Teaching and Educational Methods

Production Economics and Efficiency: An Overview

Jean-Paul Chavas[a]
[a] *University of Wisconsin*

JEL Codes: D2, D4, D6, L1
Keywords: Diversification, Efficiency, Nonconvexity, Nonlinear pricing, Production

Abstract
This paper offers an overview of production economics and its usefulness in economic analysis. It presents all key arguments of production economics and efficiency under general conditions and in an integrated manner. An empirical example illustrates how the methods can be applied. The paper also investigates four somewhat unexplored topics in production economics: (1) the effects of a nonconvex technology, (2) the role of profit maximization, (3) economies of diversification, and (4) pricing efficiency and the role of nonlinear pricing. While it is well-known in economics that competitive markets support efficient allocations under a convex technology, this result does not apply under nonconvexity. In this context, we argue that restoring efficiency can require nonlinear pricing. This seems important in evaluating economies of diversification; under nonconvexity, competitive markets can lead to inefficient specialization (which may be particularly concerning applied to environmental management). Implications for management and policy are discussed.

1 Introduction

This paper presents an overview of production economics and its usefulness in economic and welfare analysis. Much has been written on production technology, production decisions, and their implications for economic efficiency. Coelli, Rao, and Battese (2005) and Ray, Chambers, and Kumbhakar (2022) provide a good literature review on these topics. So, this paper does not provide another literature review, and its objectives are different. The first objective is to present production economics in a broad context. Our analysis holds under general conditions: it allows for market goods as well as nonmarket goods; it allows for nonconvex technology; it allows for market institutions, governments as well as contracts; and it considers the role of market pricing as well as shadow pricing for nonmarket goods (e.g., pollution, environmental services). Within this broad context, we provide some new insights into the determinants of economic efficiency.

A second objective of the paper is to provide all arguments in a unified manner. As a result, all results are presented with explicit proof. While some proofs are new, many of them are not. In this context, the reader should realize that proofs are presented for the sake of completeness (and not for the sake of novelty). A third objective is to try to reach a broad audience, including students, economists, and policymakers interested in issues related to the economics and efficiency of production activities. As a result, the discussion in the main body of the paper focuses on "story telling" and graphical illustrations. And proofs and technical arguments are presented either in footnotes or in the Appendix. This format was chosen as a compromise between being thorough and being accessible to a large audience. Hopefully, the readers will appreciate the delicate nature of this compromise and will find the presentation and discussion useful.

Our paper makes extensive use of the directional distance function proposed and discussed in Chambers, Chung, and Fare (1996, 1998). Much of our discussion builds on the usefulness of this function providing insights into the nature of technology and the evaluation of production efficiency,

including technical, allocative, and scale efficiency. As noted above, one objective of this paper is to present all arguments in an integrated manner. The analysis is general and applies under broad conditions. At the micro level, it could apply to a subsistence farm where land and family labor are used to produce food to feed the family. Or it could apply to the goods and services produced by firms in an economy, in which case the linkages between industry structure (including the size and specialization of firms) and production efficiency are of significant interest.

Our approach is general and allows us to investigate several somewhat unexplored topics in production economics: (1) the effects of a nonconvex technology, (2) the role of profit maximization, (3) economies of diversification, and (4) pricing efficiency and the role of nonlinear pricing. Here, we summarize some of our interesting results. First, we argue that profit maximization is a necessary condition of economic efficiency under general conditions. While this result is well-known under competitive markets, it continues to apply in the presence of nonmarket goods provided that their shadow prices reflect consumers' willingness-to-pay. When evaluating these prices is difficult, nonmarket mechanisms (including contracts and government policy) can help guide production decisions toward efficiency. Second, it is well-known in economics that competitive markets support efficient allocations under a convex technology (where diminishing marginal productivity holds). But this result does not apply under nonconvexity. A nonconvex technology can arise in the presence of fixed cost and/or when there are productivity gains from specialization. In such situations, competitive markets can be inefficient. We show how this inefficiency is due to uniform pricing in competitive markets. Third, we argue that, under nonconvexity, achieving efficiency can require nonlinear pricing. This is important in evaluating economies of diversification: under nonconvexity, competitive markets can lead to inefficient specialization (which may be particularly concerning when applied to environmental management). In this case, nonlinear pricing is required to restore efficiency. In general, our analysis stresses the need to consider nonlinear pricing in the efficiency analysis of production systems.

Finally, this paper is intended as a teaching and educational piece. As such, it discusses how the methods can be applied. For that purpose, it includes an empirical example presented in two files available as supplemental material to this paper. Using R software, the example starts with data used to estimate production technology, and it proceeds with an empirical assessment of production efficiency. The example provides useful guidance to students and economists on how to assess economic efficiency, information that can help inform managers and policymakers in finding ways to improve firm management, resource allocation, and the functioning of the economy.

The rest of the paper proceeds as follows. The nature of production technology and its evaluation are presented in section 2. Section 3 discusses economic efficiency and its implications for production decisions. Organizational efficiency is examined in section 4, with a focus on both scale efficiency and economies of diversification. An empirical example is discussed in section 5. Finally, some implications of our results are discussed in section 6.

2 Production Technology

Consider a production process involving m netputs $y = (y_1, \dots, y_m) \in \mathbb{R}^m$. We use the netput notation: "good outputs" which benefit consumers are positive (with $y_j \geq 0$ for the j-th output) and inputs (e.g., labor) or "bad outputs" (e.g., pollution having adverse effects on consumer welfare) are negative (with $y_k \leq 0$ for the k-th input). The production technology is denoted by the set T, where $y \in T$ means that producing netputs y is feasible under technology T. Below, we assume that the set T is closed and has a non-empty interior. However, we do not assume that the set T is necessarily convex. Note the generality of the approach. First, the set T can represent a multi-input, multi-output technology where several outputs are produced from a joint production process. As noted in the introduction, the analysis is

general. For example, the technology T could represent the case of a subsistence farm where land and family labor are used to produce food to feed the family. Or it could represent aggregate production in an economy where firms produce goods and services affecting consumer welfare. Second, the netputs y can include market goods as well as nonmarket goods. When y represents market goods, then outputs are the goods that are sold in the marketplace, while inputs are purchased in the marketplace. The vector y can also include nonmarket goods (e.g., quality attributes) as well as "nonmarket bads" (e.g., externalities such as pollution that have adverse effects on productivity or human welfare).

Third, the approach covers situations where the production process is dynamic, in which case y would involve netputs in different time periods with the feasible set T capturing intertemporal effects of input-output choices. Finally, production uncertainty can be introduced in the analysis by defining netputs to be state-contingent, allowing uncertain "states of nature" to affect production possibilities (Debreu 1959).

We are looking for some convenient representation of the technology T. For that purpose, let $g = (g_1, \dots, g_m) \in \mathbb{R}_+^m$ be a reference bundle satisfying $g \neq 0$. Following Chambers et al. (1996), for a given g, define the directional distance function:

$$\begin{aligned} D(y,T) &= \max_{\beta} \{\beta : (y + \beta\, g) \in T\} \text{ if a maximum exists,} \\ &= -\infty \text{ otherwise,} \end{aligned} \qquad (1)$$

where β is a scalar. When finite, the directional distance function $D(y,T)$ in (1) can be interpreted as the number of units of the reference bundle g measuring the distance between point y and the upper bound of the feasible set T. While the choice of the reference bundle g is somewhat flexible, our discussion below focuses on the case where g is chosen such that its non-zero elements include only "good outputs." In this context, defining g such that one unit of g is worth \$1 gives the following nice interpretation: $D(y,T)$ in (1) would become a measure of the income gain that can be obtained moving from point y to the upper bound of the feasible set T.[1]

The directional distance function $D(y,T)$ in (1) has two general properties:

Property P1: $D(y,T) \geq 0$ when $y \in T$.[2]

Property P2: $D(y - \alpha\, g, T) = \alpha + D(y,T)$ for $\alpha \in \mathbb{R}$.[3]

As discussed below, the additivity property P2 will be very useful in the evaluation of industry organization.

---

[1] When applied to an economy in the presence of externalities across firms, the production technology T must be defined at the aggregate level to capture the productivity effects of externalities across firms. Situations where there are no externalities across firms are a special case where $T = \sum_{j \in N} T_j$ and $y^j \in T_j$ denotes the netputs of the j-th firm under technology T_j, N being the set of producing firms in the economy. In this case, the directional distance function in (1) can be defined at the firm level, where

$$D_j(y^j, T_j) = \max_{\beta_j} \{\beta_j : (y^j + \beta_j\, g) \in T_j\} \qquad (1')$$

assuming that a maximum exists at $y^j, j \in N$, with $D(y,T) = \sum_{j \in N} D_j(y^j, T_j)$, and $y = \sum_{j \in N}\, y^j$.

[2] Property P1 is obtained from equation (1) noting that $y \in T$ implies that $\beta = 0$ is feasible in the maximization problem (1).

[3] To obtain property P2, note that $D(y - \alpha\, g, T) = \max_{\beta} \{\beta : (y - \alpha\, g + \beta\, g) \in T\} = \max_{\gamma} \{\alpha + \gamma : (y + \gamma\, g) \in T\} = \alpha + D(y,T)$, where $\gamma \equiv \beta - \alpha$.

Other properties of the directional distance function $D(y,T)$ will also be of interest. By definition, we say that the technology T exhibits free g-disposal if $y \in T$ implies that $(y - k\ g) \in T$ for any $y \in T$ and any $k > 0$. Three specific properties of $D(y,T)$ will prove useful.

Property P3: Under free g-disposal, $y \in T$ if and only if $D(y,T) \geq 0$.[4]

Property P4: If the set T is convex,[5] then $D(y,T)$ is a concave function of y.[6]

Property P5: Let $g = (1, 0, \ldots, 0)$ where $y = (y_1, y_c)$, y_1 is an output and y_c denotes other netputs (besides y_1). Then[7]

$$D(y,T) = f_1(y_c,T) - y_1, \tag{2}$$

where

$$f_1(y_c,T) = \max_{\gamma} \{y_1 : (\gamma, y_c) \in T\}. \tag{3}$$

Under technology T, equation (3) defines $f_1(y_c,T)$ as a standard production function (or production frontier) measuring the largest feasible output y_1 that can be produced given y_c. Property P5 has supported much empirical research in production economics. Indeed, from property P3, under free y_1-disposal, the production function $f_1(y_c,T)$ in (3) provides a complete representation of the technology T. In this case, $(y_1, y_c) \in T$ is equivalent to $y_1 \leq f_1(y_c,T)$, an intuitive result that is commonly used in the teaching of production economics. Importantly, this result holds with or without convexity. Finally, from property P4 and equation (2), the set T being convex implies that the production function $f_1(y_c,T)$ is concave in y_c, corresponding to the concept of diminishing marginal productivity. While convex technology (or diminishing marginal productivity) is often assumed in production economics, there are situations where this assumption does not hold.

---

[4] The proof of property P3 is obtained in two steps. First, from P1, note that $y \in T$ implies that $D(y,T) \geq 0$. To prove the converse, assume that $D(y,T) \geq 0$. Then equation (1) implies that $(y + D(y,T)\ g) \in T$. Under free g-disposal, this implies that $y \in T$.

[5] The set T is convex if it satisfies $[\theta\ y + (1-\theta)\ y'] \in T$ for any $y \in T$, $y' \in T$ and any $\theta \in [0,1]$.

[6] To prove P4, choose any two points y_a and $y_b \in \mathbb{R}^m$. First, assume that $D(y_a,T)$ and $D(y_b,T)$ are finite. It follows from (1) that $y'_a \equiv (y_a + D(y_a,T)\ g) \in T$ and $y'_b \equiv (y_b + D(y_b,T)\ g) \in T$. The set T being convex implies that $[\theta\ y'_a + (1-\theta)\ y'_b] \in T$ for any $\theta \in [0,1]$. From property P1, it follows that $D[\theta\ y'_a + (1-\theta)\ y'_b, T] \geq 0$. From property P2, $D[\theta\ y'_a + (1-\theta) y'_b, T]$ can be written as

$$\begin{aligned} D[\theta\ y'_a + (1-\theta)\ y'_b, T] &\equiv D[\theta\ (y_a + D(y_a,T)\ g) + (1-\theta)\ (y_b + D(y_b,T)\ g), T] \\ &= D[\theta\ y_a + (1-\theta)\ y_b, T] - \theta\ D(y_a,T) - (1-\theta)\ D(y_b,T). \end{aligned}$$

Combining these two expressions gives

$$D[\theta\ y_a + (1-\theta)\ y_b, T] \geq \theta\ D(y_a,T) + (1-\theta)\ D(y_b,T),$$

which proves that $D(y,T)$ is concave in $y \in \mathbb{R}^m$ when $D(y_a,T)$ and $D(y_b,T)$ are finite.

Second, assume $D(y_a,T) = -\infty$ and/or $D(y_b,T) = -\infty$. Then the above inequality continues to hold, which concludes the proof.

[7] To prove P5, let $g = (1, 0, \ldots, 0)$ and $y = (y_1, y_c)$. Then,
$D(y,T) = \max_{\beta} \{\beta : (y_1 + \beta, y_c) \in T\} = \max_{\gamma} \{\gamma - y_1 : (\gamma, y_c) \in T\} = \max_{\gamma} \{\gamma : (\gamma, y_c) \in T\} - y_1$, where $\gamma \equiv y_1 + \beta$.

This is illustrated in Figure 1, where $y = (y_1, y_2)$, y_1 being an output, y_2 being an input, and the production function $f_1(y_2, T)$ is given by the line AF as an upper bound to the feasible set T. In Figure 1, the production function exhibits a region of increasing marginal product (line AB), a region of decreasing marginal product (line BF), a maximum (at point E), and a region of negative marginal product (line EF). With y_2 being an input (written as $y_2 \leq 0$), note that the presence of negative marginal product means that the technology T does not satisfy free y_2-disposal. Finally, given $g = (1,0)$ and when evaluated at point G, the directional distance function $D(y, T)$ in (2) is given by the distance (DG).

