the planner's analogue of the market zero cutoff profit condition (9).

Finally, substituting (17) and (16) in (14) yields

$$\frac{1}{2}\int_0^{c^o} (c^o - c)^2\, dG(c) = \frac{\gamma f}{L}, \tag{18}$$

which is the planner's analogue of the market free entry condition (7).

4 Equilibrium versus Unconstrained Optimum

The efficiency of the market outcome with respect to the optimum depends on how the number of varieties actually produced in the market N^m and the (conditional) distribution of their unit labor requirements dictated by the cutoff c^m compare to the optimal ones as implied by N^o and c^o. In particular, as already highlighted by Nocco, Ottaviano and Salto (2014), in the market equilibrium firm selection is too weak due to $c^o < c^m$, average firm size is too small, low cost firms are too small and high cost firms are too large. Moreover, product variety is too rich due to $N^o < N^m$ (too poor due to $N^o > N^m$) when varieties are close (far) substitutes, the sunk entry cost is small (large), market size is large (small) and the difference c_M between the highest and the lowest possible marginal cost realizations is small (large).

While we will reproduce these findings for completeness, our aim here is to understand how the extent of inefficiency is affected by the degree of firm heterogeneity. To this aim, we will first study the extent of inefficiency in the limit case of homogeneous firms and then discuss what firm heterogeneity adds to the picture. In so doing, we will show that our framework encompasses previous results obtained by Ottaviano and Thisse (1999) with homogenous firms and similar preferences.

4.1 Homogeneous Firms

The case of homogeneous firms is represented by a degenerate distribution in which all firms have the same unit labor requirement. In this case there is, therefore, no variance in labor requirements ($\sigma_c^2 = 0$), all entrants produce ($G(c^m) = 1$), and c is exogenous, common to all firms and thus equal to the average unit labor requirement ($c = \bar{c}$).

To help comparison with Ottaviano and Thisse (1999), we express the model in terms of conditional average $\bar{c}(x)$ and conditional variance $\sigma_c^2(x)$ of the unit labor requirement distribution for $x = \{c^m, c^o\}$ where, by definition, we have $\bar{c}(x) = \left[\int_0^x c dG(c)\right] / G(x)$ and $\sigma_c^2(x) = \left[\int_0^x c^2 dG(c)\right] / G(x) - \bar{c}(x)^2$. The free entry condition (7) in the market equilibrium can then be rewritten as

$$\frac{1}{4}G(c^m)\left\{\left[c^m - \bar{c}(c^m)\right]^2 + \sigma_c^2(c^m)\right\} = \frac{\gamma f}{L} \tag{19}$$

while the corresponding condition (18) for the planner becomes

$$\frac{1}{2}G(c^o)\left\{\left[c^o - \bar{c}(c^o)\right]^2 + \sigma_c^2(c^o)\right\} = \frac{\gamma f}{L}. \tag{20}$$

As for the market outcome, solving expression (19) under $G(c^m) = 1$ and $c(c^m) = \bar{c}$ yields the equilibrium cutoff with homogeneous firms ('willingness to pay')

$$c^m = \bar{c} + 2\sqrt{\frac{\gamma f}{L}}. \tag{21}$$

As c^m is larger than the common marginal cost $\bar{c}$, all entrants produce. Given (5), the equilibrium quantity produced of each good is $q^m(\bar{c}) = L\left(c^m - \bar{c}\right) / \left(2\gamma\right)$, which can then be used together with (21) to find the equilibrium output of each firm

$$q^m = \sqrt{\frac{fL}{\gamma}}. \tag{22}$$

Finally, the equilibrium number of varieties supplied can be found by plugging the cutoff value (21) into the zero cutoff profit condition (9) to obtain

$$N^m = \frac{(\alpha - \bar{c})\sqrt{\frac{\gamma L}{f}} - 2\gamma}{\eta}, \tag{23}$$

so that in equilibrium total supply of the differentiated good evaluates to

$$q^m N^m = \frac{(\alpha - \bar{c})}{\eta}L - 2\frac{1}{\eta}\sqrt{\gamma f L}. \tag{24}$$

Turning to the planner outcome, condition (20) can be solved under $G(c^o) = 1$ and $c(c^o) = \bar{c}$ to find the optimal cutoff

$$c^o = \bar{c} + \sqrt{2}\sqrt{\frac{\gamma f}{L}}. \tag{25}$$

As c^o is larger than the common marginal cost $\bar{c}$, also in the optimum all entrants produce. Plugging the cutoff value (25) into the cutoff rule (16) determines the optimal output of each good

$$q^o = \sqrt{2}\sqrt{\frac{fL}{\gamma}}. \tag{26}$$

The corresponding optimal number of varieties supplied can be retrieved by combining (17) and (25) to yield

$$N^o = \frac{(\alpha - \bar{c})\sqrt{\frac{\gamma L}{2f}} - \gamma}{\eta}. \tag{27}$$

