Journal of Regulatory Economics (2024) 66:52–61
https://doi.org/10.1007/s11149-024-09479-0

ORIGINAL ARTICLE

Natural monopoly revisited

Oriol Carbonell-Nicolau[1]

Accepted: 7 May 2024 / Published online: 18 May 2024

Abstract
We study the conditions under which production processes exhibit a decreasing average cost function in the absence of perfectly competitive input markets and discuss some implications for regulatory policy.

Keywords Natural monopoly · Monopsony · Returns to scale · Average cost function

JEL Classification L12 · L13

Natural monopolies are typically defined as industries in which the average cost of production decreases with output. The property of increasing returns to scale fully characterizes natural monopolies in the sense that it is necessary and sufficient for the existence of decreasing average costs.[1] This result rests on the important assumption, often omitted, that input markets are perfectly competitive.

In the simplest case when there is only one input, say labor, denoted by l, the production function $f(l)$ exhibits increasing returns to scale if and only if

$$f(\lambda l) > \lambda f(l), \quad \text{for all } \lambda > 1 \text{ and all } l > 0,$$

which can be equivalently stated as

$$c(f(\lambda l)) > c(\lambda f(l)), \quad \text{for all } \lambda > 1 \text{ and all } l > 0, \tag{1}$$

where $c(\cdot)$ denotes the cost function. Letting $w > 0$ be the competitive wage rate, we have

$$c(f(\lambda l)) = w\lambda l = \lambda c(f(l)),$$

[1] See, e.g., Kreps (2004, pp. 219–220) and Serrano and Feldman (2013, pp. 152–155).

Valuable comments from an anonymous referee and the editor are gratefully acknowledged.

✉ Oriol Carbonell-Nicolau
carbonell-nicolau@rutgers.edu

[1] Department of Economics, Rutgers University, 75 Hamilton St., New Brunswick, NJ 08901, USA

and so (1) can be restated as

$$\lambda c(f(l)) > c(\lambda f(l)), \quad \text{for all } \lambda > 1 \text{ and all } l > 0,$$

or, equivalently,

$$\lambda c(x) > c(\lambda x), \quad \text{for all } \lambda > 1 \text{ and all } x > 0,$$

which holds if, and only if,

$$\frac{c(x)}{x} > \frac{c(\lambda x)}{\lambda x}, \quad \text{for all } \lambda > 1 \text{ and all } x > 0,$$

i.e., if, and only if, the average cost function is decreasing.

The purpose of this note is to understand the conditions under which production processes exhibit a decreasing average cost function in the absence of perfectly competitive input markets. The relevance of this line of inquiry rests on the observation that monopoly power tends to go hand-in-hand with monopsony power.[2]

It will be shown that increasing returns to scale are necessary but not sufficient for a decreasing average cost function, and that measures of the degree of oligopsony and market power in input markets are important additional factors determining the shape of average costs. The implications of these results for regulatory policy will be discussed at the end.

The analysis is framed in terms of market power in upstream labor and capital markets. To keep the analysis as simple as possible, we first confine attention to the case of a single-input production function, treating labor and capital separately, and then show that the analysis extends to the standard two-input case.

To begin, consider a monopolist that produces a private good using labor, l, as its only input. The production function is denoted by $f(l)$. Suppose that $w(l)$ is an increasing function representing the *firm-level* inverse labor supply. This function describes how workers employed by the monopolist react to changes in wages. When $w(\cdot)$ is flat (i.e., completely elastic), workers quit (moving to other firms, for example) when the monopolist lowers their wages. This extreme case represents the textbook case of competitive input markets, i.e., a complete absence of the monopolist's labor market power. When $w(\cdot)$ is very steep, the monopolist's labor force tends to be unresponsive to even large wage declines. In this case, the monopolist holds significant labor market power.