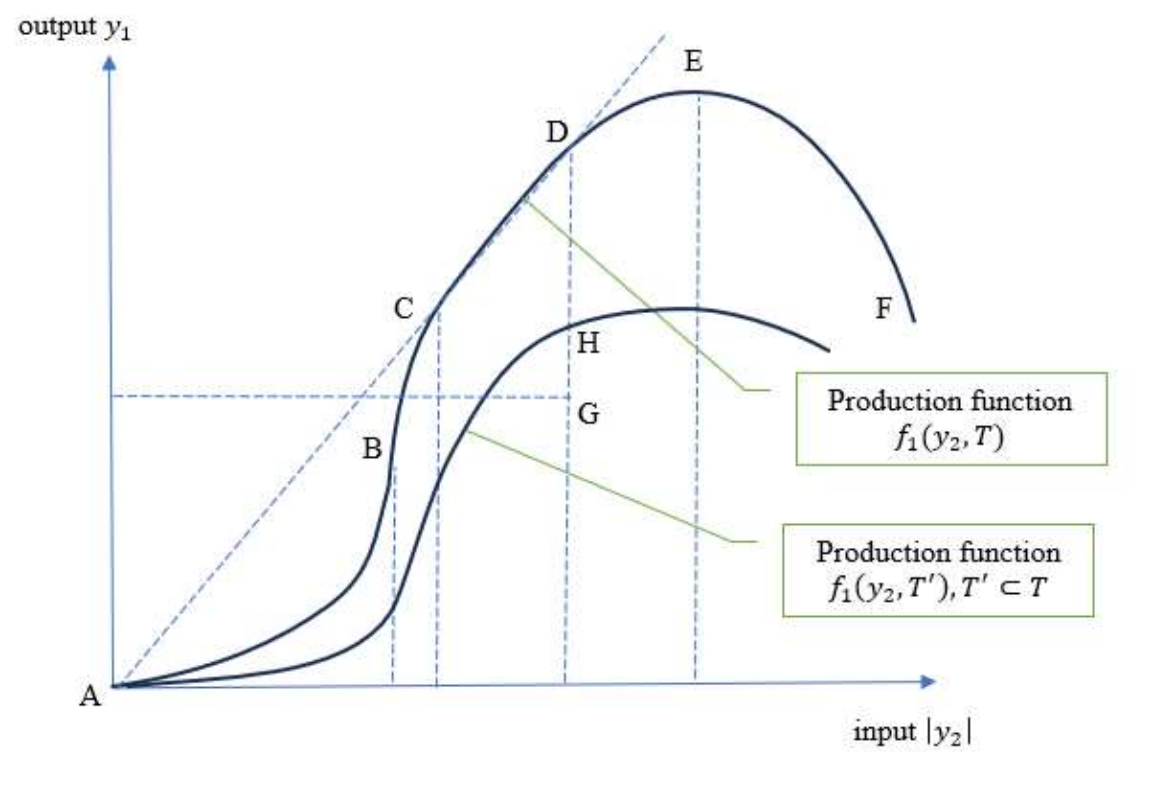


Figure 1. The production function f_1 (y_2,T) as a representation of the production technology

Technology T can also be characterized by the effects of changing the scale of operation. By definition, the technology T is said to exhibit:

$$\left\{\begin{matrix}\text{increasing return to scale (IRS)}\\ \text{constant return to scale (CRS)}\\ \text{decreasing return to scale (DRS}\end{matrix}\right\} \text{if } T \text{ satisfies } \left\{\begin{matrix} k\,T \left\{\begin{matrix}\subset\\ \supset\end{matrix}\right\} T \text{ for } \left\{\begin{matrix} k > 1 \\ k \in (0,1)\end{matrix}\right\} \\ k\,T = T \text{ for any } k > 0 \\ k\,T \left\{\begin{matrix}\supset\\ \subset\end{matrix}\right\} T \text{ for } \left\{\begin{matrix} k > 1 \\ k \in (0,1)\end{matrix}\right\} \end{matrix}\right\}, \qquad (4)$$

where $k > 0$ is a scalar rescaling feasible netputs in the set T. A characterization of returns to scale for technology T is given in the following property.

Property P6: For any $\gamma > 1$, we have[8]

$$\frac{1}{\gamma}\, D(\gamma\, y, T) \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} D(y, T) \text{ when } T \text{ exhibits } \begin{Bmatrix} \text{IRS} \\ \text{CRS} \\ \text{DRS} \end{Bmatrix}. \tag{5}$$

The term $(\gamma\, y) = (\gamma\, y_1, \ldots, \gamma\, y_m)$ in (5) reflects a change in the scale of operation, $\gamma > 0$ being a rescaling factor applied to all netputs in y. In this context, the function $[\frac{1}{\gamma}\, D(\gamma\, y, T)]$ in (5) is a measure of *ray-average productivity* under a proportional rescaling of y. A special case is obtained using P5 and equation (2), where $y = (y_1, y_c)$, $g = (1, 0, \ldots, 0)$, and $D(y, T) = f_1(y_c, T) - y_1$. Then, for any $\gamma > 1$, equation (5) can be written as:

$$\frac{1}{\gamma}\, f_1(\gamma\, y_c, T) \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} f_1(y_c, T) \text{ when the technology } T \text{ exhibits } \begin{Bmatrix} \text{IRS} \\ \text{CRS} \\ \text{DRS} \end{Bmatrix}, \tag{5'}$$

where $[\frac{1}{\gamma}\, f_1(\gamma\, y_c, T)]$ is a measure of the *ray-average product* of y_1 under a proportional rescaling of all netputs in y.

While equation (4) states global scale properties of T, there are technologies where such properties do not hold globally. Such technologies are said to exhibit variable returns to scale (VRS). Under a VRS technology, scale properties can be defined to hold "locally," thus allowing IRS, CRS, or DRS to arise in different parts of the feasible set T. For example, there are VRS technologies where the function $[\frac{1}{\gamma}\, f_1(\gamma\, y_c, T)]$ has an inverted-U shape with respect to γ.[9] In such situations, equation (5') indicates that ray-average product would be locally $\begin{Bmatrix} \text{increasing} \\ \text{constant} \\ \text{decreasing} \end{Bmatrix}$ in the scale factor γ around points exhibiting $\begin{Bmatrix} \text{IRS} \\ \text{CRS} \\ \text{DRS} \end{Bmatrix}$. This result has two useful implications. First, CRS would hold locally at scales where the ray-average product is maximized. Second, assuming differentiability at point $y = (y_1, y_c)$ where $f_1(y_c, T) > 0$, taking the derivative of $ln[\frac{1}{\gamma}\, f_1(\gamma\, y_c, T)]$ with respect to $\ln(\gamma)$ and evaluated at $\gamma = 1$ gives:

---

[8] The proof of P6 proceeds in several steps. First, note that, for some $\delta > 0$ and using (1), we have:

$$\begin{aligned} D(\delta\, y, T) &= \max_{\beta} \{\beta : (\delta\, y + \beta\, g) \in T\} \\ &= \max_{\alpha} \left\{\delta\, \alpha : (y + \alpha\, g) \in \tfrac{1}{\delta}\, T\right\} \text{ where } \alpha \equiv \tfrac{\beta}{\delta}, \\ &= \delta\, D(y, \tfrac{1}{\delta}\, T\}. \end{aligned}$$

Second, let $\gamma > 1$. Using equation (4), we have $\frac{1}{\gamma} \in (0, 1)$ and $\frac{1}{\gamma}\, T \begin{Bmatrix} \supset \\ = \\ \subset \end{Bmatrix} T$ under $\begin{Bmatrix} \text{IRS} \\ \text{CRS} \\ \text{DRS} \end{Bmatrix}$. Using equation (1), it follows that $D(y, \frac{1}{\gamma}\, T) \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} D(y, T)$ under $\begin{Bmatrix} \text{IRS} \\ \text{CRS} \\ \text{DRS} \end{Bmatrix}$.

Finally, letting $\delta = \gamma$, combining these two expressions gives equation (5).

[9] Note an inverted-U-shape ray-average product with respect to γ is not a general property of the technology T. Indeed, there exist nonconvex technologies T for which the ray-average product is not concave in the scale factor γ.

$$SE(y,T)\ \begin{Bmatrix}\geq\\=\\\leq\end{Bmatrix} 1 \text{ when the technology } T \text{ exhibits } \begin{Bmatrix}\text{IRS}\\\text{CRS}\\\text{DRS}\end{Bmatrix} \text{ around point } y, \qquad (6)$$

where $SE(y,T)$ is the scale elasticity defined as $SE(y,T) \equiv \sum_{j=2}^{m} \frac{\partial \ln [f_1(y_c,T)]}{\partial |y_j|} |y_j|$. From (6), the scale elasticity $SE(y,T)$ provides a convenient way to assess the nature of returns to scale.

These arguments are illustrated in Figure 1. With y_1 being an output and y_2 being an input, $[\frac{1}{\gamma} f_1(\gamma\, y_2, T)]$ becomes a simple average product. In Figure 1, the technology T exhibits VRS: IRS holds along the line AC where the average product is increasing in γ, CRS holds along the line CD where the average product is constant, and DRS holds along the line DEF where the average product is decreasing in γ. In this case, the average product $[\frac{1}{\gamma} f_1(\gamma\, y_2, T)]$ has an inverted-U shape with respect to γ. As expected, CRS is located at scales where the average product is maximized.[10]

Another example of production function is: $f_1(y_2,T) = \begin{Bmatrix}0\\g(|y_2|)\end{Bmatrix}$ when $|y_2| \begin{Bmatrix}<\\\geq\end{Bmatrix} k_2$, where y_2 is an input and $g(|y_2|)$ is an increasing and concave function of $|y_2|$ satisfying $g(|k_2|) = 0$. Here $k_2 \geq 0$ can be interpreted as "fixed cost" reflecting the smallest input $|y_2|$ required to produce a positive output y_1. When $k_2 = 0$ (i.e., no "fixed cost"), then the technology T is convex: it exhibits diminishing marginal productivity, and either CRS or DRS. When $k_2 > 0$ (i.e., in the presence of fixed cost), the technology T is nonconvex, the nonconvexity arising in the region where y_1 is "small." In this case, T would exhibit IRS globally when $g(|y_2|)$ is linear (i.e., when marginal product is constant), and it would exhibit VRS (with an inverted-U average product) when $g(|y_2|)$ is strictly concave. This illustrates two important effects of fixed costs: (1) any technology that exhibits fixed costs would be nonconvex, and (2) fixed costs contribute to the presence of (at least local) IRS when production is "small."

Properties P1–P6 make the directional distance function $D(y,T)$ in (1) very useful in production economics. Below, we explore the many ways $D(y,T)$ can be used.[11]

3 Production Efficiency

A key concept in economics is efficiency. An allocation is said to be Pareto efficient if there is no other feasible allocation that can make a consumer better off without making anyone else worse off. There is a broad consensus that Pareto efficiency is desirable (Graaff 1967; Luenberger 1995). In this section, we explore the implications of efficiency for production decisions. As noted above, our analysis applies under a general technology T. It covers market economies. It also allows for nonmarket goods (e.g., product quality) and "nonmarket bads" (e.g., pollution).

[10] Figure 1 can also be used to motivate the definition of returns to scale given in equation (4). Indeed, as the scaling factor $k \geq 1$ increases, the subset of $(k\ T)$ below the line (AC) (where IRS holds) becomes smaller, the subset of $(k\ T)$ below the line (CD) (where CRS holds) remains constant, and the subset of $(k\ T)$ below the line (DEF) (where DRS holds) expands.

[11] Note that other related approaches have also appeared in the literature. Letting $y = (y_O, y_I)$ where y_O are outputs and y_I are inputs, related functions include the Farrell input distance function $D_F(y,T) = \inf_{\beta>0} \{\beta: (y_O, \beta\ y_I) \in T\}$ (Farrell 1957), the Shephard input distance function $D_I(y,T) = \sup_{\beta>0} \{\beta: \left(y_O, \frac{1}{\beta}\ y_I\right) \in T\}$, the Shephard output distance function $D_O(y,T) = \inf_{\beta>0} \{\beta: \left(\frac{1}{\beta}\ y_O, y_I\right) \in T\}$ (Shephard 1970), and the shortage function $S(y,T) = \inf_{\beta} \{\beta: (y - \beta\ g) \in T\}$ (Luenberger 1995). When well defined, these functions are all related to each other as they satisfy $D_F(z,T) = 1 - D(z,T)$ and $D_I(z,T) = \frac{1}{1-D(z,T)}$ when $g = (0, |y_I|)$; $D_O(z,T) = \frac{1}{1+D(z,T)}$ when $g = (y_O, 0)$; and $S(y,T) = -D(y,T)$ (Chambers et al. 1996, 1998).

A simple way to characterize economic efficiency is to consider the case where there is a representative consumer choosing goods $x \in X \subset \mathbb{R}^m$.[12] The vector $x = (x_1, \ldots, x_m)$ includes consumer goods defined to be positive (with $x_k \geq 0$ for the k-th consumer good); it also includes labor and externalities (such as pollution) defined to be negative (e.g., with $x_k \leq 0$ when x_k denotes labor supply from the consumer). An allocation (x, y) is said to be feasible if it satisfies $x \in X$ and $x \leq y$, where y denotes production satisfying $y \in T$. When applied to consumer goods, the constraint $x \leq y$ states that consumption cannot exceed production. And when the k-th good is labor (treated as negative), $x \leq y$ means that labor demand $|y_k|$ cannot exceed labor supply $|x_k|$. We assume that initial endowments are included implicitly in T. With x and y including labor and consumer goods, it follows that the production set T represents the aggregate feasibility of combining labor and initial endowments to produce consumer goods for the representative consumer. While this involves no loss of generality, this representation will help simplify our analysis.[13]

In this context, economic efficiency can be defined in terms of allocations that maximize consumer welfare subject to feasibility. Let consumer preferences be represented by the utility function $u(x): X \to \mathbb{R}$. We assume that the set X is closed and that the utility function is continuous. Again, our analysis makes use of the reference bundle $g \in \mathbb{R}^m_+$ satisfying $g \neq 0$. We assume that consumer preferences are non-satiated in g, with $u(x + \beta\, g) > u(x)$ for all $x \in X$ and all $\beta > 1$. Below, we choose the bundle g so that one unit of g is worth \$1.

Assuming a representative consumer, an allocation (x^*, y^*) is said to be efficient if it maximizes consumer welfare (i.e., if it solves the following maximization problem):

$$U^* = \max_{x,y} \{u(x): x \leq y, x \in X, y \in T\}. \tag{7}$$

Below, we explore the implications of economic efficiency defined in (7) for production decisions. Our discussion examines several forms of efficiency, including technical efficiency, allocative efficiency, and pricing efficiency.[14]

[12] The focus on a representative consumer simplifies our efficiency analysis in an important way: it avoids issues related to income distribution. The generalization to multiple consumers is presented in Luenberger (1992, 1995) and Chavas (2015). Note that considering multiple consumers would be required in the investigation of regional trade, net exports are defined in each region as regional production minus regional consumption. But this would raise the issue of how income and welfare is distributed among consumers; while important, addressing this issue is beyond the scope of this paper.