Optimal total supply of the differentiated good then evaluates to

$$q^o N^o = \frac{(\alpha - \bar{c})}{\eta} L - \frac{1}{\eta}\sqrt{2\gamma f L}. \tag{28}$$

Comparisons of the equilibrium and the optimal outcomes reveal several ways in which the former is inefficient. Comparing (26) and (24) with (22) and (28) respectively show that in the market equilibrium each firm and the differentiated good sector as a whole are smaller than optimal ($q^m < q^o$ and $q^m N^m < q^o N^o$). Hence, we can write:

Proposition 1 ***(Between-sector misallocation)*** *When firms are homogeneous, in the market equilibrium the total supply of the differentiated good sector as well as the quantity supplied of each variety are smaller than optimal.*

The intuition behind this proposition can be gauged by recalling that, as discussed in Section 3.1, in the market equilibrium the markup of a firm with marginal cost c equals $\mu^m(c) = p^m(c) - c = (c^m - c)/2$. With homogeneous firms we have $c = \bar{c}$ and thus $\mu^m(\bar{c}) = (c^m - \bar{c})/2$ so that all firms supply the same quantity and quote the same positive markup above marginal cost. Accordingly, consumption is inefficiently biased against the differentiated varieties and in favor of the numeraire good as the prices of the former are inefficiently above marginal cost.

Turning to product variety, (27) and (23) imply that the number of varieties supplied by the market is inefficiently large ($N^m > N^o$) when varieties are close substitutes (γ is small), the entry labor requirement f is low relative to market size as measured by α and L, and when the (average) unit labor requirement $\bar{c}$ is low, that is when

$$\alpha > \bar{c} + \frac{\sqrt{2}}{\sqrt{2} - 1}\sqrt{\frac{\gamma f}{L}}.$$

Vice versa, the market tends to underprovide variety ($N^m < N^o$) when the opposite inequality holds. These results concur with those in Ottaviano and Thisse (1999) once we account for their parametrization $L = 1$ and $\bar{c} = 0$. We can summarize the foregoing as:

Proposition 2 ***(Inefficient product variety)*** *When firms are homogeneous, in the market equilibrium product variety is richer (poorer) than in the optimum when varieties are close (far) substitutes, the entry cost is small (large), market size is large (small) and average unit input requirement is low (high).*

From the demand functions it is apparent that these propositions are driven by the fact that (21) and (25) imply $c^m > c^o$, leading to:

Proposition 3 ***(Inefficient willingness to pay)*** *When firms are homogeneous, in the market equilibrium the cost cutoff is inefficiently large.*

Clearly, with homogeneous firm selection is immaterial. This is the main difference with respect to a situation where firms are heterogeneous, to which we turn next.

4.2 Heterogeneous Firms

To study the role of firm heterogeneity we follow Melitz and Ottaviano (2008) and Nocco, Ottaviano and Salto (2014) in assuming that firms draw their marginal productivity $1/c$ from a Pareto distribution with shape parameter $k \geq 1$ over the support $[1/c_M, \infty)$ and thus their unit input requirement c from a distribution with cumulative density function

$$G(c) = \left(\frac{c}{c_M}\right)^k, \ c \in [0, c_M]. \tag{29}$$

For $k = 1$ the distribution is uniform. As k rises, density is skewed towards the upper bound of the support.

As for the market outcome, under this distributional assumption the 'free entry condition' (7) becomes

$$\left(\frac{c^m}{c_M}\right)^k \frac{L\left(c^m\right)^2}{2\gamma(k+1)(k+2)} = f, \tag{30}$$

where, due to the law of large numbers, the two terms of (30) have a double interpretation: $G(c^m) = (c^m/c_M)^k$ is the *ex ante* probability that an entrant will produce as well as the *ex post* share of entrants that eventually produce; and $L\left(c^m\right)^2/[2\gamma(k+1)(k+2)]$ is the *ex ante* expected profit conditional on producing as well as the *ex post* average profit of producers.

Condition (30) can be solved for the unique equilibrium cutoff marginal cost

$$c^m = \left[\frac{2\gamma(k+1)(k+2)\left(c_M\right)^k f}{L}\right]^{\frac{1}{k+2}}. \tag{31}$$

This implies that the average unit labor requirent of producers is $\bar{c}^m = c^m k/(k+1)$, which can be substituted into the 'zero cutoff profit condition' (9) to obtain the equilibrium number of producers (and varieties produced)

$$N^m = \frac{2\gamma(k+1)}{\eta}\frac{\alpha - c^m}{c^m}, \tag{32}$$

with the corresponding equilibrium number of entrants given by $N_E^m = N^m/G(c^m) = N^m\left(c_M/c^m\right)^k$ given that the conditional distribution satisfies $G^m(c) = G(c)/G(c^m) = \left(c/c^m\right)^k$.