The monopolist's cost function, expressed in terms of hours of hired labor, l, is given by $w(l)l$. The *elasticity of the cost function with respect to* l is expressible as

$$\frac{d(w(l)l)}{dl} \cdot \frac{l}{w(l)l} = \frac{w'(l)l + w(l)}{w(l)} = 1 + \frac{1}{\xi_l}, \tag{2}$$

[2] The case of Amazon is detailed in Khan (2016). See also the empirical analysis in Azar et al. (2022), which shows that "[g]iven high concentration, mergers of employers have the potential to significantly increase *labor* market power." Naidu et al. (2018 pp. 546–547) survey empirical literature suggesting that "industry consolidation has given employers greater bargaining power in labor markets."

where ξ_l denotes the wage elasticity of the labor supply facing the monopolist, i.e.,

$$\xi_l = \frac{1}{w'(l)} \cdot \frac{w(l)}{l}.$$

This elasticity is called "*residual labor supply elasticity*" in Naidu et al. (2018). It takes values in the range $[0, \infty)$. The interval's lower bound (resp., upper bound) represents the case of absolute labor market power (resp., the absence of labor market power).

The elasticity of the cost function with respect to l given in (2) measures the relative responsiveness of the monopolist's cost to a one-percent increase in the quantity of hired labor.

The *elasticity of scale*,

$$\xi_f(l) = f'(l) \cdot \frac{l}{f(l)}$$

takes values in the interval $(0, \infty)$ and measures the relative responsiveness of output to a one-percent increase in the quantity of hired labor. Note that

$\xi_f(l) > 1$ for all l if and only if the production function $f(\cdot)$ exhibits increasing returns to scale;

$\xi_f(l) = 1$ for all l if and only if the production function $f(\cdot)$ exhibits constant returns to scale;

$\xi_f(l) < 1$ for all l if and only if the production function $f(\cdot)$ exhibits decreasing returns to scale.

The ratio of the elasticity of the cost function with respect to l to the elasticity of scale,

$$\theta(l) = \frac{1 + (1/\xi_l)}{\xi_f(l)},$$

measures the relative responsiveness of the total cost to the relative responsiveness of output (with respect to a one-percent increase in the quantity of hired labor). Note that

$\theta(l) < 1$ for all l if and only if the monopolist's average cost function is decreasing;

$\theta(l) = 1$ for all l if and only if the monopolist's average cost function is constant;

$\theta(l) > 1$ for all l if and only if the monopolist's average cost function is increasing.

Thus, when the labor market is not competitive, the shape of the average cost function is determined by the elasticity of scale *and* the residual labor supply elasticity. Standard textbooks consider the extreme case of a competitive labor market, i.e., the case when $\xi_l = \infty$, which yields $\theta(l) = 1/\xi_f(l)$, and so $\theta(l)$ becomes the inverse of the elasticity of scale. In this case, increasing returns to scale are necessary and sufficient for a decreasing average cost curve.

In general, increasing returns to scale are necessary but not sufficient for a decreasing average cost curve. This is because $\theta(l) < 1$ requires $\xi_f(l) > 1$ (i.e., increasing returns to scale). However, if the monopolist holds significant labor market power, so that the residual labor supply elasticity is relatively low and the elasticity of the cost function relatively high, $\theta(l)$ tends to be large—in particular, greater than one. In fact, in the extreme case of absolute labor market power, i.e., the case when $\xi_l \approx 0$ for a range of hired labor l, we have $\theta(l) \approx \infty$, implying that the monopolist's average cost function is increasing (for said range of l).

Let us now consider the case of capital input markets.[3] Suppose that the monopolist owns the capital used in its production process and chooses how much of its capital stock, K, is used as an input in its own production. Alternatively, the monopolist can loan capital to other agents.

The production function is denoted by $f(k)$, where k measures 'capital.' Let $r(k)$ be a decreasing function representing the inverse demand for capital facing the monopolist. The price elasticity of the firm-level demand for capital,

$$\epsilon_k = \frac{1}{r'(k)} \cdot \frac{r(k)}{k},$$

which ranges between 0 and $-\infty$, can be taken as a measure of the monopolist's capital market power as a supplier of capital.[4] A perfectly inelastic (resp., elastic) demand represents the case of absolute market power (resp., the absence of market power).