[13] The generality of the approach can be illustrated in two examples. The first example introduces intermediate goods y_M in the analysis. Consider an economy comprised of two vertically related industries A and B: industry A uses labor y_{IA} to produce intermediate outputs y_{MA} and final outputs y_{OA} under technology T_A with $(y_{IA}, y_{MA}, y_{OA}) \in T_A$, and industry B uses labor y_{IB} and intermediate inputs y_{MB} to produce final outputs y_{OB} under technology T_B with $(y_{IB}, y_{MB}, y_{OB}) \in T_B$. Then, assuming that all goods are private goods, the aggregate technology T can be written in general as $T \equiv \{y_{IA} + y_{IB}, y_{OA} + y_{OB}: (y_{IA}, y_{MA}, y_{OA}) \in T_A, (y_{IB}, y_{MB}, y_{OB}) \in T_B, |y_{MB}| \leq y_{AM}\}$. The inequalities $|y_{MB}| \leq y_{AM}$ are part of the feasibility conditions stating that industry B cannot use more intermediate goods y_M than was produced by industry A. The constraints $|y_{MB}| \leq y_{AM}$ have shadow prices that can be analyzed like the ones associated with $x \leq y$ (including associated profit maximization conditions involving intermediate goods and the possible need for nonlinear pricing under nonconvexity). In this case, the set T provides an implicit representation of allocation decisions related to intermediate goods, thus showing how the analysis presented in this section would apply to economies involving multi-stage production processes. Second, consider the example where the feasible set T represents a joint production process of both good outputs and "bad outputs." In this case, the technology T can also be expressed in terms of two production processes: one producing good outputs while the other produces "bad outputs" (e.g., Murty, Russell, and Levkoff 2012).

[14] See Chambers et al. (1998), Coelli et al. (2005), and Ray et al. (2022) for an overview of the literature on production efficiency. Note that our approach is more general in the sense that we also evaluate pricing efficiency and examine the role of nonlinear pricing under a nonconvex technology.

3.1 Technical Efficiency and Productivity

From property P1, recall that production feasibility $y \in T$ implies that $D(y,T) \geq 0$. Consider an allocation where netputs $y' \in Y$ satisfy $D(y',T) > 0$. Using equation (1), this would imply that $[y' + D(y',T)\ g] \in T$, meaning that both production and consumption can be increased by $[D(y',T)\ g]$. Under non-satiation in g, this increase would make the consumer better off, implying that y' cannot be a solution in (7). Thus, having $D(y',T) > 0$ implies that y' cannot be efficient. This argument generates the following definition.

Definition D1: Using the directional distance function $D(y,T)$ in (1), a feasible allocation $y \in T$ is said to be *technically efficient* if it satisfies $D(y,T) = 0$, and it exhibits technical inefficiency if $D(y,T) > 0$.

From the above discussion, under non-satiation in g, economic efficiency implies technical efficiency. In addition, definition D1 identifies $D(y,T)$ in (1) as a convenient measure of technical inefficiency. In general, $D(y,T) = 0$ means that allocation y is technically efficient (as y is located on the upper bound of the feasible set T in the direction g). And $D(y,T) > 0$ means that y is technically inefficient, $D(y,T)$ measuring the number of units of g that separates the point y from the upper bound of the set T. When g is chosen such that one unit of g is worth \$1, the directional distance function has a nice economic interpretation: when $y \in T$, $D(y,T)$ in (1) provides a measure of the income loss due to technical inefficiency.

The directional distance function $D(y,T)$ is also useful in the evaluation of technological change. Consider a situation where the technology changes from T' to T, reflecting technological progress as $T \supset T'$ (i.e., as the new technology enlarges the feasible set). In this context, consider the associated change in $D(y,T)$:

$$\Delta D(y,T',T) = D(y,T) - D(y,T') \geq 0. \qquad (8)$$

Evaluated at point y, the term $\Delta D(y,T',T)$ in (8) is a measure of productivity growth reflecting the shift in technology from T' to T in the direction g. When g is chosen to have a unit price, $\Delta D(y,T',T)$ can be interpreted as the income gain generated by technological progress. Note that equation (8) can be alternatively written as

$$D(y,T) = D(y,T') + \Delta D(y,T',T) \geq 0. \qquad (8')$$

Evaluated at point y, equation (8') decomposes the technical inefficiency measure $D(y,T)$ into two additive parts: $D(y,T')$ measuring the distance to the frontier technology under T', and $\Delta D(y,T',T)$ measuring productivity growth due to technological progress. This is illustrated in Figure 1, where point G is a technically inefficient point under both T' and T. The distance (HG) is $D(y,T')$ measuring technical inefficiency under T', while (DH) is $\Delta D(y,T',T)$ measuring productivity growth due to a change from T' to T. In a way consistent with (8'), Figure 1 illustrates that the distance (DG) has two interpretations: it is $D(y,T)$ as a measure technical inefficiency under T, and it is the sum of $D(y,T')$ measuring technical inefficiency under T' and $\Delta D(y,T',T)$ measuring productivity growth. This makes it clear that technological progress and improving technical efficiency both contribute to increasing productivity.

3.2 Allocative and Pricing Efficiency

While technical efficiency requires that y is located on the boundary of T, it does not identify which point on this boundary would be efficient. Economic efficiency (as given in equation (7)) involves choosing a

point on the boundary of the set T. This means that an efficient allocation involves a choice among all possible technically efficient allocations. Since this choice requires information about the value of goods, this motivates the need to assess economic values.

Assessing the economic value of goods is implicit in the maximization problem (7). Our analysis of allocative efficiency will rely on a dual formulation to (7). Following Luenberger (1992, 1995), this dual formulation makes the valuation of goods explicit. Importantly, our analysis holds under general conditions: it applies whether the technology T is convex or not, and it allows for nonmarket goods. The efficiency formulation dual to (7) is presented in Appendix A.

This dual formulation involves a monetary valuation of netputs y given by the function $h(y) \in H$, where H is the class of functions h: $\mathbb{R}^m \to \mathbb{R}_+$, where $h(y)$ is absolutely continuous and increasing in y, and satisfying $h(0) = 0$ and $h(y + \alpha\, g) = \alpha + h(y)$ for $y \in \mathbb{R}^m$ and $\alpha \in \mathbb{R}$. Note that this last condition is a price normalization rule as it implies that $\frac{\partial h(y+\alpha\, g)}{\partial \alpha} = \frac{\partial h(y+\alpha\, g)}{\partial y}\, g = 1$. When one unit of g is worth \$1, this means that $h(y)$ provides a monetary valuation of y expressed in dollars. In this context, the function h represents a pricing scheme. As shown in Appendix A (lemma 3), under efficiency, the derivatives of $h(y)$, $\frac{\partial h(y)}{\partial y} = [\frac{\partial h(y)}{\partial y_1}, \dots, \frac{\partial h(y)}{\partial y_m}]$, can be interpreted as the (shadow) prices of netputs reflecting their marginal benefit to society. This interpretation is general: when y_j is a market good, $\frac{\partial h(y)}{\partial y_j}$ would be the market price of y_j; alternatively, it would be its shadow price when y_j is a nonmarket good.

While our analysis allows $h(y)$ to be nonlinear, there is an important special case: the situation where $h(y)$ is linear: $h(y) = p\, y = \sum_{j=1}^{m}\ p_j\, y_j$ where $p_j = \frac{\partial h(y)}{\partial y_j} > 0$ is the price of the j-th netput. Having $h(y) = p\, y$ represents a situation of uniform pricing where $p = (p_1, \dots, p_m)$ are the prices for the netputs y. Such prices are called "uniform" because p_j, the unit price of y_j, does not depend on $y_j, j = 1, \dots, m$. For example, competitive markets involve "uniform prices" as such prices apply to all market transactions. More generally, we allow $h(y)$ to be a nonlinear function. This corresponds to nonlinear pricing where the prices $\frac{\partial h(y)}{\partial y}$ can change with y. As discussed in Wilson (1993), there are many examples where nonlinear pricing arises in economics (e.g., peak load pricing, volume discount pricing, two-part tariff, bundle pricing). A simple example is the pricing scheme associated with the linear spline function $h(y) = \sum_{j=1}^{m}\ p_{aj}\, y_j + \sum_{j=1}^{m}\ (p_{bj} - p_{aj})\, Q_j(y_j)$, where $Q_j(y_j) = \begin{Bmatrix} y_j - \bar{\bar{y}}_j \\ 0 \end{Bmatrix}$ if $\begin{matrix} y_j \geq \bar{\bar{y}}_j \\ y_j < \bar{\bar{y}}_j \end{matrix}$, where $\bar{\bar{y}} = (\bar{\bar{y}}_1, \dots, \bar{\bar{y}}_m)$ are spline knots. On the one hand, this includes uniform pricing as a special case when $p_{aj} = p_{bj}, j \in \{1, \dots, m\}$. On the other hand, having $p_{aj} \neq p_{bj}$ for some $j \in \{1, \dots m\}$ represents a nonlinear pricing scheme involving two-part tariffs where $p_{bj} \begin{Bmatrix} > \\ < \end{Bmatrix} p_{aj}$ means that $\begin{Bmatrix} p_{bj} - p_{aj} \\ p_{bj} - p_{bj} \end{Bmatrix}$ is a $\begin{Bmatrix} \text{price premium} \\ \text{price discount} \end{Bmatrix}$ applied when the quantity y_j is larger than $\bar{\bar{y}}_j$. As discussed below, the choice of the pricing scheme $h(y)$ is relevant in the evaluation of production efficiency.

Recall that we use the netput notation for y where outputs are positive and inputs are negative. It follows that $h(y)$ denotes profit under pricing scheme $h \in H$. As stated in lemma 4 in Appendix A, economic efficiency implies that efficient production y^* must satisfy the profit maximization condition:

$$y^* \in argmax_y \{h^*(y): y \in T\}, \tag{9}$$

where h^* is the efficient pricing scheme $h^* \in H$ satisfying:

$$h^* \in argmin_h \{\pi(h) - E(h): h \in H\}, \tag{10}$$

where $\pi(h) = max_y \{h(y): y \in T\}$ is the profit function given in (A16), and $E(h)$ is the expenditure function defined in (A15). Equations (9) and (10) establish conditions for production efficiency and pricing efficiency, respectively.

We start with a discussion of pricing efficiency. From equation (10), it is clear evaluating pricing efficiency requires information about both production technology and consumer preferences. Note that $0 = B^* = min_h \{\pi(h) - E(h): h \in H\}$ from (A17). This suggests the following measure and definition:

$$PRI(h) = \pi(h) - E(h) \geq 0, h \in H. \quad (11)$$

Definition D2: A pricing scheme $h \in H$ is said to exhibit *pricing efficiency* if it satisfies $PRI(h) = 0$, and $h \in H$ is said to be price inefficient if $PRI(h) > 0$.

Equations (10)–(11) involve the profit function $\pi(h)$ and the expenditure function $E(h)$. First, conditional on h, maximizing profit and minimizing expenditure supports maximizing the purchasing power of the consumer, a necessary condition for economic efficiency. Second, the choice of h in equation (10) ensures that the pricing rule h^* satisfies supply-demand equilibrium: $x^* \leq y^*$ and $h^*(x^*) = h^*(y^*)$, conditions which are also necessary for efficiency. Third, equation (11) can be interpreted as a budget constraint stating that consumer expenditure $E(h)$ cannot exceed profit $\pi(h)$. Under economic efficiency, maximizing consumer welfare requires that all profit be redistributed to support consumption activities. This motivates the pricing inefficiency measure $PRI(h) = \pi(h) - E(h)$ given in equation (11), along with its use in definition D2: $PRI(h) = 0$ under pricing efficiency, while $PRI(h) > 0$ under pricing inefficiency. This provides the following intuitive interpretation of the pricing inefficiency measure in (11): $PRI(h)$ is the potential increase in the consumer's purchasing power that can be generated by choosing a pricing scheme $h \in H$ supporting economic efficiency.

Next, consider the analysis of production efficiency associated with the profit maximizing condition stated in equation (9). Note that in the presence of externalities across firms, equation (9) must apply at the aggregate level to capture the productivity effects of externalities across firms. This raises the question: can profit maximization apply at the firm level and support efficiency? The answer is yes it can under two conditions: (1) there are no externalities across firms (where $T = \sum_{j \in N} T_J$, T_j denoting the production technology of the j-th firm and N being the set of producing firms), and (2) uniform pricing is efficient (where $h^*(y) = p^* y$). Indeed, under these two conditions, equation (9) can be written as the firm-level profit maximization problem

$$y^{j*} \in argmax_{y^j} \{p^* y^j: y^j \in T_j\}, j \in N, \quad (9')$$

where $y^* = \sum_{j \in N} y^{j*}$. Yet, without these two conditions, (9) would not imply (9'), in which case firm-level decentralized profit maximization could fail to support an efficient allocation (as further discussed below).

Equation (9) suggests the following measure of production inefficiency:

$$PDI(y) = h^*(y^*) - h^*(y). \quad (12)$$

It is clear from (9) that $PDI(y) \geq 0$ for any $y \in T$. This motivates the following definition:

Definition D3: A feasible allocation $y \in T$ is said to be *production efficient* if it satisfies $PDI(y) = 0$, and it exhibits allocative inefficiency if $PDI(y) > 0$.

In general, $PDI(y)$ in (1) is a monetary measure of the profit loss associated with production inefficiency. This profit loss is zero under production efficiency, but it is positive under production inefficiency (when y does not maximize profit). Note that $h \in H$ satisfies the property $h(y + \alpha\, g) = \alpha + h(y)$. Thus, evaluated at the technically efficient point $[y + D(y,T)\, g]$, we have $h^*(y + D(y,T)\, g) = h^*(y) + D(y,T)$ for any $y \in T$. When the bundle g is chosen to have a unit price of \$1 and for any $y \in T$, this suggests that the production inefficiency measure $PDI(y)$ given in (12) can be decomposed as follows:

$$PDI(y) \equiv TEI(y) + ALI(y) \geq 0, \tag{13}$$

where

$$\begin{aligned} TEI(y) &\equiv h^*(y + D(y,T)\, g) - h^*(y) \\ &= D(y,T) \geq 0 \end{aligned} \tag{14a}$$

is a measure of technical inefficiency, and

$$ALI(y) \equiv h^*(y^*) - h^*(y + D(y,T)\, g) \geq 0 \tag{14b}$$

is a measure of *allocative inefficiency*.

Equation (13) decomposes the production inefficiency measure $PRI(y)$ into two additive components: technical inefficiency $TEI(y)$ measured by the directional distance function $D(y,T)$ and reflecting the potential profit loss associated with y being located below the upper bound of the feasible set T; and allocative inefficiency $ALI(y)$ measuring the potential profit loss generated when the technically efficient point $[y + D(y,T)g]$ does not maximize profit. Measuring technical inefficiency $TEI(y)$ by the directional distance function $D(y,T)$ is consistent with the discussion presented in section 3.1. The decomposition given in (13) implies that production efficiency ($PDI(y) = 0$) requires both technical efficiency ($TEI(y) = D(y,T) = 0$) and allocative efficiency ($ALI(y) = 0$). It also indicates that production inefficiency ($PDI(y) > 0$) can come from technical inefficiency ($TEI(y) = D(y,T) > 0$) or allocative inefficiency ($ALI(y) > 0$) or both. This last result has stimulated much research assessing the relative role of technical efficiency versus allocative efficiency in production decisions.