Turning to the optimum, using the distributional assumption (29) in condition (18) gives

$$\left(\frac{c^o}{c_M}\right)^k \frac{L\,(c^o)^2}{\gamma(k+1)(k+2)} = f \tag{33}$$

so that the optimal cutoff evaluates to

$$c^o = \left[\frac{\gamma(k+1)(k+2)\,(c_M)^k\, f}{L}\right]^{\frac{1}{k+2}}. \tag{34}$$

Assumption (29) also implies that the planner's cutoff condition (17) becomes

$$N^o = N_E^o G(c^o) = \frac{\gamma(k+1)}{\eta}\frac{\alpha - c^o}{c^o} \tag{35}$$

with the corresponding $N_E^o = N^o/G(c^o) = N^o\,(c_M/c^o)^k$ given $G^o(c) = G(c)/G(c^o) = (c/c^o)^k$.

Comparing the equilibrium cutoff, number of varieties and quantities supplied with the optimal ones finds results with firm heterogeneity (already highlighted by Nocco, Ottaviano and Salto, 2014) broadly in line with those reported in Propositions 1, 2 and 3 with homogeneous firms. Yet, there are some differences.

Proposition 2 applies to the case of heterogeneous firms qualitatively unchanged. As in the homogeneous firms case, with firm heterogeneity we have $N^m > N^o$ if and only if varieties are close substitutes (γ small), the entry requirement f is low and the mean of the (unconditional) distribution of unit labor requirements (dictated by c_M) are small compared with market size as measured by α and L. Indeed, from direct comparison between (32) and (35) given (31) and (34), $c^m = 2^{1/(k+2)}c^o$ implies $N^m > N^o$ as long as

$$\alpha > \alpha_1 \equiv \frac{c^o}{2^{\frac{k+1}{k+2}} - 1} = \frac{1}{2^{\frac{k+1}{k+2}} - 1}\left[\frac{\gamma(k+1)(k+2)\,(c_M)^k\, f}{L}\right]^{\frac{1}{k+2}}, \tag{36}$$

which is the case when α as well as L are large and when not only γ and f but also c_M are small. However, when firms are heterogeneous there is a difference between firms that enter and firms that produce as some firms that enter the market may decide not to produce after observing their high unit labor requirement. If we use N_E^j for $j \in \{m, o\}$ to denote the equilibrium and optimal number of 'entrants' (i.e. varieties that are designed but not necessarily produced) in the market and the optimal outcomes, then we have

$$N_E^j = \frac{N^j}{G(c^j)} = N^j\left(\frac{c_M}{c^j}\right)^k. \tag{37}$$

This can be used together with (32) and (35) as well as (31) and (34), to show that $c^m = 2^{1/(k+2)}c^o$ imply $N_E^m > N_E^o$ as long as

$$\alpha > \alpha_2 \equiv \frac{2^{2/(k+2)} - 1}{2^{1/(k+2)} - 1}c^o = \frac{2^{2/(k+2)} - 1}{2^{1/(k+2)} - 1}\left[\frac{\gamma(k+1)(k+2)\,(c_M)^k\, f}{L}\right]^{\frac{1}{k+2}} \tag{38}$$

with $\alpha_2 > \alpha_1$. Condition (38) holds in the same (qualitative) situations as condition (36), that is, when α as well as L are large and γ, f as well as c_M are small.

Also Proposition 3 still holds qualitatively unchanged as (31) and (34) imply that in the market equilibrium the cutoff is inefficiently large ($c^o < c^m$). However, with firm heterogeneity $c^o < c^m$ entails inefficient selection as some varieties with high unit input requirement $c \in [c^o, c^m]$ are supplied by the market but not by the planner ('extensive margin misallocation'). Moreover, while with firm heterogeneity in the market equilibrium the total supply of the differentiated good sector is still smaller than optimal, among the varieties supplied both by the market and the planner, the cutoff rules (5) and (16) imply that the output of the varieties with high unit labor requirement is inefficiently large while the output of those with low unit labor requirement is inefficiently small ('intensive margin misallocation'). In detail, given $c^m = 2^{1/(k+2)}c^o$, it is readily verified that $q^m(c) > q^o(c)$ if and only if $c > \left(2 - 2^{1/(k+2)}\right) c^o$, which falls in the relevant interval $[0, c^o]$ as we have $0 < \left(2 - \sqrt[3]{2}\right) < \left(2 - 2^{1/(k+2)}\right) < 1$. Hence, with respect to the optimum, the market equilibrium undersupplies varieties with marginal cost $c \in [0, \left(2 - 2^{1/(k+2)}\right) c^o)$ and oversupplies varieties with marginal cost $c \in (\left(2 - 2^{1/(k+2)}\right) c^o, c^m]$. The intuition behind this result is that in equilibrium firms with lower unit labor requirement do not pass on their entire cost advantage to consumers as they absorb part of it in the markup and this decreases with the unit labor requirement. Hence, the price ratio of higher cost firms to lower cost firms is smaller than the ratio of their marginal costs and thus the quantities sold by the former are too large from an efficiency point of view relative to those sold by the latter.