If the monopolist loans κ units of capital and uses $K - \kappa$ units of capital in the production of its own final good, the cost of using an extra (marginal) unit of capital in the production of the final good is the opportunity cost of loaning that unit, i.e.,

$$MR(\kappa) = r(\kappa) + r'(\kappa)\kappa,$$

which represents the monopolist's marginal revenue evaluated at the quantity of capital loaned, κ. Intuitively, this opportunity cost consists of the forgone unit price of capital, $r(\kappa)$, minus the extra revenue from the increase in the price of capital paid for the inframarginal units, $r'(\kappa)\kappa$; the increase in the price of capital results from the marginal reduction in the supply of loanable funds.

Hence, the monopolist's marginal economic cost of using k units of capital in the production of the final good (which gives a corresponding loan size of $K - k$ units of capital) is given by

$$c'(k) = MR(K - k),$$

[3] The essence of the argument that follows also applies to the case of 'land.'

[4] The Lerner index of a firm's market power (Lerner, 1934) can be expressed solely in terms of the firm-level price elasticity of demand.

while the monopolist's total economic cost of using k units of capital in the production of the final good is given by

$$c(k) = \int_{K-k}^{K} MR(\kappa)d\kappa. \tag{3}$$

Note that the marginal cost is negative for those values of k for which $MR(K-k)$ is negative. However, any loan supply of size $K-k$, where $MR(K-k)$ is negative, is not profit maximizing, since, at those levels, reducing the loan supply brings about extra revenue.

The *elasticity of the cost function with respect to* k, which measures the relative responsiveness of the monopolist's cost to a one-percent increase in the quantity of capital, can be written as

$$c'(k) \cdot \frac{k}{c(k)} = \frac{MR(K-k)k}{c(k)}. \tag{4}$$

Note that if the $MR(\cdot)$ function is decreasing, then this elasticity is greater than 1, since, in this case, the marginal cost exceeds the average cost:[5]

$$\frac{MR(K-k)k}{c(k)} > 1 \Leftrightarrow \frac{c'(k)}{c(k)/k} = \frac{MR(K-k)}{c(k)/k} = \frac{MR(K-k)}{(\int_{K-k}^{K} MR(\kappa)d\kappa)/k} > 1.$$

Note also that $MR(K-k)$ is expressible, in terms of the price elasticity of the firm-level demand for capital, as

$$MR(K-k) = \left(1 + \frac{1}{\epsilon_{K-k}}\right) r(K-k). \tag{5}$$

The ratio of the elasticity of the cost function with respect to k to the elasticity of scale,

$$\theta(k) = \frac{MR(K-k)k}{c(k)} \Big/ \xi_f(k) = \frac{\left(1 + \frac{1}{\epsilon_{K-k}}\right) r(K-k)k}{c(k)} \Big/ \xi_f(k),$$

measures the relative responsiveness of the total cost to the relative responsiveness of output (with respect to a one-percent increase in the quantity of hired labor). Note that

$\theta(k) < 1$ for all k if and only if the monopolist's average cost function is decreasing;
$\theta(k) = 1$ for all k if and only if the monopolist's average cost function is constant;
$\theta(k) > 1$ for all k if and only if the monopolist's average cost function is increasing.

[5] A downward sloping $MR(\cdot)$ function is sufficient but not necessary for the elasticity of the cost function with respect to k to be greater than one.

Thus, when the labor market is not competitive, the shape of the average cost function is determined by the elasticity of scale *and* the firm-level price elasticity of demand.

The extreme case of a perfectly competitive capital market, i.e., the case when the price elasticity of the firm-level demand for capital is $-\infty$, corresponds to the case of a flat inverse demand function $r(\cdot)$ (hence a flat $MR(\cdot)$ function which coincides with $r(\cdot)$), implying that the elasticity of the cost function with respect to k is equal to one. In this case, $\theta(k) = 1/\xi_f(k)$, implying that increasing returns to scale are necessary and sufficient for a decreasing average cost curve.

Under complete capital market power, $\epsilon_k = 0$, which gives an infinite elasticity of the cost function with respect to k, implying that the monopolist's operating average costs are increasing.