These results are general and apply under any technology T, whether it is convex or not. And they apply in the presence of nonmarket goods as long as the pricing scheme $h^*(y)$ reflects the shadow pricing of these goods. This is important: it means that our analysis covers efficiency analysis in a broad context, including both market economies and situations where markets are either incomplete or imperfect. As discussed below, production efficiency can be supported by markets, by contracts, and/or by government policy.

Our analysis includes as special cases some well-known results in economics. They include standard welfare theorems stating that profit maximizing behavior in competitive markets supports an efficient allocation (e.g., Arrow 1951). Indeed, when the technology T is convex, then uniform pricing is efficient. Under such conditions, competitive markets and profit maximization implement an efficient allocation (Arrow 1951; Luenberger 1995, pp. 180–181). This result has been used in favor of competitive markets due to their ability to support efficient allocations. But as discussed below, such results do not necessarily hold when the technology is not convex. And its validity in the presence of nonmarket goods has been debated (e.g., Pigou 1920; Coase 1960; Baumol and Oates 1988).

The above discussion indicates that our analysis leaves open the possibilities of supporting an efficient allocation by relying on markets (including new markets; as suggested by the welfare theorems); by using government regulations and/or pricing policy (e.g., taxes or subsidies) that can

eliminate the inefficiency created by market imperfections (Pigou 1920) and/or by using contracts (Coase 1960). Choosing the efficient netputs y^* can be done directly through government regulation or contracts. Such options have the advantage of not requiring an explicit evaluation of the pricing scheme h^* (and associated shadow prices in the presence of nonmarket goods). But choosing y^* directly requires extensive information about the nature of the production process, an issue that can limit the ability of government or contracts to achieve efficiency. Alternatively, the choice of the efficient netputs y^* can be guided by profit maximization incentives under the pricing scheme h^*. This option is simplest when markets are complete and uniform pricing is efficient: competitive markets and market-clearing prices then support an efficient decentralized allocation (as stated in welfare theorems). The option of relying on pricing to guide production decisions toward efficiency remains available when markets are imperfect and/or incomplete. This is relevant in designing government pricing policy and in choosing pricing rules in noncompetitive markets. But this option can also be more challenging for two reasons: (1) it requires an explicit evaluation of shadow prices and (2) it may require nonlinear pricing when uniform pricing is inefficient. The need for nonlinear pricing under a nonconvex technology is perhaps less well understood and is further discussed below.

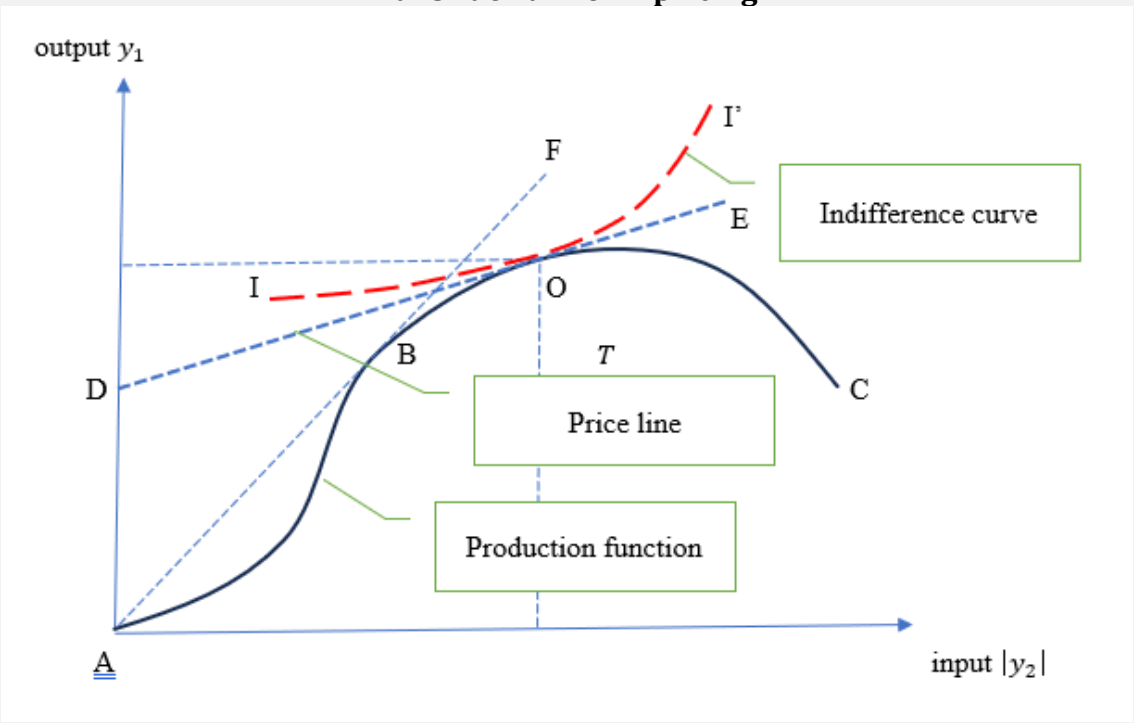
These arguments are illustrated in Figures 2 and 3 under a nonconvex technology, T. In these figures, the efficient choice is given by point O. At the efficient point O, the slope of the indifference curve (i.e., the marginal rate of substitution) is equal to the slope of the production function (i.e., the marginal rate of transformation). Figures 2a and 2b illustrate the case of input/output choices, where the technology T is nonconvex in a way similar to Figure 1. In Figures 2a and 2b, there is a market-clearing price line that goes through point O and that is tangent to both the production function (line ABOC) and the indifference curve (line I'I). However, there is an important difference: uniform pricing is efficient in Figure 2a but not in Figure 2b. In Figure 2a, the price line (DOE) corresponds to uniform pricing, the slope of (DOE) being constant. In this case, uniform pricing supports both profit maximization and economic efficiency: point O corresponds to the largest possible profit obtained under uniform pricing. This is a scenario where an efficient allocation is attained under competitive markets.

In contrast, Figure 2b presents a situation where uniform pricing is not efficient. Indeed, under uniform pricing, the price line would be (DOE). But this price line intersects the feasible set T, implying that profit maximization under this pricing scheme would provide an incentive to produce at point A (where $y_2 = 0$). But point A is inefficient. Thus, in this case, profit maximization under uniform pricing is inefficient. Yet, there is a nonlinear pricing line (D"D'OE) that would support point O as a profit maximizing point. Figure 2b is an example of a nonconvex technology where achieving efficiency requires nonlinear pricing. The efficient form of nonlinear pricing reported in Figure 2b involves a two-part tariff: the slope of the pricing line is steeper along (D'OE) but less steep along (D'D"). This means that the input-output price ratio must be lower along (D'D") than along (D'OE). Importantly, in this case, efficiency cannot be attained under any uniform pricing scheme. Indeed, a government pricing policy that generates a uniform increase in output price (or decrease in input price) would be inefficient. An example would be uniform pricing represented by the line ABF in Figure 2b: such a policy would induce choosing point B as a profit maximizing choice. But point B is inefficient (it generates lower utility than point O). This is a scenario where any uniform pricing (including pricing in competitive markets) would fail to support an efficient allocation. Note that this is also a situation where the efficient point O is located in the region of IRS.

Figures 3a and 3b illustrate the case of output choices under a nonconvex technology T. Figures 3a and 3b represent a situation where the nonconvexity of the technology T arises from large productivity gains associated with specialized production of output y_2. Again, in Figures 3a and 3b, there is a price line that goes through the efficient point O and that is tangent to both the production function (line AOC) and the indifference curve (line I'I). Uniform pricing is efficient in Figure 3a but not in Figure 3b. In Figure 3a, the price line (DOE) corresponds to uniform pricing, which supports efficient

production under profit maximization. In this case, efficiency would be attained under competitive markets. In contrast, Figure 3b presents a situation where uniform pricing is inefficient: under the price line DOE, profit maximization would provide incentives to specialize in y_2 and produce at the inefficient point C. Yet, in Figure 3b, the nonlinear pricing scheme (DOE'E") would support point O as a profit

2a. Under uniform pricing



2b. Under nonlinear pricing

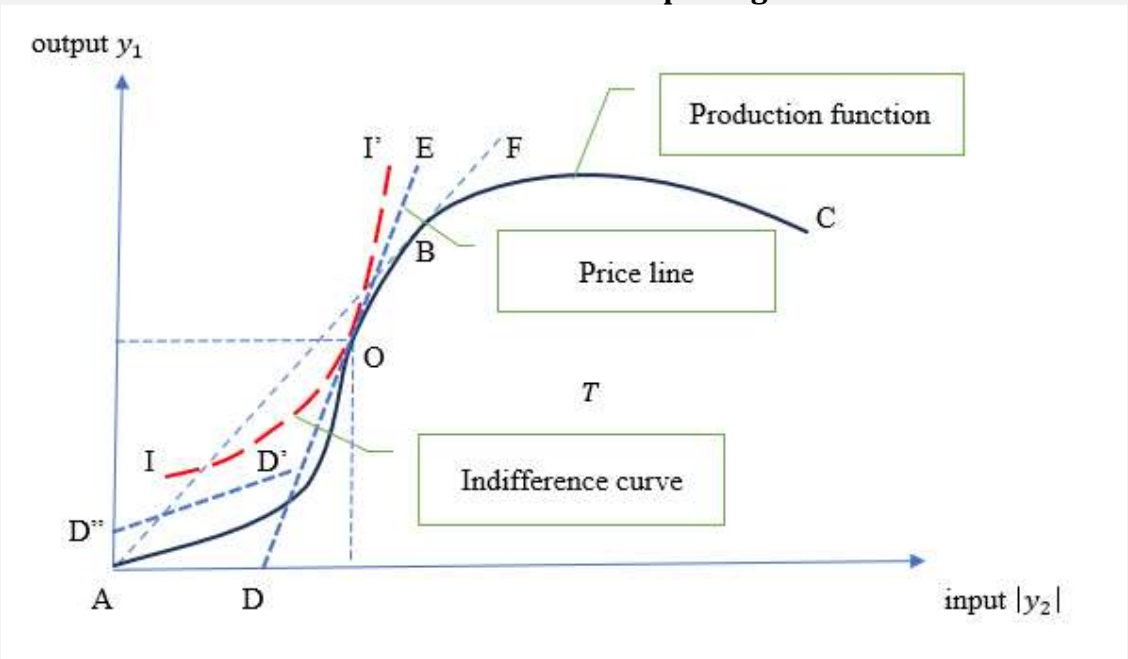
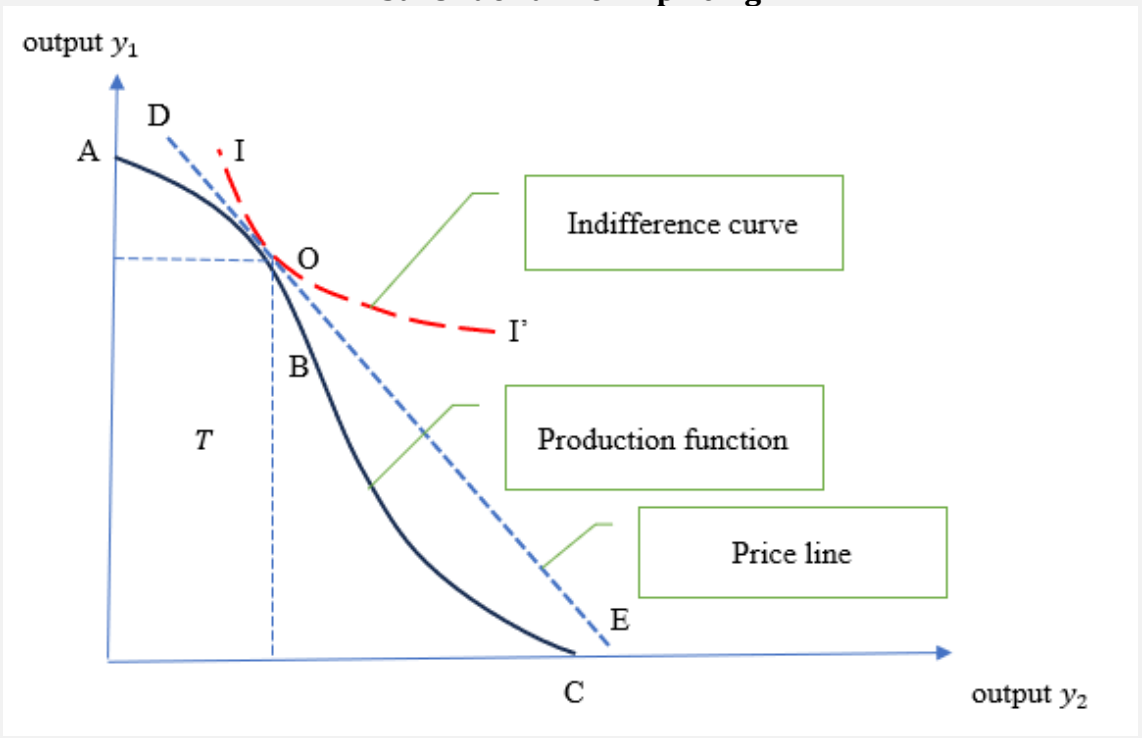


Figure 2. Efficient input-output allocation under technology *T*

3a. Under uniform pricing



3b. Under nonlinear pricing

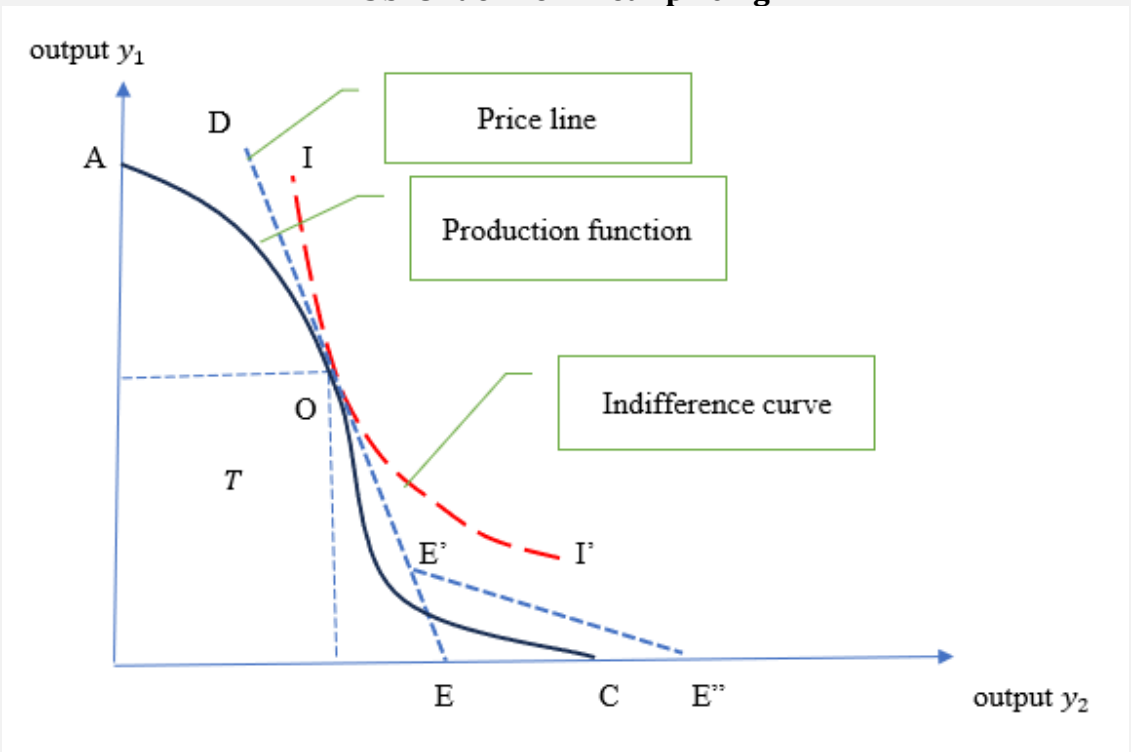


Figure 3. Efficient output allocation under technology *T*

maximizing point. These results have two implications. First, the efficient point O occurs in the nonconvex region of T (where diminishing marginal productivity does not hold), illustrating that the failure of uniform (or competitive) pricing in supporting efficiency comes from nonconvexity in the

technology. Second, point O is efficient because consumer preferences favor diversification in the goods (y_1, y_2). By preferring point O over point C, the consumer is willing to forgo the productivity gains associated with the specialized production of y_2 at point C. This raises the question: how to evaluate the economic tradeoffs between efficiency and the productivity benefits of specialization? This issue is further discussed below.