To summarize what is different from the homogeneous firm case, we can state:

Proposition 4 ***(Between- and within-sector misallocation)*** *When firms are heterogeneous, in the market equilibrium the total supply of the differentiated good sector is smaller than optimal. However: (i) some varieties with high unit input requirement are supplied while they should not from an efficiency point of view; (ii) as for varieties that should be supplied also from an efficiency viewpoint, the output of those with lower unit input requirement is smaller than optimal while the output of those with higher unit input requirement is larger than optimal.*

5 Firm Heterogeneity and Market Inefficiency

The previous section has shown that firm heterogeneity is an important source of misallocation at the extensive and intensive margins across firms within the differentiated good sector. We now discuss how the extent of inefficiency is affected by the degree of firm heterogeneity. The key question here is whether or not the inefficiency of the market equilibrium is largest when selection is needed most, that is, when there are a lot of firms with high unit labor requirement and only few with low unit labor requirement.

As discussed by Ottaviano (2012), the scale and shape parameters of the Pareto distribution (29) regulate the 'heterogeneity' of cost draws along two dimensions: 'richness' and 'evenness' (Maignan, Ottaviano, Pinelli and Rullani, 2003). First, the scale parameter c_M quantifies 'richness', defined as the measure ('number') of different unit labor requirements that can be drawn. Larger c_M leads to a rise in heterogeneity along the richness dimension, and this is achieved by making it possible to draw also larger unit labor requirements. Second, the shape parameter k is an inverse measure of 'evenness', defined as the similarity between the probabilities of those different draws to happen. When $k = 1$, the unit labor requirement distribution is uniform on $[0, c_M]$ with maximum evenness. As k increases, the distribution becomes more concentrated at higher unit labor requirements close to c_M: evenness falls. As k goes to infinity, the distribution becomes degenerate at c_M: all draws deliver a unit labor requirement c_M with probability one. Hence, smaller k leads to a rise in heterogeneity along the evenness dimension, and this is achieved by making low unit labor requirements more likely without changing the unit labor requirements that are possible. Accordingly, more richness (larger c_M) comes with higher average unit labor requirement ('cost-increasing richness'), more evenness (smaller k) comes with lower average unit labor requirement ('cost-decreasing evenness').

Given the cutoff expressions (31) and (34), more heterogeneity has different impacts on selection depending on whether it comes through more richness or evenness. To see this, rewrite (31) and (34) as

$$\left(\frac{c^m}{c_M}\right)^k \left[\frac{L}{4\gamma} \frac{2\left(c^m\right)^2}{(k+2)(k+1)}\right] = f$$

and

$$\left(\frac{c^o}{c_M}\right)^k \left[\frac{L}{2\gamma} \frac{2\left(c^m\right)^2}{(k+2)(k+1)}\right] = f,$$

where $\left(c^m/c_M\right)^k$ and $\left(c^o/c_M\right)^k$ are the shares of viable varieties and the bracketed terms are average firm profit for the market equilibrium and average surplus per variety for the optimum respectively. For any given cutoffs, more cost-increasing richness (larger c_M) decreases the left hand sides of both expressions through its depressing effect on the share of viable varieties. As the right hand sides are constant, (31) and (34) can keep on holding only if the cutoffs rise. Differently, for any given cutoffs (smaller than c_M), more cost-decreasing evenness (smaller k) increases the left hand sides of both expressions through its enhancing effect on both the share of viable varieties and average profit or surplus. Again, as the right hand sides are constant, (31) and (34) can keep on holding only if the cutoffs fall. Hence, while more cost-increasing richness makes selection softer, more cost-decreasing evenness makes it tougher.

When we focus on the percentage deviation of the market equilibrium from the optimum, only the change in evenness matters for several outcomes. Specifically, given $c^m = 2^{1/(k+2)} c^o$, more evenness (smaller k) leads to a larger percentage gap in the cutoffs between the market equilibrium and the optimum ($\left(c^m - c^o\right)/c^o$ rises) whereas more richness is immaterial. Hence, we have:

Proposition 5 ***(Heterogeneity and extensive margin misallocation)*** *More cost-decreasing evenness increases the percentage gap in the cutoffs between the market equilibrium and the optimum. Cost-increasing richness has no impact on this gap.*

Several implications can be derived from Proposition 5. First, recall that, with respect to the optimum, the market equilibrium undersupplies varieties with marginal cost $c \in [0, \left(2 - 2^{1/(k+2)}\right) c^o)$ and oversupplies varieties with marginal cost $c \in (\left(2 - 2^{1/(k+2)}\right) c^o, c^m]$. When k falls $\left(2 - 2^{1/(k+2)}\right) c^o / c^m$ also falls whereas it does not change when c_M changes. This leads to:

Corollary 6 ***(Heterogeneity and intensive margin misallocation)*** *More cost-decreasing evenness makes the overprovision of varieties relatively more likely than its underprovision in the market equilibrium. Cost-increasing richness has no impact on this.*

Second, we know that the overall size of the differentiated good sector in the market equilibrium is smaller than optimal. However, expressions $N^m \overline{q}^m = (L/\eta)(\alpha - c^m)$ and $N^o \overline{q}^o = (L/\eta)(\alpha - c^o)$ imply that there can exist a threshold value k^* of k such that we have:

Corollary 7 ***(Heterogeneity and between-sector misallocation)*** *More cost-decreasing evenness decreases the gap in the total output of the differentiated varieties between the market equilibrium and the optimum for $k > k^*$ and, viceversa, it increases the gap for $k < k^*$. The threshold value k^* increases with L and c_M, and decreases with γ and f. However, if L and c_M are sufficiently small and/or γ and f are sufficiently large, more cost-decreasing evenness does always decrease the gap. More cost-increasing richness always increases the gap.*[4]

Third, expressions (32) and (35) with the associated condition (36) lead to:

Corollary 8 ***(Heterogeneity and product variety)*** *Less cost-decreasing evenness and more cost-increasing richness makes the underprovision of variety relatively more likely than its overprovision in the market equilibrium.*

Analogously, given (37) and the associated condition (38), we can write:

Corollary 9 ***(Heterogeneity and entry)*** *Less cost-decreasing evenness and more cost-increasing richness makes insufficient entry relatively more likely than excess entry in the market equilibrium.*

Finally, we can look at the relation between heterogeneity and welfare. We prove in the Appendix that we have:

Corollary 10 ***(Heterogeneity and welfare)*** *More cost increasing richness increases the welfare gap between the optimum and the market equilibrium if the preference for the differentiated good (α) is sufficiently large; otherwise it decreases that gap.*

[4] More precisely, more cost-decreasing evenness increases the gap if $\left[2^{\frac{1}{k+2}} \ln 2 - (2k+3)\left(2^{\frac{1}{k+2}} - 1\right)/(k+1)\right] / \left(2^{\frac{1}{k+2}} - 1\right) + \ln((k+1)(k+2)) > \ln\left[Lc_M^2/(\gamma f)\right]$; viceversa, it decreases the gap if the opposite inequality sign holds. The threshold k^* corresponds to the value at which the function $\left[2^{\frac{1}{k+2}} \ln 2 - (2k+3)\left(2^{\frac{1}{k+2}} - 1\right)/(k+1)\right] / \left(2^{\frac{1}{k+2}} - 1\right) + \ln((k+1)(k+2))$, (which is increasing in k) crosses $\ln\left[Lc_M^2/(\gamma f)\right]$, and it exists only if L and c_M are sufficiently large and/or γ and f are sufficiently small.

Moreover, numerical analysis shows that when α is sufficiently large, more cost-decreasing evenness decreases (increases) the welfare gap if k is initially low (high). In other words, from a welfare point of view, when the preference for the differentiated good is sufficiently large, the inefficiency of the market equilibrium is largest when selection is needed most: many firms with high unit labor requirement coexist with few firms with high labor requirement.

6 Constrained Optimum

The decentralization of the optimal outcome with firm heterogeneity we have discussed so far requires a rich set of policy tools to correct for both between- and within-sector misallocation. In particular, firm-specific per unit transfers are needed to deal with within-sector misallocation at the intensive margin and lump sum trasfers for firms are needed to deal with within-sector misallocation at the extensive margin.

When such tools are not available, policy makers cannot affect firms' pricing and selection but can only target product variety. In this case the most they can do from a welfare viewpoint is to implement the 'second best' allocation of a constrained planner who maximizes (12) with respect to N_E subject to two constraints: profit maximizing output (5) and the free entry condition (30). These impose the planner the market cutoff (31). Therefore, substituting (5) and (30) in (12) allows us to rewrite the constrained problem as the maximization of

$$W = L + \overline{q}_0 L + \frac{2\alpha\left(k+2\right) - \left(2k+3\right)c^m}{2c^m} f N_E - \frac{\eta\left(k+2\right)\left(c^m\right)^k}{4\gamma\left(k+1\right)\left(c_M\right)^k} f\left(N_E\right)^2 \quad (39)$$

with respect to N_E. Using $N_E = N\left(c_M/c^m\right)^k$ to substitute for N_E in the first order condition of the planner's problem yields

$$N^s = \frac{2\gamma\left(k+1\right)}{\eta}\frac{\alpha - \frac{2k+3}{2(k+2)}c^m}{c^m}. \quad (40)$$