If the elasticity of the cost function with respect to k is greater than 1 (which is true, for example, if the $MR(\cdot)$ function is decreasing), then increasing returns to scale are necessary, but not sufficient, for a decreasing average cost function.

The two-input case can be handled as a simple extension of the preceding analysis.

Suppose that the production function is given by $f(l, k)$, a function of labor, l, and capital, k. We maintain the assumption that the monopolist owns the capital used in its production process and chooses how much of its capital stock, K, is used as an input in its own production. Alternatively, the monopolist can loan capital to other agents.

The monopolist's total cost function is now

$$w(l)l + c(k),$$

where $c(\cdot)$ represents the capital cost function given in (3). The elasticity of the cost function with respect to the labor input,

$$\frac{\partial(w(l)l + c(k))}{\partial l} \cdot \frac{l}{w(l)l + c(k)},$$

measures the percentage change in the total cost resulting from a one-percent increase in the quantity of labor hired. Similarly,

$$\frac{\partial(w(l)l + c(k))}{\partial k} \cdot \frac{k}{w(l)l + c(k)}.$$

measures the percentage change in the total cost resulting from a one-percent increase in the quantity of capital used in the firm's production process.

Note that

$$\frac{\partial(w(l)l + c(k))}{\partial l} \cdot \frac{l}{w(l)l + c(k)} = \left(\frac{1}{\xi_l} + 1\right) \frac{w(l)l}{w(l)l + c(k)} = \left(\frac{1}{\xi_l} + 1\right) \alpha_l(l, k),$$

where ξ_l is the familiar residual labor supply elasticity and $\alpha_l(l,k)$ represents the cost share of the labor input. Similarly,

$$\begin{aligned}
&\frac{\partial(w(l)l+c(k))}{\partial k}\cdot\frac{k}{w(l)l+c(k)}\\
&=\left(c'(k)\cdot\frac{k}{c(k)}\right)\frac{c(k)}{w(l)l+c(k)}=\left(c'(k)\cdot\frac{k}{c(k)}\right)\alpha_k(l,k)\\
&=\left(\frac{MR(K-k)k}{c(k)}\right)\alpha_k(l,k)=\left(\frac{\left(1+\frac{1}{\epsilon_{K-k}}\right)r(K-k)k}{c(k)}\right)\alpha_k(l,k),
\end{aligned}$$

where $\alpha_l(l,k)$ represents the cost share of capital and the last two equalities follow from (4) and (5), respectively.

Consequently, the relative responsiveness of the total cost to a one-percent increase in the quantity of all inputs is given by

$$\begin{aligned}
&\frac{\partial(w(l)l+c(k))}{\partial l}\cdot\frac{l}{w(l)l+c(k)}+\frac{\partial(w(l)l+c(k))}{\partial k}\cdot\frac{k}{w(l)l+c(k)}\\
&=\left(\frac{1}{\xi_l}+1\right)\alpha_l(l,k)+\left(\frac{MR(K-k)k}{c(k)}\right)\alpha_k(l,k)\\
&=\left(\frac{1}{\xi_l}+1\right)\alpha_l(l,k)+\left(\frac{\left(1+\frac{1}{\epsilon_{K-k}}\right)r(K-k)k}{c(k)}\right)\alpha_k(l,k).
\end{aligned}$$

The elasticity of scale is now given by

$$\xi_f(l,k)=\frac{\partial f(l,k)}{\partial l}\cdot\frac{l}{f(l,k)}+\frac{\partial f(l,k)}{\partial k}\cdot\frac{k}{f(l,k)},$$

and it measures the relative responsiveness of output to a one-percent increase in the quantity of all inputs.[6] Note that

$\xi_f(l,k)>1$ for all (l,k) if and only if the production function $f(\cdot)$ exhibits increasing returns to scale;

$\xi_f(l,k)=1$ for all (l,k) if and only if the production function $f(\cdot)$ exhibits constant returns to scale;

$\xi_f(l,k)<1$ for all (l,k) if and only if the production function $f(\cdot)$ exhibits decreasing returns to scale.