4 Organizational Efficiency

In previous sections, the feasible set T is taken as a general representation of the technology. When applied to an economy, T represents the feasibility of aggregate production, leaving open questions about the efficient organization of industries in a production economy. In this section, we focus attention on two sets of issues: the efficient number of firms in an industry, and the efficiency implications of diversification (or specialization) within the economy. Again, the analysis applies under general conditions, allowing for multi-output production, nonconvex technology, and the presence of nonmarket goods.

4.1 Scale Efficiency

Evaluating the efficiency of the number of firms in an industry is closely linked with the nature of returns to scale. Under technology T, consider an industry where the netputs $\bar{y} \in T$ are produced by n active firms, where y^j denotes the netputs of the j-th firm, $j \in N \equiv \{1, \ldots, n\}$. Below, we assume that all netputs are private goods, that there is no externality across firms, and that each firm faces the same technology T.[15] Consider the case where $y^j = s_j\, \bar{y} \in T$, where $s_j \in (0, 1]$ is the market share (or scale) of the j-th firm satisfying $\sum_{j \in N}\ s_j = 1$. Then, using equation (5) evaluated at $\bar{y} = \frac{y^j}{s_j} \in T$ and $s_j = \frac{1}{\gamma} \in (0, 1]$ gives[16]

$$D(\bar{y}, T) \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} \sum_{j \in N}\ D(s_j\, \bar{y}, T) \text{ when } T \text{ exhibits } \begin{Bmatrix} \text{IRS} \\ \text{CRS} \\ \text{DRS} \end{Bmatrix}. \tag{15}$$

Equation (15) shows that the productivity effects of different industry structures (as reflected by the number of active firms n) depend on the nature of returns to scale. Using the directional distance function, equation (15) evaluates the distance to the frontier technology comparing having one firm $(D(\bar{y}, T))$ versus n firms $(\sum_{j \in N}\ D(s_j\, \bar{y}, T))$. For given industry netputs $\bar{y}$, a larger distance means that the frontier technology is higher, corresponding to a more technically efficient industry structure. Equation (15) indicates that industry productivity (as measured by the distance to the frontier technology) would tend to $\begin{Bmatrix} \text{decrease} \\ \text{not change} \\ \text{increase} \end{Bmatrix}$ with the number of firms n under $\begin{Bmatrix} \text{IRS} \\ \text{CRS} \\ \text{DRS} \end{Bmatrix}$.

[15] Note that these assumptions can be relaxed. For example, we could introduce public goods z in the analysis by letting the industry technology be $T_0(z)$, where $y \in T_0(z)$ means that the private netputs y can be produced conditional on public goods z. In this case, after replacing T by $T_0(z)$, the analysis of scale efficiency presented below would hold conditional on the public goods z.

[16] Evaluating equation (5) at $\bar{y} = y^j/s_j$ and $s_j = \frac{1}{\gamma}$ gives $s_j\, D(\bar{y}, T) \begin{Bmatrix} \geq \\ = \\ \leq \end{Bmatrix} D(s_j\, \bar{y}, T)$ when T exhibits $\begin{Bmatrix} \text{IRS} \\ \text{CRS} \\ \text{DRS} \end{Bmatrix}$. Summing this expression across all $j \in N$ and using $\sum_{j \in N}\ s_j = 1$ yield equation (15).

This result has several important implications under IRS or CRS.[17] First, consider the case where IRS holds globally. Under global IRS, equation (15) shows that the frontier technology is highest when there is a single firm in the industry. This gives the well-known result: a natural monopoly is the most technically efficient way to produce under global IRS. Monopoly situations are problematic in two ways: (1) in market economies, monopolies have incentives to choose noncompetitive prices that have adverse effects on consumer welfare, and (2) IRS is a form of nonconvexity where uniform pricing is inefficient (as discussed in section 3). Government regulations offer a possible solution. Other options include using contracts and/or implementing nonlinear pricing.

Second, consider the case where the technology exhibits constant return to scale (CRS) globally. Then, equation (15) means that the frontier technology is the same irrespective of the number of firms in the industry. Thus, under global CRS, the industry structure (given by the number of firms n) has no effect on the technical efficiency of the industry.

Third, consider the case where the technology exhibits variable returns to scale (VRS) (i.e., where returns to scale can vary between IRS, CRS, and DRS in different parts of the feasible set T). Define $T_c \equiv \{k\ y: y \in T \text{ for any } k \geq 0\} \supset T$, where T_c is the smallest CRS technology that contains T. Assume that the function $\{\frac{D(s\ \bar{y},T)}{s}: s > 0\}$ has a maximum for some s. Then, evaluated at $\bar{y}$, we have:[18]

$$D(\bar{y}, T_c) \geq \frac{D(s\ \bar{y},T)}{s} \text{ for any scale } s > 0. \qquad (16)$$

Under VRS, equation (16) states that $D(\bar{y}, T_c)$ is the largest possible ray-average distance $\frac{D(s\ \bar{y},T)}{s}$ for all scales $s > 0$.[19] Since T_c is the smallest CRS technology that contains T, this result has two implications. First, the ray-average distance $\frac{D(s\ \bar{y},T)}{s}$ reaches it maximum $D(\bar{y}, T_c)$ at points $(s\ \bar{y})$ where T exhibits local CRS; at such points, average productivity is highest for the firm. Second, having $D(\bar{y}, T_c) > \frac{D(s\ \bar{y},T)}{s}$ implies that the technology T cannot exhibit local CRS at point $(s\ \bar{y})$. This discussion suggests the following definition:

Definition D4: Under VRS, assume that $\{\frac{D(s\ \bar{y},T)}{s}: s > 0\}$ attains a maximum for some s. For a given industry netputs $\bar{y}$, a firm is said to be *scale efficient* if it produces at a scale $s \in (0, 1]$ satisfying $D(\bar{y}, T_c) = \frac{D(s\ \bar{y},T)}{s}$, and it is scale inefficient if $D(\bar{y}, T_c) > \frac{D(s\ \bar{y},T)}{s}$.

[17] The case where the technology exhibits decreasing return to scale (DRS) globally is less interesting for two reasons. First, under global DRS, equation (15) implies that the industry would be most productive when the number of firms is infinite, an unrealistic scenario. Second, as discussed in section 2, the presence of fixed costs contributes to local IRS at least for small scales, indicating that DRS would not be expected to hold globally.

[18] To prove the inequality in (16), note that:

$$\begin{aligned} D(y, T_c) &= \max_{\alpha} \{\alpha: (y + \alpha g) \in T_c\} \\ &= \max_{\alpha, s} \{\alpha: (y + \alpha\ g) \in \frac{1}{s} T, s > 0\} \\ &= \max_{\beta, s} \{\frac{\beta}{s}: (s\ y + \beta\ g) \in T, s > 0\} \text{ where } \beta = \alpha\ s, \\ &\geq \frac{1}{s} \max_{\beta} \{\beta: (s\ y + \beta\ g) \in T\} \text{ for any } s > 0 \\ &= \frac{D(s\ y,T)}{s}. \end{aligned}$$

[19] From P5 and letting $g = (1,0, \dots ,0)$, these results would also apply using the ray-average product $\frac{f_1(s\ y_c,T)}{s}$, where $f_1(y_c, T)$ is the production function defined in equation (3).

For a firm producing at scale s, definition D4 suggests the following measure of scale inefficiency:

$$SCI(s, \bar{y}) = D(\bar{y}, T_c) - \frac{D(s\,\bar{y}, T)}{s} \geq 0. \qquad (17)$$

Under VRS, the measure $SCI(s, \bar{y})$ in (17) is the distance between the ray-average product $\frac{D(s\,\bar{y}, T)}{s}$ and $D(\bar{y}, T_c)$. Its interpretation becomes simple when the ray-average product $\frac{D(s\,\bar{y}, T)}{s}$ has an inverted-U shape with respect to s. In this context, s is an efficient scale if $SCI(s, \bar{y}) = 0$, in which case the firm produces in the region of CRS where ray-average product is maximized. And s would be deemed an inefficient firm scale if $SCI(s, \bar{y}) > 0$, in which case the firm would produce either in the region of IRS (where the ray-average product is increasing and the scale is "too small" to be efficient) or in the region of DRS (where the ray-average product is decreasing and the scale is "too large" to be efficient). This is illustrated in Figure 1, where the average production function has an inverted-U shape: firm scale is "too small" at IRS points along the line (AC), it is efficient at CRS points along the lines (CD), and it is "too large" at DRS points along the line (DF).

Applied to a market, note that the size of the market (as represented by $\bar{y}$) matters. Under VRS technology, a small market (e.g., a local market) would be more likely to see firms producing in the region of IRS. And a large market (e.g., a global market) is more likely to have firms producing in the region of CRS or DRS. This is important: as discussed in section 3 and illustrated in Figure 2b, uniform pricing can fail to be efficient under IRS, a scenario where competitive markets would not support an efficient allocation. As just noted, this is more likely to occur in small local markets. Alternatively, when the market is sufficiently large and under VRS technology, the industry would achieve its highest level of productivity when all firms produce in the region of CRS (i.e., when all active firms are scale efficient).

4.2 Economies of Diversification

The economics of diversification has been the subject of much interest.[20] One line of inquiry relates to the economics of mergers. Do mergers contribute to increasing productivity? A positive answer to this question is often seen as a motivation for a merger between two firms. Another line of inquiry comes from ecology: does increasing diversity help improve the functioning of an ecological system? In general, there is interest in evaluating the productivity effects of alternative diversification strategies.

Consider that netputs $\bar{y}$ can be produced in two ways: in an integrated manner where all inputs and outputs are parts of a single production system or in a more specialized manner where $\bar{y}$ is produced by k separate units, y^j denoting the netputs produced by the j-th unit, $j \in K \equiv \{1, \dots, k\}$. Below, we assume that all netputs y are private goods, that there are no externalities across units, and that each unit has access to the same technology T.[21] To make the evaluation of diversification meaningful, we compare the productivity of the integrated system and the specialized system assuming that they produce the same netputs $\bar{y}$, with $\sum_{j \in K}\ y^j = \bar{y}$. We also assume that $y^j \neq \bar{y}/k$ so that each y^j exhibits some form of specialization. The specialization scheme can be partial and covers both inputs and outputs. Or it can involve complete specialization among outputs. To see this, letting $y = (y_O, y_I)$, where $y_O = (y_{O1}, \dots, y_{Ok})$ are outputs and y_I are netputs, complete output specialization would correspond to $y^j = \left(y_O^j, y_I^j\right), y_I^j = \frac{\bar{y}_I}{k}$ and $y_O^1 = (\bar{y}_{O1}, 0, \dots, 0, 0), y_O^2 = (0, \bar{y}_{O2}, \dots, 0, 0)$, ...,

[20] See Baumol, Panzar, and Willig (1982) for a cost-based approach to the economics of diversification and its linkages with industry structure. Our approach is more general in the sense that we do not assume that all inputs are market goods.

[21] These assumptions could be relaxed. For example, public goods z can be introduced by replacing the technology T by $T_0(z)$, in which case our analysis of economies of diversification would still apply conditional on z.

$y_O^k = (0,0,\dots,0,\bar{y}_{Ok})$, where the j-th unit is completely specialized in the production of outputs $y_{Oj}, j \in K$, while inputs are equally divided across units.

For a given $\bar{y}$ and a specialization scheme $(y^1,\dots,y^k)$ satisfying $\sum_{j\in K}\ y^j = \bar{y}$, consider the following measure:

$$DIV = D(\bar{y},T) - \sum_{j\in K}\ D(y^j,T). \tag{18}$$

For a given technology T and using the directional distance function, DIV in equation (18) is the distance to the frontier technology comparing one integrated system ($D(\bar{y},T)$) versus k specialized units ($\sum_{j\in N}\ D(y^j,T)$). For given netputs $\bar{y}$, a larger distance means that the frontier technology is higher, corresponding to a more productive system. This suggests the following definition of economies of diversification:

Definition D5: Evaluated at netputs $\bar{y}$ and $(y^1,\dots,y^k)$ satisfying $\sum_{j\in K}\ y^j = \bar{y}$, a technology T is said to exhibit *economies of diversification* if DIV in (18) satisfies $DIV > 0$, and it exhibits diseconomies of diversification if $DIV < 0$.