Comparing this expression with (32) reveals that product variety is richer in the constrained optimum than in the market equilibrium, while each variety's output is the same as implied under (5) by the fact that the cutoff is identical in both outcomes. Specifically, one can compute the difference between the numbers of varieties in the constrained optimum and in the market equilibrium as

$$N^s - N^m = \frac{\gamma\left(k+1\right)}{\eta\left(k+2\right)},$$

which is always positive. The impact of heterogeneity on the inefficiency of the market equilibrium with respect to the constrained optimum can be assessed by noticing that richness c_M does not affect $N^s - N^m$ whereas $\partial\left(N^s - N^m\right)/\partial k$ is also always positive. This leads to:

Proposition 11 ***(Heterogeneity, product variety and constrained efficiency)*** *While by definition the output of each variety and the average output per variety in the market equilibrium and in the constrained optimum coincide,* *product variety is poorer in the former than in the latter outcome. Less cost-decreasing evenness fosters the underprovision of variety in the market equilibrium with respect to the constrained optimum. Cost-increasing richness does not impact the extent of such underprovision.*

In order to assess the impact of heterogeneity on welfare, expression (40) can be used together with $N_E = N\left(c_M/c^m\right)^k$ to substitute for N_E while (30) can be used to substitute for f in the planner's objective. These substitutions allow us to express welfare in the constrained optimum as a function of the market cutoff

$$W^s = L + \overline{q}_0 L + \frac{L}{2\eta}\left[\alpha - \frac{2k+3}{2(k+2)}c^m\right]^2,$$

which is smaller than W^o but larger than W^m. The welfare gap between the constrained optimum and the market equilibrium evaluates to

$$W^s - W^m = \frac{L}{8\eta}\left(\frac{c^m}{k+2}\right)^2$$

and, given (31), it is readily verified that we have $\partial\left(W^s - W^m\right)/\partial c_M > 0$. Hence, less cost-increasing richness (smaller c_M) reduces the welfare gap between the market outcome and the constrained optimum. Analogously, given (31), derivation of $W^s - W^m$ with respect to k yields

$$\frac{\partial\left(W^s - W^m\right)}{\partial k} = \frac{L}{4\eta}\frac{(c^m)^2}{(k+2)^4(k+1)}\left[(k+1)\ln\frac{(c_M)^2 L}{2\gamma(k+1)(k+2)f} - \left(k + k^2 - 1\right)\right],$$

which is positive (negative) for $k < k^{**}$ $(k > k^{**})$ when $\ln\left[(c_M)^2 L/\left(12\gamma f\right)\right] > 1/2$, where $k^{**} > 1$ is the value of k that solves

$$\ln\left\{(c_M)^2 L/\left[2\gamma(k+1)(k+2)f\right]\right\} = \left(k - 1 + k^2\right)/\left(k+1\right).$$

Otherwise, when $\ln\left[(c_M)^2 L/\left(12\gamma f\right)\right] < 1/2$, $\partial\left(W^s - W^m\right)/\partial k < 0$ holds.[5] Hence, we can conclude that:

Corollary 12 ***(Heterogeneity, welfare and constrained efficiency)*** *Less cost-increasing richness (smaller c_M) reduces the welfare gap between the constrained optimum and the market equilibrium. The same happens in the case of more cost-decreasing evenness (smaller k) when initial evenness is high. Differently, when initial evenness is low, more cost-decreasing evenness raises the gap.*

As in the case of the 'unconstrained' optimum, the inefficiency of the market equilibrium is largest when selection is needed most.

[5]The term $(c_M)^2 L/\left[2\gamma(k+1)(k+2)f\right]$ is larger than 1 to ensure $c^m < c_M$. Then the left hand side of the equation, $\ln\left\{(c_M)^2 L/\left[2\gamma(k+1)(k+2)f\right]\right\}$, is a positive and decreasing function of k. The right hand side, $\left(k - 1 + k^2\right)/\left(k+1\right)$, is instead a positive and increasing function of k, attaining value $1/2$ at $k = 1$. Hence, the two functions cross only once at $k = k^{**} > 1$ if $\ln\left[(c_M)^2 L/\left(12\gamma f\right)\right] > 1/2$.

7 Conclusion

After some decades of relative oblivion, the interest in the optimality properties of monopolistic competition has recently re-emerged due to the 'heterogeneous firms revolution' in international trade theory initiated by Melitz (2003). The availability of an appropriate and parsimonious framework to deal with firm heterogeneity allows one to bring back into the normative debate the full set of questions the canonical formalization of the Chamberlinian model by Spence (1976) and Dixit and Stiglitz (1977) was designed to answer. In particular, it provides a useful analytical tool to address the question whether in the market equilibrium the products are supplied by the right set of firms, or there are rather 'errors' in the choice of technique.

We have contributed to this debate by showing that the market outcome errs in many ways. In particular we proved that heterogeneity in firms's production costs add specific distortions to the models with homogeneous firms on the one hand and with CES utility function on the other, with respect to the selection of firms, number of products, the size and the selection of producers. More crucially with respect to the existing literature, we have also shown that the extent of the errors depends on the degree of firm heterogeneity. In particular, we have found that the inefficiency of the market equilibrium tends to be largest when selection is needed most, that is, when there are relatively many firms with low productivity and relatively few firms with high productivity. This holds from the viewpoints of both unconstrained and constrained efficiency.