[6] Note that we are adopting a "long-run" perspective, since no input is fixed over the underlying time horizon.

The ratio of the elasticity of the total cost function with respect to both inputs to the elasticity of scale,

$$\theta(l,k) = \frac{\left(\frac{1}{\xi_l}+1\right)\alpha_l(l,k) + \left(\frac{MR(K-k)k}{c(k)}\right)\alpha_k(l,k)}{\xi_f(l,k)} = \frac{\left(\frac{1}{\xi_l}+1\right)\alpha_l(l,k) + \left(\frac{\left(1+\frac{1}{\epsilon_{K-k}}\right)r(K-k)k}{c(k)}\right)\alpha_k(l,k)}{\xi_f(l,k)}, \tag{6}$$

determines the shape of the average cost function:

$\theta(l,k) < 1$ for all (l,k) if and only if the monopolist's average cost function is decreasing;

$\theta(l,k) = 1$ for all (l,k) if and only if the monopolist's average cost function is constant;

$\theta(l,k) > 1$ for all (l,k) if and only if the monopolist's average cost function is increasing.

Note that, in the case of a perfectly elastic firm-level capital demand curve, the $MR(\cdot)$ is flat, implying that

$$c(k) = \int_{K-k}^{K} MR(\kappa)d\kappa = MR(K-k)k, \quad \text{for all } k.$$

Thus, in the extreme case of competitive input markets, we have $\xi_l = \infty$ (a perfectly elastic firm-level labor supply curve) and $\epsilon_{K-k} = -\infty$ (a perfectly elastic firm-level capital demand curve), and $\theta(l,k)$ reduces to

$$\theta(l,k) = 1/\xi_f(l,k),$$

implying that the shape of the average cost curve is fully determined by the elasticity of scale. In this case, the average cost function is decreasing if and only if the production function exhibits increasing returns to scale.

In general, if $\frac{MR(K-k)k}{c(k)} > 1$ for all k (i.e., if the marginal cost of capital exceeds the average cost of capital, which is true, for example, if the $MR(\cdot)$ curve is decreasing), since $1 + \frac{1}{\xi_l} > 1$ for all l, (6) implies that $\theta(l,k) < 1$ for all (l,k) (i.e., the firm's average cost function is decreasing) only if $\xi_f(l,k) > 1$ for all (l,k), i.e., only if the production function exhibits increasing returns to scale. While, in this case, increasing returns to scale are necessary for a decreasing average cost function, they are not, in general, sufficient.

In the case when the firm holds significant market power in the market for inputs, so that $\xi_l \approx 0 \approx \epsilon_{K-k}$ for a range of input levels, the firm's average cost function is increasing over said range.

We conclude this note with a brief discussion of the implications of the analysis for regulatory policy. The textbook example of a natural monopoly is an industry whose production processes exhibit declining average costs. Such industries are argued to be more efficiently served by monopolies, given their cost advantages, which also act as a barrier to entry, since an incumbent firm can always undercut the prices charged by an entrant. In the presence of decreasing average costs, standard economic theory advocates the awarding of an exclusive franchise to serve the market to a single firm. Such exclusive rights typically coexist with other regulatory policies (e.g., price regulation) aimed at restraining market power.[7]

On the other hand, the modern literature on merger efficiencies, which dates back to Williamson (1968), often refers to merger efficiencies as average cost reductions resulting from economies of scale. This interpretation has been underscored by law scholars and economists alike to this day.[8]

Thus, economies of scale and declining average costs play a central role in regulatory policy and are often regarded as a prominent rationale for legal promotion and consolidation of monopoly power via the awarding of exclusive franchises or by means of (vertical or horizontal) mergers. However, if legal entry restrictions and corporate amalgamation are accompanied by a significant outgrowth of market power in input markets, such regulations are likely to bring about a complete reversal of the conditions warranting their implementation, ultimately resulting in diseconomies of scale.