By definition D5, economies of diversification arise when $DIV > 0$ in (18), corresponding to the distance function $D(y,T)$ being superadditive in y.[22] Alternatively, diseconomies of diversification arise when $DIV < 0$ in (18), corresponding to the distance function $D(y,T)$ being subadditive in y. To gain insights into the determinants of economies of diversification, note that equation (18) can be equivalently written as:

$$DIV = DIV_S + DIV_C, \tag{19}$$

where[23]

$$DIV_S = D(\bar{y},T) - \ k\ D(\tfrac{\bar{y}}{k},T)\begin{Bmatrix}\geq\\=\\\leq\end{Bmatrix} 0 \text{ when } T \text{ exhibits } \begin{Bmatrix}\text{IRS}\\\text{CRS}\\\text{DRS}\end{Bmatrix}, \tag{20a}$$

$$DIV_C = k\ D\left(\tfrac{\bar{y}}{k},T\right) - \sum_{j\in K}\ D\left(y^j,T\right)\begin{Bmatrix}\geq\\=\\\leq\end{Bmatrix} 0 \text{ if } D(y,T) \text{ is } \begin{Bmatrix}\text{concave}\\\text{linear}\\\text{convex}\end{Bmatrix} \text{ in } y. \tag{20b}$$

Equation (19) decomposes economies of diversification DIV into two additive terms: a scale component DIV_S given in (20a) and a C-component DIV_C given in (20b). From (20a), the scale component DIV_S is positive, zero, or negative when the technology exhibits IRS, CRS, or DRS, respectively. This implies that scale effects contribute positively to economies of diversification under IRS but negatively under DRS. This is intuitive: given the condition $\sum_{j\in K}\ y^j = \bar{y}$, specialized units tend

[22] The function $D(y,.)$ is said to be $\begin{Bmatrix}\text{superadditive}\\\text{subadditive}\end{Bmatrix}$ in y if it satisfies $D(\sum_{j\in K}\ y^j,.)\ \begin{Bmatrix}\geq\\\leq\end{Bmatrix}\sum_{j\in K}\ D(y^j,.)$ for all $\{y_j: j \in K\}$.

[23] Signing equation (20a) follows from equation (5), letting $\gamma = k$ and $y = \bar{y}/k$. To sign equation (20b), note that the function $D(y,T)$ being $\begin{Bmatrix}\text{concave}\\\text{linear}\\\text{convex}\end{Bmatrix}$ in y would satisfy $D(\sum_{j\in K}\ \alpha_j\ y^j,T)\begin{Bmatrix}\geq\\=\\\leq\end{Bmatrix}\sum_{j\in K}\ \alpha_j\ D(y^j,T)$ for any y^j and any $\alpha_j \in [0,1]$ where $\sum_{j\in K}\ \alpha_j = 1$. Letting $\alpha_j = 1/k$ and $\sum_{j\in K} y^j = \bar{y}$, it follows that $k\ D(\tfrac{\bar{y}}{k},T)\begin{Bmatrix}\geq\\=\\\leq\end{Bmatrix}\sum_{j\in K} D(y^j,T)$ if $D(y,T)$ is $\begin{Bmatrix}\text{concave}\\\text{linear}\\\text{convex}\end{Bmatrix}$ in y.

to be smaller than the integrated unit, implying that they would exhibit lower (higher) scale efficiency under IRS (DRS). And scale effects play no role in the economies of diversification under CRS.

As stated in (20b), the C-component DIV_C is positive, zero, or negative when the directional distance function $D(y,T)$ is respectively concave, linear, or convex in y. From property P4, $D(y,T)$ is a concave function of y when the technology T is convex. This has two implications: (1) the C-component DIV_C is necessarily non-negative when the technology T is convex, and (2) the C-component DIV_C can be negative only if the technology T is nonconvex.

From (19) and (20a)–(20b), a necessary and sufficient condition to have $DIV \geq 0$ is that $DIV_S + DIV_C \geq 0$. And sufficient conditions to have $DIV \geq 0$ are that the technology T is convex (implying $DIV_C \geq 0$ from P4 and (20b)) and exhibits CRS (implying $DIV_S = 0$ from (20a)).[24] Under such conditions, producing $\bar{y}$ in an integrated system is always more productive (as there is no productivity gain from specialization). This argument applies to ecological systems as well as firms. Applied to ecology, this means that any attempt to implement specialized production schemes in an ecological system where T is convex and exhibit CRS would be inefficient. And applied to private firms, firms would benefit from merging as mergers contribute to increasing industry productivity. But this argument seems inconsistent with the observations that many firms choose to specialize in their choice of output mix.

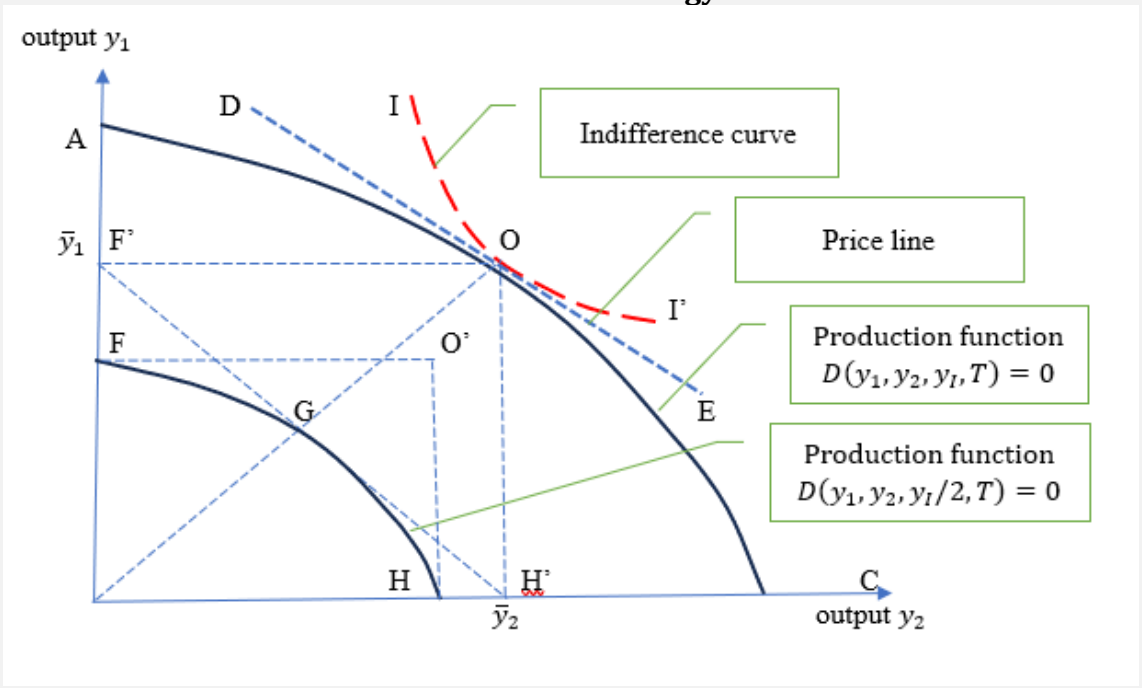
Observing that many firms and industries tend to be specialized suggests that there must exist some gains from specialization. In our analysis, this would correspond to diseconomies of diversification: $DIV < 0$. From equations (19) and (20a)–(20b), these diseconomies can arise from two situations: from DRS (implying $DIV_S \leq 0$ from (20a)) and/or from a nonconvex technology (which is required to obtain $DIV_C < 0$ from P4 and (20b)). As discussed above, having $DIV_S < 0$ can arise only under DRS and scale inefficiency. This suggests that investigating the presence of diseconomies of diversification should focus on the role of nonconvex technology, as further discussed below.

The linkages between technology and economies of diversification are illustrated in Figures 4a–4c, with $y = (y_1, y_2, y_I)$, (y_1, y_2) being two outputs and y_I being inputs. Figures 4a–4c focus on productivity effects associated with diversification/specialization among outputs (y_1, y_2). In all cases, the analysis compares an integrated system $(\bar{y}_1, \bar{y}_2, y_I)$ with a specialized system involving two units $y^1 = (\bar{y}_1, 0, y_I/2)$ and $y^2 = (0, \bar{y}_2, y_I/2)$, total netputs being the same in both systems. The main differences across figures come from the production technology: T is convex in Figure 4a; T is nonconvex in Figure 4b, the nonconvexity being associated with productivity gains from specialization in y_2; and T is nonconvex in Figure 4c, the nonconvexity being associated with productivity gains from specialization in y_1 as well as y_2. In all Figures 4, point O denotes the optimal allocation in an integrated system, located along the production function AOC (in a way similar to Figures 3). Figures 4 also show the production function for the two specialized units as represented by the line (FGH) (corresponding to the equation $D(\bar{y}_1, \bar{y}_2, y_I/2) = 0$), one unit being located at point F (producing $(\bar{y}_1, 0)$) and the other at point H (producing $(0, \bar{y}_2)$). In all Figures 4, we start with integrated production $(\bar{y}_1, \bar{y}_2)$ at point O where $D(\bar{y}_1, \bar{y}_2, y_I) = 0$. We also have $D\left(\frac{\bar{y}}{k}, T\right) = 0$ at point G, implying from (19) to (20) that $DIV_S = 0$ and $DIV = DIV_C = -D(y^1, T) - D(y^2, T)$. Thus, with $DIV_S = 0$, scale effects do not play any role, and the evaluation presented in Figures 4a–4c focuses on the role of the DIV_C component.

Figure 4a presents the effects of diversification under convex technology and CRS. Noting that point F' (corresponding to producing $(\bar{y}_1, 0)$) is above the production function (FGH), it follows that $D(y^1, T) < 0$. Similarly, $D(y^2, T) < 0$ since point H' (corresponding to producing $(0, \bar{y}_2)$) is above the production function (FGH). This implies that $DIV = DIV_C = -D(y^1, T) - D(y^2, T) > 0$ (i.e., that there are benefits

[24] Note that, in general, having strict IRS cannot arise in a technology T that is globally convex. In this sense, global convexity rules out IRS.

4a. Under a convex technology *T* and CRS



4b. Under a nonconvex technology *T* and CRS

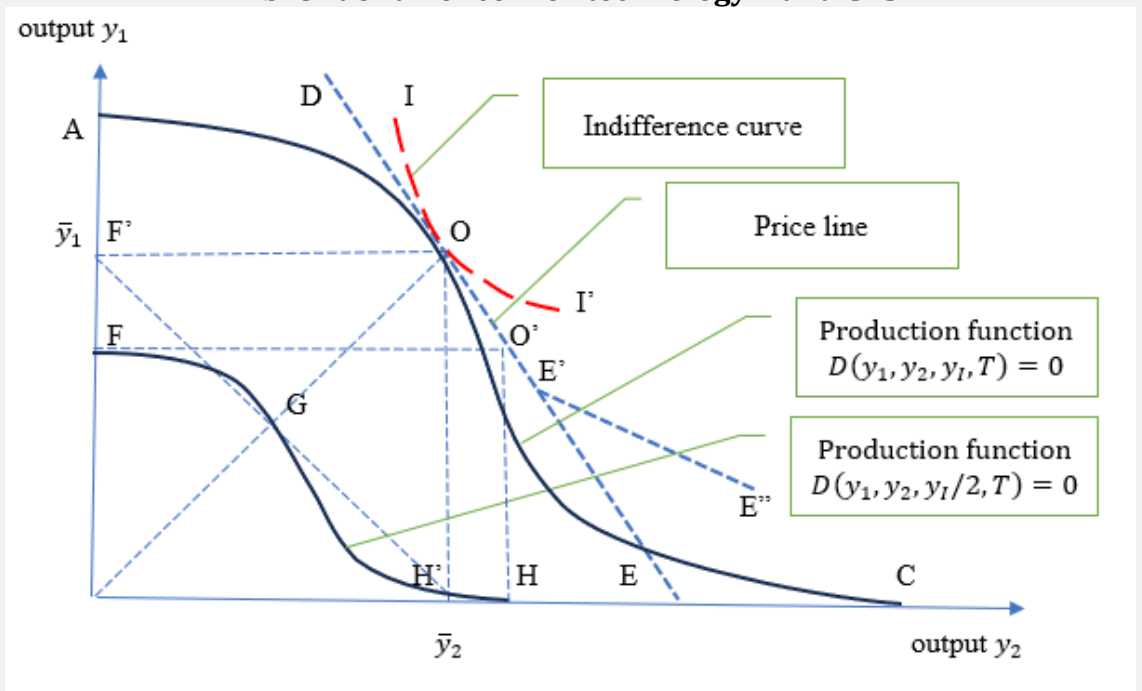


Figure 4. Economies of diversification under technology *T*

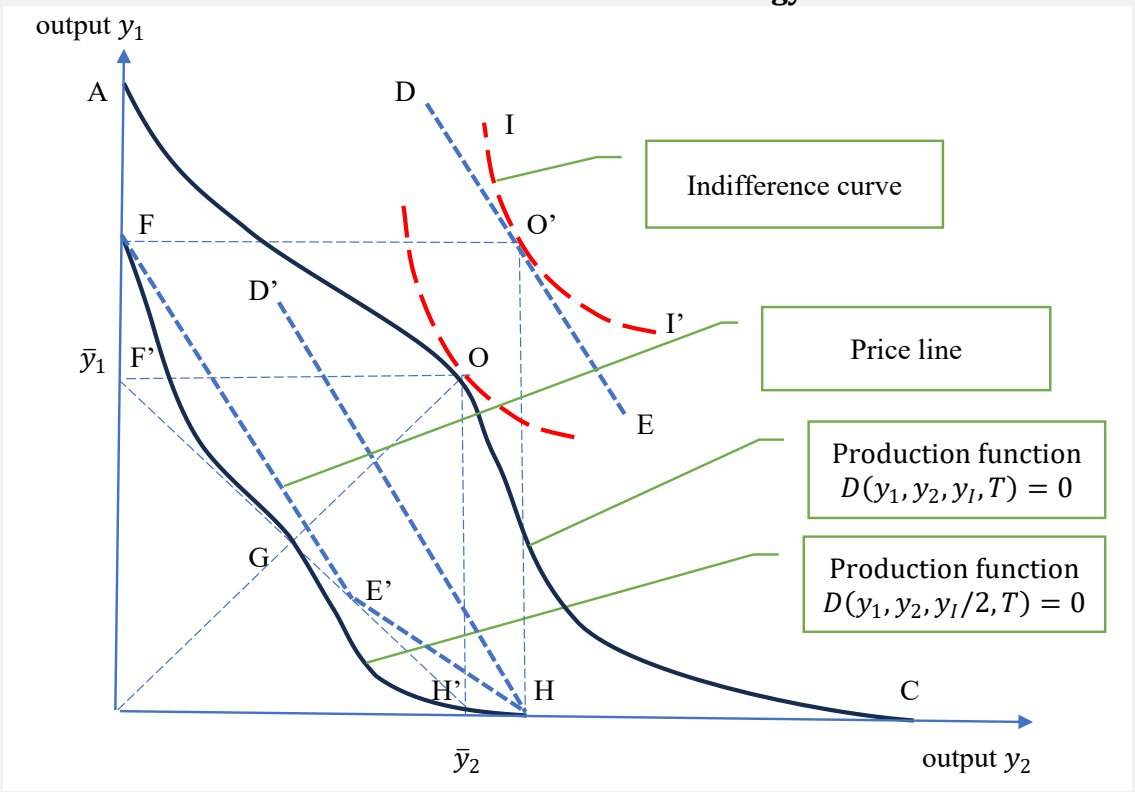


Figure 4 continued.

of diversification: the integrated system is more productive). This is consistent with our earlier discussion: an integrated system is always more efficient under convex technology and CRS.