These insights have been obtained for a parametrization of demand that is admittedly specific but still non-separable and more flexible than the CES. Also the adopted parametrization of the distribution of firm productivity is quite specific. While it would be important to understand whether our results would apply to a more general setup, this is left to future research.

References

[1] Anderson S., A. de Palma and J.-F. Thisse (1992) *Discrete Choice Theory and Product Differentiation* (Cambride MA: MIT Press).

[2] Arkolakis, C., A. Costinot, and A. Rodriguez-Clare (2012) New trade models, same old gains?, *American Economic Review* 102, 94-130.

[3] Behrens K. and Y. Murata (2012) Trade, competition, and efficiency, *Journal of International Economics* 87, 1-17.

[4] Bertoletti P. and F. Etro (2016) Preferences, Entry and Market Structure, RAND Journal of Economics 47(4), 792-821.

[5] Bertoletti P., F. Etro and I. Simonovska (2017) International Trade with Indirect Additivity, forthcoming in American Economic Journal: Microeconomics.

[6] Bishop R. (1967) Monopolistic competition and welfare economics, in Kuenne R. (ed.), *Monopolistic Competition Theory: Studies in Impact: Essays in Honor of Edward H. Chamberlin* (New York: John Wiley).

[7] Brakman and Heijdra (2004) Introduction, in Brakman R. and B. Heijdra (eds.), *The Monopolistic Competition Revolution in Retrospect* (Cambridge: Cambridge University Press).

[8] Chamberlin E. (1933) *The Theory of Monopolistic Competition* (Cambride MA: Harvard University Press).

[9] Dhingra, S., and J. Morrow (2016). Monopolistic competition and optimum product diversity under firm heterogeneity, LSE Department of Economics, mimeo.

[10] Dixit A. (2004) Some reflections on theories and applications of monopolistic competition, in Brakman R. and B. Heijdra (eds.), *The Monopolistic Competition Revolution in Retrospect* (Cambridge: Cambridge University Press).

[11] Dixit A. and J. Stiglitz (1975) Monopolistic competition and optimum product diversity, The Warwick Economics Research Series 64, University of Warwick, Department of Economics.

[12] Dixit A. and J. Stiglitz (1977) Monopolistic competition and optimum product diversity, *American Economic Review* 67, 297-308.

[13] Kaldor N. (1935). Market imperfection and excess capacity, *Economica* 2, 33-50.

[14] Maignan C., G. Ottaviano, D. Pinelli and F. Rullani (2003) Bio-ecological diversity vs. socio-economic diversity: A comparison of existing measures, FEEM Working Paper 58.2003.

[15] Melitz M. and G. Ottaviano (2008) Market size, trade, and productivity, *Review of Economic Studies* 75, 295-316.

[16] Neary P. (2004) Monopolistic competition and international trade theory, in Brakman R. and B. Heijdra (eds.), *The Monopolistic Competition Revolution in Retrospect* (Cambridge: Cambridge University Press).

[17] Nocco, A., G.I.P. Ottaviano and M. Salto (2014). Monopolistic competition and optimum product selection, *American Economic Review* 104, 304-309.

[18] Ottaviano G. (2012) Agglomeration, trade and selection, *Regional Science and Urban Economics*, vol. 42, issue 6, 987-997.

[19] Ottaviano G. and J.-F. Thisse (1999) Monopolistic competition, multiproduct firms and optimum product diversity, CEPR Discussion Paper 2151.

[20] Ottaviano G., T. Tabuchi and J.-F. Thisse (2002) Agglomeration and trade revisited, *International Economic Review* 43, 409-436.

[21] Pascoa M. R. (1997) Monopolistic Competition and Non-Neighboring-Goods, *Economic Theory* 9, 129-142.

[22] Spence M. (1976) Product selection, fixed costs and monopolistic competition, *Review of Economic Studies* 43, 217-235.

[23] Stiglitz J. (1975) Monopolistic competition and the capital market, in Brakman R. and B. Heijdra (eds.), *The Monopolistic Competition Revolution in Retrospect* (Cambridge: Cambridge University Press).

[24] Melitz M. (2003) The impact of trade on intra-industry reallocations and aggregate industry productivity, *Econometrica* 71, 1695-1725.

[25] Melitz M. and S. Redding (2012) Heterogeneous firms and trade, in Gopinath G., G. Grossman and K. Rogoff (eds.) *Handbook of International Economics* (Amsterdam: North Holland), forthcoming.

[26] Zhelobodko, E., S. Kokovin, M. Parenti and J.-F. Thisse (2012). Monopolistic competition: beyond the constant elasticity of substitution, *Econometrica* 80, 2765-2784.