This prediction is consistent with evidence reported in the recent empirical literature on labor markets, which documents a negative correlation between the residual labor supply elasticity and measures of labor market concentration; and a tendency for markets with higher concentration or lower residual labor supply elasticity to have significantly lower wages.[9] In this context, our analysis puts forward a novel rationale for antitrust enforcement as a means to restraining monopsony power.

Author contributions OCN wrote the entire article.

Declarations

Competing interests The authors declare no competing interests.

References

Azar, J., Marinescu, I., & Steinbaum, M. (2019). Measuring labor market power two ways. *AEA Papers and Proceedings, 109*, 317–21. https://doi.org/10.1257/pandp.20191068

Azar, J., Marinescu, I., & Steinbaum, M. (2022). Labor market concentration. *Journal of Human Resources, 57*, S167–S199. https://doi.org/10.3368/jhr.monopsony.1218-9914R1

Coate, M., & Andrew, H. (2009). *Merger Efficiencies at the Federal Trade Commission 1997-2007*. https://www.ftc.gov/reports/merger-efficiencies-federal-trade-commission-1997-2007.

[7] See, e.g., Joskow (2007) for a survey.

[8] See, e.g., U.S. Department of Justice (1968, 1982, 1984, 1992, 1997, 2023), Williamson (1968), Muris (1980), Fisher and Lande (1983), Werden (1997), Kolasky and Dick (2003), and Coate and Andrew (2009).

[9] See, e.g., Azar et al. (2019).

Fisher, A. A., & Lande, R. H. (1983). Efficiency considerations in merger enforcement. *California Law Review, 71*, 1580–1696. https://doi.org/10.2307/3480297
Joskow, P. L. (2007). Chapter 16 regulation of natural monopoly. *Handbook of Law and Economics, 2*, 1227–1348. https://doi.org/10.1016/S1574-0730(07)02016-6
Khan, L.H. (2016). Amazon's antitrust paradox. *Yale Law Journal*, *126*, https://digitalcommons.law.yale.edu/ylj/vol126/iss3/3.
Kolasky, W. J., & Dick, A. R. (2003). The merger guidelines and the integration of efficiencies into antitrust review of horizontal mergers. *Antitrust Law Journal, 71*, 207–251.
Kreps, D.M. (2004). *Microeconomics for Managers*: Norton, 1–652.
Lerner, A. P. (1934). The concept of monopoly and the measurement of monopoly power. *The Review of Economic Studies, 1*, 157–175. https://doi.org/10.2307/2967480
Muris, T. J. (1980). The efficiency defense under section 7 of the clayton act. *Case Western Reserve Law Review, 30*, 381–432.
Naidu, S., Posner, E., & Weyl, G. (2018). Antitrust remedies for labor market power. *Harvard Law Review, 132*, 536–601.
Serrano, R., & Feldman, A. M. (2013). *A Short Course in Intermediate Microeconomics with Calculus*. Cambridge: Cambridge University Press.
U.S. Department of Justice. (1968). *1968 Merger Guidelines*. https://www.justice.gov/archives/atr/1968-merger-guidelines.
U.S. Department of Justice. (1982). *The 1982 Merger Guidelines And The Ascent Of The Hypothetical Monopolist Paradigm*. https://www.justice.gov/archives/atr/1982-merger-guidelines-and-ascent-hypothetical-monopolist-paradigm.
U.S. Department of Justice. (1984). *1984 Merger Guidelines*. https://www.justice.gov/archives/atr/1984-merger-guidelines.
U.S. Department of Justice. (1992). *1992 Merger Guidelines*. https://www.justice.gov/archives/atr/1992-merger-guidelines.
U.S. Department of Justice. (1997). *1997 Merger Guidelines.* https://www.justice.gov/archives/atr/1997-merger-guidelines.
U.S. Department of Justice. (2023). *2023 Merger Guidelines.* https://www.justice.gov/atr/2023-merger-guidelines.
Werden, G. (1997). An economic perspective on the analysis of merger efficiencies. *Antitrust*, *12*.
Williamson, O. E. (1968). Economies as an antitrust defense: The welfare tradeoffs. *The American Economic Review, 58*, 18–36.