Figure 4b presents the analysis of diversification under a nonconvex technology and CRS. In Figure 4b, the point F' (corresponding to producing $(\bar{y}_1, 0)$) is above the production function (FGH), implying that $D(y^1, T) < 0$. However, the point H' (corresponding to producing $(0, \bar{y}_2)$) is now below the production function (FGH). It follows that $D(y^2, T) > 0$, as the nonconvex technology represents a situation where specialization in y_2 increases productivity. Figure 4b shows that $DIV = DIV_C = -D(y^1, T) - D(y^2, T) > 0$ (i.e., that $D(y^1, T)$ is the dominant effect). In this case, even under nonconvex technology, the productivity gains from specialization in y_2 are not large enough: a diversified production system is more productive. The efficient allocation is then at point O obtained under an integrated production system. In this context, under nonconvexity in T, the discussion associated with Figure 3b applies: uniform pricing cannot support efficiency at point O, and nonlinear pricing (along the line DOE'E") is required. This has two implications: (1) under a nonconvex technology T, having some incentive to specialize is not sufficient to imply that specialization is efficient, and (2) nonlinear pricing may be required to support efficiency under a diversified production system.

Like Figure 4b, Figure 4c presents the effects of diversification under a nonconvex technology but under stronger productivity gains from specialization. In Figure 4c, both points F' (corresponding to producing $(\bar{y}_1, 0)$) and H' (corresponding to producing $(0, \bar{y}_2)$ are now below the production function (FGH), implying that $D(y^1, T) > 0$ and $D(y^2, T) > 0$. In this case, nonconvexity generates stronger productivity effects of specialization. Figure 4c illustrates a case where the nonconvex technology implies that $DIV = DIV_C = -D(y^1, T) - D(y^2, T) < 0$ (i.e., that specialization generates large productivity gains that make the specialized production system more efficient). In such a situation, the efficient allocation would now switch from point O to point O'. This is a scenario where efficient

production would be done by specialized units, unit 1 being completely specialized in producing y_1 (at point F) while unit 2 is completely specialized in producing y_2 (at point H). What are the implications for efficient pricing? In Figure 4c, the line (DOE) would be the pricing line under uniform pricing (as (DOE) is tangent to the indifference curve (IO'I')). But this line (or the parallel price line (D'H)) would provide incentives for both decentralized units to produce at point H, which is inefficient. To provide incentives for unit 1 to produce at point F and unit 2 at point H, nonlinear pricing is required (such as the pricing line (FE'H) representing a two-part tariff). Thus, achieving efficiency at point O' requires nonlinear pricing. This has two implications: (1) specialization in production activities can be efficient under a nonconvex technology T (when the benefits from specialization are sufficiently large), and (2) supporting efficiency under a specialized production system can require nonlinear pricing.

5 An Empirical Example

To make the analysis useful, it must be empirically tractable. This section briefly discusses an example providing some guidance to students and applied economists interested in the assessment of production efficiency. The example is presented in two files available as supplemental material to this paper: one file includes the data and analytical code (using the software package R); the other is an output file that contains the results. The application involves corn production as a function of two soil nutrients: nitrogen and phosphorus. The analysis starts with the estimation of a production function. As reported in the output file, the estimated production function is very similar to the one shown in Figures 1 and 2. First, the technology exhibits variable return to scale (VRS), making the assessment of returns to scale important. Second, it exhibits nonconvexity when inputs are "small," stressing the need to examine the role of nonconvexity in production analysis. The analysis reported in the output file proceeds with the empirical evaluation of technical efficiency, allocative efficiency, and scale efficiency. As expected, it shows how higher input prices reduce the incentive to produce. It also documents how nonlinear pricing can help restore the incentive to produce. Overall, the R analysis illustrates how the methods discussed in previous sections can be practical and applied to the investigation of economic performance. It provides useful guidance to students and economists on how to assess economic efficiency.

6 Discussion

The analysis presented in previous sections provides valuable insights into the economics of production. Production systems are often complex, making the evaluation of their management difficult. Developed under general conditions, our overview of production economics presents classical results as well as some new results on how efficiency, technology, and economic institutions can interact.

The first classical result relates to the role of competitive markets. In a way consistent with standard welfare theorems, complete and perfectly competitive markets support efficient allocations under convex technology (where diminishing marginal productivity holds everywhere). This argument provides support for relying on markets and for the development of economic policies favoring market globalization. Competitive markets allow for a decentralization of production decisions: taking market-clearing prices as given, profit maximizing firms can choose inputs/outputs in a decentralized way while supporting an efficient allocation of resources. In a market economy, this is the story where market prices are the "invisible hand" that guides economic agents toward economic efficiency (Smith 1776). This is a very attractive feature of competitive markets: they allow for decentralized decision-making while inducing profit maximizing firms to choose efficient production plans. But such results do not provide much guidance when markets are imperfect or incomplete and/or when the technology is not convex (as discussed below).

The second classical result relates to monopoly: a natural monopoly arises when the technology exhibits increasing returns to scale (IRS). In this case, in any market exhibiting uniform pricing,

monopolies are inefficient: they would have adverse effects on consumer welfare. This has motivated the development of institutional frameworks and policies trying to improve efficiency and protect consumers from monopolization, going from antitrust policy (making monopolization illegal), government regulation (constraining monopolies to behave in the consumer interest), and/or pricing policy (implementing pricing rules that would improve efficiency). This last option has often focused on uniform pricing rules (such as average cost pricing). Our analysis indicates that a focus on uniform pricing is too narrow: attempts to achieve efficiency under monopolies should also consider nonlinear pricing (e.g., Borenstein and Bushnell 2022).

A third classical result relates to the linkages between profit maximization and efficiency. From the welfare theorems, we know that, under competitive markets, profit maximization supports economic efficiency. But the general role of profit maximization is perhaps less well understood. As stated in equations (9) and (10), profit maximization is a general necessary condition for efficiency. Under pricing efficiency, profit is entirely redistributed to support consumption activities, meaning that any failure to maximize profit would reduce the consumer's purchasing power and thus be inefficient. But profit maximization in (9) relies on two conditions: (1) prices (or shadow prices) must reflect the willingness-to-pay of consumers and (2) in the presence of nonconvexity, nonlinear pricing must be allowed. This has two important implications. First, it is inappropriate to say that the profit motive is a cause of inefficiency (as sometimes argued in economic and political debates). Second, the problem is not with the profit motive; rather, the problem is with the inefficiency of pricing rules. Our analysis identifies two potential types of pricing inefficiency: (1) when prices do not reflect consumer willingness-to-pay and (2) when nonlinear pricing is required but only uniform pricing is employed. Both types of inefficiency must be eliminated to allow markets to support an efficient allocation. Doing so can be challenging, indicating that achieving efficiency relying on markets alone is likely to be difficult.

This argument stresses the need to consider nonmarket mechanisms in resource allocation. Such nonmarket mechanisms include both contracts and government policies. Contracts are agreements made among individuals in a group dictating the terms of economic relationships affecting resource allocation within the group. Contracts can be informal (e.g., among individuals in a family) or they can be formal, which means that they can be enforceable in the courts in case of breach of contract. Government policies are sets of rules affecting the behavior and welfare of individuals within a region (local government), a nation (national government), or the world. Both contracts and governments can affect resource allocation by choosing some production/consumption decisions in (y, x) directly and/or by modifying the associated incentive structure. Our analysis would apply to both situations. If elements of (y, x) are chosen directly, then there would be no need to know the associated shadow prices explicitly. This can be a significant advantage when applied to situations where assessing shadow prices is difficult. This is the realm where contracts are common (e.g., coordination among specialized workers within a firm or management of quality in a vertical supply chain). Alternatively, if contracts or governments can affect resource allocation by modifying the associated incentive structure, then an explicit evaluation of prices (or shadow prices when applied to nonmarket goods) would be required. Such evaluation would become more complex if nonlinear pricing is required to support efficiency. Yet, as noted in Wilson (1993), nonlinear pricing schemes are commonly found in the business world (e.g., volume discount, bundle pricing). This indicates that the historical decline in information cost makes the implementation of nonlinear pricing schemes easier for both contracts and government policy.

Our analysis has focused attention on the role of nonconvex technology. This is important as nonconvex technologies arise under two scenarios: (1) the presence of fixed cost and (2) situations where there are benefits from specialization. Both scenarios seem rather common, stressing the relevance of understanding the economics of production under nonconvexity. As discussed in section 2, fixed costs contribute to both nonconvexity and IRS. And as argued by Smith (1776), productivity gains

from specialization can be large, stressing that specialization is an important contributor to wealth creation.[25]

The presence of nonconvexity can provide useful insights into producer behavior. This was illustrated in Figure 2b, where producers in competitive markets would fail to use an input (by choosing point A) even if positive input use is efficient (point O). Such patterns (which arise only under a nonconvex technology) can help explain why African farmers seem to "underutilize" fertilizers (Duflo, Kremer, and Robinson 2011). In this case, our analysis indicates that nonlinear pricing would provide an efficient way to nudge African farmers toward efficiency.

Figures 3b and 4b illustrate the effects of nonconvexity due to the benefits from specialization. In these figures, competitive markets would provide incentive to specialize in y_2 (choosing the inefficient point C) even if the efficient allocation is to diversify (point O). This seems relevant for environmental management when y_1 denotes wildlife and y_2 is agriculture. In this case, Figures 3b and 4b predict that competitive markets would lead farmers/ranchers to specialize in agriculture at the expense of wildlife.[26] These arguments raise significant concerns about current environmental management and policy. A good historical example is the story of the North American bison: markets did nothing to stop the slaughter and virtual extinction of the bison in the nineteenth century (Taylor 2011). Our analysis indicates that this outcome was not due to profit maximization; rather, it was due to the inefficiency of uniform pricing.

Our discussion of efficiency analysis makes it clear that a narrow focus on uniform pricing is often inappropriate. It underscores the need to incorporate nonlinear pricing as an integral part of resource management, policy design, and economic evaluation.

About the Authors: Jean-Paul Chavas is a Professor with the University of Wisconsin. (Corresponding Author Email: jchavas@wisc.edu)

Acknowledgments: I would like to thank two anonymous reviewers and the AETR Editor for useful comments made on an earlier draft of the paper.

[25] Unfortunately, Smith (1776) missed the point that, under uniform pricing, specialization can create nonconvexity, which can lead competitive markets to be inefficient.

[26] A similar argument could apply to diversification within agriculture: competitive markets may induce farmers to specialize in the most productive crops, leading to "overspecialization" and inefficient use of environmental services. Such issues can be exacerbated by R&D that often favors dominant crops, thus strengthening the overspecialization argument. This discussion indicates the need to rethink current agricultural and environmental policies.

Appendix A

The analysis proceeds in several steps. First, starting with the efficient allocation given in (7), we follow Luenberger (1992, 1995) to examine a dual formulation of efficiency. Second, we use this dual formulation to define the shadow prices of the netputs y under general conditions. Third, we examine how profit maximization and efficient pricing can support an efficient allocation.

Our analysis will rely on the following assumptions.

Assumption As1 (non-satiation in g): $u(x + \beta\, g) > u(x)$ for all $x \in X$ and all $\beta > 0$.

Assumption As2 (no destitution): there is a $\gamma > 0$ such that $(x^* - \gamma\, g) \in X$.

Following Luenberger (1992, 1995), consider the dual formulation to (7):

$$B^* = \max_{x,y} \{B(x, U^*): x \le y, x \in X, y \in T\}, \tag{A1}$$

where $B(x, U)$ is the benefit function defined as

$$\begin{aligned} B(x, U) &= \max_{\beta} \{\beta: u(x - \beta\, g) \ge U): x \in X\} \text{ if a maximum exists} \\ &= -\infty \text{ otherwise.} \end{aligned} \tag{A2}$$

Under assumption As1 and evaluated at U^* in (7), note that (A1) satisfies $B^* = 0$, leading Luenberger (1992, 1995) to call (A1) "zero maximality."

Lemma 1: Under assumption As1–As2, an allocation (x^*, y^*) is efficient if and only if it satisfies (A1).

Proof: Let (x^*, y^*) be an efficient allocation satisfying (7). Assume that (x^*, y^*) does not maximize (A1). Then, there exists a feasible (x_a, y_a) such that $B\big(x_a, u(x^*)\big) > B\big(x^*, u(x^*)\big)$ where $B\big(x^*, u(x^*)\big) = 0$ under As1. This implies that $u(x_a) > u(x_a - B(x_a, u(x^*))\, g) = u(x^*)$. But this contradicts (x^*, y^*) being efficient. Thus, economic efficiency implies (A1).

Next, let (x^*, y^*) be a solution to (A1). Assume that (x^*, y^*) is not efficient (i.e., that (x^*, y^*) does not maximize (7)). Then, there exists a feasible allocation (x_b, y_b) such that $u(x_b) > u(x^*)$. Under As1–As2, this implies that $B(x_b, u(x^*)) > B\big(x^*, u(x^*)\big) = 0$. But this contradicts (x^*, y^*) being a solution to (A1). Thus, (A1) implies economic efficiency.

Q.E.D.

The function $B(x, U)$ in (A2) gives the number of units of the reference bundle g the consumer is willing to give up starting at x to reach utility level U. Note that $B(x, U)$ in (A2) satisfies $B(x + \alpha\, g, U) = \alpha + B(x, U)$, implying that $\frac{\partial B}{\partial \alpha} = \frac{\partial B}{\partial x}\, g = 1$. When the bundle g is chosen such that one unit of g is worth \$1, it follows that the benefit function $B(x, U)$ in (6) is a measure of the consumer's willingness-to-pay. In a way consistent with Luenberger (1992, 1995), the solution (x^*, y^*) to the maximization problem in (A1) is efficient. This is intuitive: economic efficiency is equivalent to choosing a feasible allocation that maximizes consumer benefit. As such, lemma 1 establishes that (A1) is a valid dual formulation of economic efficiency. Importantly, this formulation holds in general: it applies for any technology T, and it applies to market economies as well as situations where there are nonmarket goods.

Noting that equation (A1) involves the feasibility constraints $x \le y$, we are interested in evaluating the shadow prices of these constraints. This can be done considering a Lagrangian formulation associated with (A1). Let H be the class of absolutely continuous and non-decreasing

functions $h: \mathbb{R}^m \to \mathbb{R}$ satisfying $h(0) = 0$ and $h(z + \alpha\, g) = \alpha + h(z)$ for $z \in \mathbb{R}^m$ and $\alpha \in \mathbb{R}$. Following Gould (1969), consider the generalized Lagrangian:

$$L(x, y, h) = B(x, U^*) + h(y) - h(x), (x, y, h) \in X \times T \times H. \quad \text{(A3)}$$

Note that (A3) includes the standard Lagrangian as a special case when $h(z)$ is linear, with $h(z) = \sum_{j=1}^{m} p_j\, z_j$, p_j being the Lagrange multiplier for the j-th constraint, $x_j \le y_j, j = 1, \dots, m$.