8 Appendix: Proof of Corollary 10

The welfare attained in the market equilibrium can be expressed as a function of the corresponding cutoff through the following substitutions in the planner's objective (12). Expression (32) can be used together with $N_E = N\left(c_M/c^m\right)^k$ to substitute for N_E. Expression (30) can be used to substitute for f. Expression (5) can be used to substitute for $q(c)$. These substitutions give

$$W^m = L + \overline{q}_0 L + \frac{L}{2\eta}\left(\alpha - c^m\right)\left(\alpha - \frac{k+1}{k+2}c^m\right). \tag{41}$$

Analogously, the welfare level attained in the optimum can be expressed as a function of the corresponding cutoff through the following substitutions in the planner's objective (12). Expression (35) can be used together with $N_E = N\left(c_M/c^m\right)^k$ to substitute for N_E. Expression (33) can be used to substitute for f. Expression (16) can be used to substitute for $q(c)$. These substitutions give:

$$W^o = L + \overline{q}_0 L + \frac{L}{2\eta}\left(\alpha - c^o\right)^2. \tag{42}$$

Given $c^m = 2^{1/(k+2)}c^o$, the gap in the welfare level between the optimum and the market equilibrium in (42) and (41) can be written as

$$W^o - W^m = a c^m \frac{L}{2\eta}, \tag{43}$$

where $a \equiv b\alpha - dc^m$, $b \equiv (2k+3)/\left(k+2\right) - 2^{(k+1)/(k+2)}$ and $d \equiv \left(k+1\right)/\left(k+2\right) - 2^{-2/(k+2)}$. Given that $\alpha/c^m > 1 > d/b$, it is readily verified that $a > 0$ and, consequently, $W^o > W^m$.

Derivation of (43) with respect to c_M then gives

$$\frac{\partial\left(W^o - W^m\right)}{\partial c_M} = \frac{L}{2\eta}\left[\left(\frac{2k+3}{k+2} - \left(\frac{1}{2}\right)^{-\frac{k+1}{k+2}}\right)\alpha - 2\left(\frac{k+1}{k+2} - \left(\frac{1}{2}\right)^{\frac{2}{k+2}}\right)c^m\right]\frac{\partial c^m}{\partial c_M},$$

where $\partial c^m/\partial c_M > 0$ and the sign of the term in the square brackets depends on the value of α: it is positive (negative) for $\alpha > (<)\alpha_3$ with

$$\alpha_3 \equiv 2\frac{\frac{k+1}{k+2} - \left(\frac{1}{2}\right)^{\frac{2}{k+2}}}{\frac{2k+3}{k+2} - \left(\frac{1}{2}\right)^{-\frac{k+1}{k+2}}}c^m.$$

Hence, we have that, as stated in Corollary 10, more cost increasing richness increases the gap $W^o - W^m$ if the preference for the differentiated good is sufficiently large (small), i.e. $\alpha > \alpha_3$ $(\alpha < \alpha_3)$.

Moreover, derivation with respect to k gives

$$\frac{\partial\left(W^o - W^m\right)}{\partial k} =$$

$$= \quad \{[\left(\tfrac{2k+3}{k+2} - \left(\tfrac{1}{2}\right)^{-\frac{k+1}{k+2}}\right)\tfrac{(k+1)\ln\left(\frac{c_M}{c^m}\right)^{(k+2)}+2k+3}{(k+1)} - \left(2^{\frac{k+1}{k+2}}\ln 2 - 1\right)]\alpha +$$
$$+[2^{\frac{k}{k+2}}\ln 2 - 1 - 2\left(\tfrac{k+1}{k+2} - \left(\tfrac{1}{2}\right)^{\frac{2}{k+2}}\right)\tfrac{(k+1)\ln\left(\frac{c_M}{c^m}\right)^{(k+2)}+2k+3}{(k+1)}]c^m\}\tfrac{L}{2\eta(k+2)^2}c^m$$

and numerical analysis shows that, when α is sufficiently large, more cost-decreasing evenness decreases (increases) the gap $W^o - W^m$ if k is initially low (high). Indeed, the coefficient of α is positive for k initially low and thus the term in the curly brackets is positive if

$$\alpha > \frac{2\left(\frac{k+1}{k+2}-\left(\frac{1}{2}\right)^{\frac{2}{k+2}}\right)\frac{(k+1)\ln\left(\frac{c_M}{c^m}\right)^{(k+2)}+2k+3}{(k+1)}-\left(2^{\frac{k}{k+2}}\ln 2-1\right)}{\left[\frac{2k+3}{k+2}-\left(\frac{1}{2}\right)^{-\frac{k+1}{k+2}}\right]\frac{(k+1)\ln\left(\frac{c_M}{c^m}\right)^{(k+2)}+2k+3}{(k+1)}-\left(2^{\frac{k+1}{k+2}}\ln 2-1\right)}c^m.$$