Consider a saddle-point $(x^*, y^*, h^*) \in X \times T \times H$ of the Lagrangian $L(x, y, h)$ satisfying:

$$L(x, y, h^*) \le L(x^*, y^*, h^*) \le L(x^*, y^*, h), \quad \text{(A4)}$$

for all $(x, y, h) \in X \times T \times H$.

Lemma 2: Let (x^*, y^*, h^*) be a saddle-point of $L(x, y, h)$ satisfying (A4). Then, (x^*, y^*) solves (A1).

Proof: The proof follows the arguments presented in Gould (1969). From the second inequality in (A4), we have

$$h^*(y^*) - h^*(x^*) \le h(y^*) - h(x^*) \text{ for all } h \in H. \quad \text{(A5)}$$

Assume that (x^*, y^*, h^*) does not satisfy $x^* \le y^*$. This means that there exists a $j \in \{1, \dots, m\}$ such that $x_j^* > y_j^*$. Let $h_a(z) = k\, z_j$, where $k > 0$. Noting that $h_a \in H$, (A5) gives $h^*(y^*) - h^*(x^*) \le h_a(y^*) - h_a(x^*) = k\left(y_j^* - x_j^*\right) < 0$. Letting $k \to \infty$, it follows that $[h_a(y^*) - h_a(x^*)]$ does not have a lower bound. But this contradicts (A5). We conclude that:

$$x^* \le y^*. \quad \text{(A6)}$$

Choosing $h_b = 0$ and noting that $h_b \in H$, (A5) implies that $[h^*(y^*) - h^*(x^*)] \le 0$. Having $h \in H$ as a non-decreasing function, it follows from (A6) that $h^*(x^*) \le h^*(y^*)$. Combining these two results gives the complementary slackness condition:

$$h^*(x^*) = h^*(y^*). \quad \text{(A7)}$$

From the first inequality in (A4), we have

$$B(x, U^*) + h^*(y) - h^*(x) \le B(x^*, U^*) + h^*(y^*) - h^*(x^*). \quad \text{(A8)}$$

Using (A7), this gives

$$B(x^*, U^*) \ge B(x, U^*) + h^*(y) - h^*(x), x \in X, y \in Y$$

or

$$B(x^*, U^*) \ge B(x, U^*) \text{ when } x \le y, x \in X, y \in Y, \quad \text{(A9)}$$

since $x \le y$ and $h^* \in H$ being a non-decreasing function imply that $h^*(y) - h^*(x) \ge 0$. Equation (A9) proves that (x^*, y^*) is a solution to (A1). In addition, note that (A7) implies that

$L(x^*, y^*, h^*) = B(x^*, U^*) = B^*$ (i.e., that the Lagrangian $L(x^*, y^*, h^*)$ is a measure of consumer benefit at the optimum).

Q.E.D.

Combined with lemma 1, lemma 2 provides an alternative formulation of economic efficiency. We now use this formulation to evaluate the shadow prices of netputs y. As a modified version of our analysis, consider the case where $\delta \in \mathbb{R}^m$ is a vector of initial endowment for the netputs. Then, for a given δ and from (7), consider the allocation (x_δ^*, y_δ^*) satisfying:

$$\max_{x,y} \{u(x): x \leq y + \delta, x \in X, y \in T\}, \tag{7'}$$

and the dual formulation (A1) becomes:

$$B^*(\delta) = \max_{x,y} \{B(x, U^*): x \leq y + \delta\, , x \in X, y \in T\}. \tag{A1'}$$

The associated generalized Lagrangian is

$$L'(x, y, h, \delta) = B(x, U^*) + h(y + \delta) - h(x), (x, y, h) \in X \times T \times H, \tag{A3'}$$

where $L'(x, y, h, 0) = L(x, y, h)$. For a given δ, a saddle-point $(x_\delta^*, y_\delta^*, h_\delta^*) \in X \times T \times H$ of $L'(x, y, h, \delta)$ satisfies:

$$L'(x, y, h_\delta^*\, , \delta) \leq\ L'(x_\delta^*, y_\delta^*, h_\delta^*, \delta) \leq L'(x_\delta^*, y_\delta^*, h, \delta) \tag{A4'}$$

for all $(x, y, h) \in X \times T \times H$.

Lemma 3: Assume that $B^*(\delta)$ and h_δ^* are differentiable in δ. Then, evaluated at point $\delta = 0$, we have:

$$\frac{\partial B^*(\delta)}{\partial \delta} = \frac{\partial h^*(y)}{\partial y} \text{ evaluated at } y^*. \tag{A10}$$

Proof: Consider any two δ and $\delta' \in \mathbb{R}^m$. The first inequality in (A4') implies that $L'(x, y, h_\delta^*\, , \delta) \leq\ L'(x_\delta^*, y_\delta^*, h_\delta^*, \delta)$ for any $(x, y) \in X \times T$. Letting $x = x_{\delta'}^*$ and $y = y_{\delta'}^*$, this gives

$$L'(x_\delta^*, y_\delta^*, h_\delta^*, \delta) \geq L'\left(x_{\delta'}^*, y_{\delta'}^*, h_\delta^*\, , \delta\right) =\ B\left(x_{\delta'}^*, U^*\right) + h_\delta^*\left(y_{\delta'}^* + \delta\right) - h_\delta^*(x_{\delta'}^*). \tag{A11}$$

Since $\left(x_{\delta'}^*, y_{\delta'}^*\right)$ is feasible and $h_\delta^* \in H$ is a non-decreasing function, we have $x_{\delta'}^* \leq y_{\delta'}^* + \delta'$ and $h_\delta^*\left(y_{\delta'}^* + \delta'\right) - h_\delta^*\left(x_{\delta'}^*\right) \geq 0$. This implies:

$$B\left(x_{\delta'}^*, U^*\right) + h_\delta^*\left(y_{\delta'}^* + \delta'\right) - h_\delta^*\left(x_{\delta'}^*\right) \geq B\left(x_{\delta'}^*, U^*\right) = L'\left(x_{\delta'}^*, y_{\delta'}^*, h_{\delta'}^*\, , \delta'\right). \tag{A12}$$

Summing (A11) and (A12) yields:

$$L'(x_\delta^*, y_\delta^*, h_\delta^*, \delta) - L'\left(x_{\delta'}^*, y_{\delta'}^*, h_{\delta'}^*\, , \delta'\right) \geq h_\delta^*\left(y_{\delta'}^* + \delta\right) - h_\delta^*\left(y_{\delta'}^* + \delta'\right). \tag{A13}$$

Switching δ and δ' and multiplying (A13) by -1, we obtain:

$$h^*_{\delta'}(y^*_\delta + \delta) - h^*_{\delta'}(y^*_\delta + \delta') \geq L'(x^*_\delta, y^*_\delta, h^*_\delta, \delta) - L'\left(x^*_{\delta'}, y^*_{\delta'}, h^*_{\delta'}, \delta'\right) \geq h^*_\delta\left(y^*_{\delta'} + \delta\right) - h^*_\delta\left(y^*_{\delta'} + \delta'\right). \quad \text{(A14)}$$

Note that $B^*(\delta) = B(x^*_\delta, U^*) = L'(x^*_\delta, y^*_\delta, h^*_\delta, \delta)$. Then, when $B^*(\delta)$ and $h^*(y + \delta)$ are differentiable in δ and letting $\delta' \to \delta$, equation (A10) follows from (A14) evaluated at $\delta = 0$.

Q.E.D.

Lemma 3 establishes that, in efficient allocations, $\frac{\partial h^*(y)}{\partial y}$ can be interpreted as the shadow prices of the netputs y^*, shadow prices that reflect the marginal benefit $\frac{\partial B^*(\delta)}{\partial \delta}$. This result applies when y includes market goods (in which case $h^*(y)$ is the market value of the goods y) as well as nonmarket goods. When $h^*(y)$ is linear with $h^*(y) = \sum_{j=1}^{m} p^*_j\ y_j$, then $p^* = (p^*_1, \ldots, p^*_m)$ become the standard Lagrange multipliers. In this case, as a special case, we obtain the well-known results that Lagrange multipliers measure shadow prices (e.g., Takayama 1985, pp. 135–139). Importantly, this interpretation continues to hold when $h^*(y)$ is nonlinear (corresponding to nonlinear pricing). In addition, from (A6) and (A7), note that the pricing scheme $h^*(y)$ guarantees feasibility and the equilibrium of supply and demand.

Next, we evaluate the implications of our analysis for efficient production and efficient pricing.

Lemma 4: Economic efficiency implies that:

$$E(h) = \min_x\ \{h(x) - B(x, U^*) : x \in X\} \quad \text{(A15)}$$

$$\pi(h) = \max_y\ \{h(y) : y \in Y\} \quad \text{(A16)}$$

$$0 = B^* = \min_h\ \{\pi(h) - E(h) : h \in H\}. \quad \text{(A17)}$$

Proof: Given $L(x, y, h) = B(x, U^*) + h(y) - h(x)$ given in (A3), equations (A15), A(16), and A(17) follow directly from a stage-wise decomposition of the saddle-point problem in (A4).

The function $E(h)$ in (A15) can be interpreted as a consumer expenditure function. To see this, note that:

$$\min_x\ \{h(x) : u(x) \geq U^*, x \in X\}$$
$$= h(x') \text{ where } x' \in X \text{ solves the minimization problem and satisfies } u(x') \geq U^*,$$
$$\geq h(x') - B(x', U^*) \text{ since } x' \in X \text{ and } u(x') \geq U^* \text{ imply that } B(x', U^*) \geq 0.$$

Also, for any $x \in X$ where $B(x, U^*)$ is finite, we have $x'' \equiv [x - B(x, U^*)\ g] \in X$ satisfying $u(x'') \geq U^*$. It follows that:

$$\min_x\ \{h(x) : u(x) \geq U^*, x \in X\}$$
$$\leq h(x'') = h(x - B(x, U^*)\ g)$$
$$= h(x) - B(x, U^*).$$

Combining these two expressions and using (A15) imply that $\min_x\ \{h(x) : u(x) \geq U^*, x \in X\} = E(h)$ (i.e., that $E(h)$ in (A15) is a consumer expenditure function).

Q.E.D.

Conditional on h, equation (A15) states that efficient consumption minimizes consumer expenditures. Similarly, conditional on h, equation (A16) states that efficient production maximizes profit. Finally, equation (A17) identifies the pricing scheme h^* that achieves efficiency. Here, efficient pricing has two functions: (1) to clear the market (so that (x^*, y^*) satisfy (A6) and (A7)) and (2) to support an allocation that maximizes consumer benefit (as given in (A1)). Combining (A16) and (A17), it follows that profit maximizing netputs support an efficient allocation under pricing scheme h^*. Under convexity assumptions, uniform pricing (where $h(y)$ is linear) can always be used to attain efficiency (e.g., Takayama 1985, pp. 66–74). However, as discussed in the text, uniform pricing can be inefficient under nonconvexity, in which case nonlinear pricing (where $h(y)$ is nonlinear) would be required to support an efficient allocation.

References

Arrow, K. 1951. "An Extension of the Basic Theorems of Classical Welfare Economics." In J. Neyman, ed. *Proceedings of the Second Berkely Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press, pp. 507–532.

Baumol, W.J., J.C. Panzar, and R.D. Willig. 1982. *Contestable Markets and the Theory of Industry Structure*. New York: Harcourt Brace Jovanovich, Inc.

Baumol, W.J., and W.E. Oates. 1988. *The Theory of Environmental Policy*, 2nd ed. Cambridge: Cambridge University Press.

Borenstein, S., and J.B. Bushnell. 2022. "Do Two Electric Pricing Wrongs Make a Right? Cost Recovery, Externalities and Efficiency." *American Economic Journal: Economic Policy* 14(4):80–110.

Chambers, R., Y. Chung, and R. Fare. 1996. "Benefit and Distance Functions." *Journal of Economic Theory* 70(2):407–419.

Chambers, R., Y. Chung, and R. Fare. 1998. "Profit, Directional Distance Functions and Nerlovian Efficiency." *Journal of Optimization Theory and Applications* 98(2):351–364.

Chavas, J.P. 2015. "Coase Revisited: Economic Efficiency under Externalities, Transaction Costs and Nonconvexity." *Journal of Institutional and Theoretical Economics* 171:709–734.

Coase, R.H. 1960. "The Problem of Social Cost." *Journal of Law and Economics* 3:1–44.

Coelli, T.J., D.S.P. Rao, and G.E. Battese. 2005. *An Introduction to Efficiency and Productivity Measurement*. New York: Springer.

Debreu, G. 1959. *Theory of Value*. New York: Wiley.

Duflo, E., M. Kremer, and J. Robinson. 2011. "Nudging Farmers to Use Fertilizer: Theory and Experimental Evidence from Kenya." *American Economic Review* 101:2350–2390.

Farrell, M.J. 1957. "The Measurement of Productive Efficiency." *Journal of the Royal Statistical Society (ser. A, General)* 120:253–281.

Gould, F.J. 1969. "Extensions of Lagrangian Multipliers in Nonlinear Programming." *SIAM Journal of Applied Mathematics* 17:1280–1297.

Graaff, J. de V. 1967. *Theoretical Welfare Economics*. Cambridge, England: Cambridge University Press.

Luenberger, D.G. 1992. "New Optimality Principles for Economic Efficiency and Equilibrium." *Journal of Optimization Theory and Applications* 75:221–264.

Luenberger, D.G. 1995. *Microeconomic Theory*. New York: McGraw-Hill.

Murty, S., R.R. Russell, and S.B. Levkoff. 2012. "On Modeling Pollution-Generating Technologies." *Journal of Environmental Economics and Management* 64(1):117–135.

Pigou, A.C. 1920. *The Economics of Welfare*. London: Macmillan.

Ray, S.C., R.G. Chambers, and S.C. Kumbhakar. 2022. *Handbook of Production Economics*. Singapore: Springer Nature.

Shephard, R.W. 1970. *Theory of Cost and Production Functions*. Princeton, NJ: Princeton University Press.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. Vol. II, 1st ed. London: W. Strahan & T. Cadell.

Takayama, A. 1985. *Mathematical Economics*, 2nd ed. Cambridge, England: Cambridge University Press.

Taylor, M.S. 2011. "Buffalo Hunt: International Trade and the Virtual Extinction of the North American Bison." *American Economic Review* 101(7):3162–3195.

Wilson, R.B. 1993. *Nonlinear Pricing*. Oxford: Oxford University Press.

DOI: https://doi.org/10.71162/aetr.539094